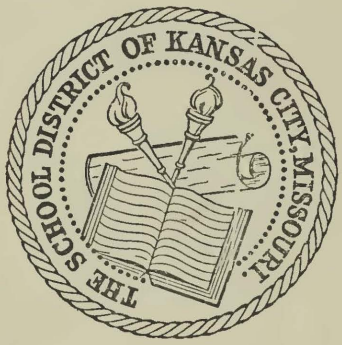
Periodical 8:53:625

Kansas City Public Library



This Volume is for

REFERENCE USE ONLY

8-33—6m—P

From the collection of the



San Francisco, California
2008

PUBLIC LIBRARY
KANSAS CITY
MO

VOLUME XIII JANUARY, 1934 NUMBER 1

THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

Stabilized Feedback Amplifiers—*H. S. Black* . . . 1

Open-Wire Crosstalk—*A. G. Chapman* 19

Vacuum Tube Electronics at Ultra-high Frequencies—*F. B. Llewellyn* 59

Contemporary Advances in Physics, XXVII—The Nucleus, Second Part—*Karl K. Darrow* 102

Abstracts of Technical Papers 159

Contributors to this Issue 161

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

50c per Copy *$1.50 per Year*

VANDERBILT UNIV
VTIC LIBRARY
ON

THE BELL SYSTEM TECHNICAL JOURNAL

Published quarterly by the
American Telephone and Telegraph Company
195 Broadway, New York, N. Y.

EDITORIAL BOARD

Bancroft Gherardi	H. P. Charlesworth	F. B. Jewett
L. F. Morehouse	E. H. Colpitts	O. B. Blackwell
D. Levinger	O. E. Buckley	H. S. Osborne

Philander Norton, *Editor* J. O. Perrine, *Associate Editor*

SUBSCRIPTIONS

Subscriptions are accepted at $1.50 per year. Single copies are fifty cents each. The foreign postage is 35 cents per year or 9 cents per copy.

Copyright, 1934

PRINTED IN U. S. A.

THE BELL SYSTEM TECHNICAL JOURNAL

A JOURNAL DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

EDITORIAL BOARD

BANCROFT GHERARDI H. P. CHARLESWORTH F. B. JEWETT
L. F. MOREHOUSE E. H. COLPITTS O. B. BLACKWELL
D. LEVINGER O. E. BUCKLEY H. S. OSBORNE
PHILANDER NORTON, *Editor* J. O. PERRINE, *Associate Editor*

TABLE OF CONTENTS AND INDEX

VOLUME XIII

1934

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

Bound
Periodical

PRINTED IN U. S. A.

DEC 21 '34

THE BELL SYSTEM TECHNICAL JOURNAL

VOLUME XIII, 1934

Table of Contents

JANUARY, 1934

Stabilized Feedback Amplifiers—*H. S. Black* 1
Open-Wire Crosstalk—*A. G. Chapman* 19
Vacuum Tube Electronics at Ultra-high Frequencies—*F. B. Llewellyn* 59
Contemporary Advances in Physics, XXVII—The Nucleus, Second Part—*Karl K. Darrow* 102

APRIL, 1934

The Carbon Microphone: An Account of Some Researches Bearing on Its Action—*F. S. Goucher* 163
Open-Wire Crosstalk—*A. G. Chapman* 195
Symposium on Wire Transmission of Symphonic Music and Its Reproduction in Auditory Perspective:
Basic Requirements—*Harvey Fletcher* 239
Physical Factors—*J. C. Steinberg and W. B. Snow* 245
Loud Speakers and Microphones—*E. C. Wente and A. L. Thuras* 259
Amplifiers—*E. O. Scriven* 278
Transmission Lines—*H. A. Affel, R. W. Chesnut and R. H. Mills* 285
System Adaptation—*E. H. Bedell and Iden Kerney* 301

JULY, 1934

The Compandor—An Aid Against Static in Radio Telephony—*R. C. Mathes and S. B. Wright* 315
The Effect of Background Noise in Shared Channel Broadcasting—*C. B. Aiken* 333
Wide-Band Open-Wire Program System—*H. S. Hamilton* 351
Line Filter for Program System—*A. W. Clement* 382
Contemporary Advances in Physics, XXVIII—The Nucleus, Third Part—*Karl K. Darrow* 391
Electrical Wave Filters Employing Quartz Crystals as Elements—*W. P. Mason* 405
Some Improvements in Quartz Crystal Circuit Elements—*F. R. Lack, G. W. Willard and I. E. Fair* 453
A Theory of Scanning and Its Relation to the Characteristics of the Transmitted Signal in Telephotography and Television—*Pierre Mertz and Frank Gray* 464

OCTOBER, 1934

An Extension of the Theory of Three-Electrode Vacuum Tube Circuits—*S. A. Levin and Liss C. Peterson* 523
The Electromagnetic Theory of Coaxial Transmission Lines and Cylindrical Shields—*S. A. Schelkunoff* 532
Contemporary Advances in Physics, XXVIII—The Nucleus, Third Part—*Karl K. Darrow* 580
The Measurement and Reduction of Microphonic Noise in Vacuum Tubes—*D. B. Penick* 614
Fluctuation Noise in Vacuum Tubes—*G. L. Pearson* 634
Systems for Wide-Band Transmission Over Coaxial Lines—*L. Espenschied and M. E. Strieby* 654
Regeneration Theory and Experiment—*E. Peterson, J. G. Kreer and L. A. Ware* 680

Index to Volume XIII

A

Affel, H. A., R. W. Chesnut and R. H. Mills, Transmission Lines, page 285.
Aiken, C. B., The Effect of Background Noise in Shared Channel Broadcasting, page 333.
Amplifiers, *E. O. Scriven*, page 278.
Amplifiers, Stabilized Feedback, *H. S. Black*, page 1.
Auditory Perspective, Symposium on Wire Transmission of Symphonic Music and Its Reproduction in, pages 239–301.

B

Basic Requirements (of Wire Transmission of Symphonic Music and Its Reproduction in Auditory Perspective), *Harvey Fletcher*, page 239.
Bedell, E. H. and Iden Kerney, System Adaptation (of Symposium on Wire Transmission of Symphonic Music and Its Reproduction in Auditory Perspective), page 301.
Black, H. S., Stabilized Feedback Amplifiers, page 1.
Broadcasting, Shared Channel, The Effect of Background Noise in, *C. B. Aiken*, page 333.

C

Chapman, A. G., Open-Wire Crosstalk, pages 19 and 195.
Chesnut, R. W., H. A. Affel and R. H. Mills, Transmission Lines, page 285.
Clement, A. W., Line Filter for Program System, page 382.
Coaxial Lines, Systems for Wide-Band Transmission Over, *L. Espenschied and M. E. Strieby*, page 654.
Coaxial Transmission Lines and Cylindrical Shields, The Electromagnetic Theory of, *S. A. Schelkunoff*, page 532.
Compandor, The—An Aid Against Static in Radio Telephony, *R. C. Mathes and S. B. Wright*, page 315.
Contemporary Advances in Physics, XXVII—The Nucleus, Second Part, *Karl K. Darrow*, page 102.
Contemporary Advances in Physics, XXVIII—The Nucleus, Third Part, *Karl K. Darrow*, pages 391 and 580.
Crosstalk, Open-Wire, *A. G. Chapman*, pages 19 and 195.

D

Darrow, Karl K., Contemporary Advances in Physics, XXVII—The Nucleus, Second Part, page 102.
Contemporary Advances in Physics, XXVIII—The Nucleus, Third Part, pages 391 and 580.

E

Electromagnetic Theory of Coaxial Transmission Lines and Cylindrical Shields, The, *S. A. Schelkunoff*, page 532.
Espenschied, L. and M. E. Strieby, Systems for Wide-Band Transmission Over Coaxial Lines, page 654.

F

Fair, I. E., F. R. Lack and G. W. Willard, Some Improvements in Quartz Crystal Circuit Elements, page 453.
Filter, Line, for Program System, *A. W. Clement*, page 382.

Filters, Electrical Wave, Employing Quartz Crystals as Elements, *W. P. Mason* page 405.
Fletcher, Harvey, Basic Requirements (of Wire Transmission of Symphonic Music and its Reproduction in Auditory Perspective), page 239.

G

Goucher, F. S., The Carbon Microphone: An Account of Some Researches Bearing on its Action, page 163.
Gray, Frank and Pierre Mertz, A Theory of Scanning and Its Relation to the Characteristics of the Transmitted Signal in Telephotography and Television, page 464.

H

Hamilton, H. S., Wide-Band Open-Wire Program System, page 351.

K

Kerney, Iden and E. H. Bedell, System Adaptation (of Symposium on Wire Transmission of Symphonic Music and Its Reproduction in Auditory Perspective), page 301.
Kreer, J. G., L. A. Ware and E. Peterson, Regeneration Theory and Experiment, page 680.

L

Lack, F. R., G. W. Willard and I. E. Fair, Some Improvements in Quartz Crystal Circuit Elements, page 453.
Levin, S. A. and Liss C. Peterson, An Extension of the Theory of Three-Electrode Vacuum Tube Circuits, page 523.
Llewellyn, F. B., Vacuum Tube Electronics at Ultra-high Frequencies, page 59.
Loud Speakers and Microphones, *E. C. Wente and A. L. Thuras*, page 259.

M

Mason, W. P., Electrical Wave Filters Employing Quartz Crystals as Elements, page 405.
Mathes, R. C. and S. B. Wright, The Compandor—An Aid Against Static in Radio Telephony, page 315.
Mertz, Pierre and Frank Gray, A Theory of Scanning and Its Relation to the Characteristics of the Transmitted Signal in Telephotography and Television, page 464.
Microphone, The Carbon: An Account of Some Researches Bearing on Its Action. *F. S. Goucher*, page 163.
Microphones, Loud Speakers and, *E. C. Wente and A. L. Thuras*, page 259.
Microphonic Noise in Vacuum Tubes, The Measurement and Reduction of, *D. B. Penick*, page 614.
Mills, R. H., H. A. Affel and R. W. Chesnut, Transmission Lines, page 285.
Music, Symphonic, Symposium on Wire Transmission of and Its Reproduction in Auditory Perspective. pages 239–301.

N

Noise, Background, The Effect of in Shared Channel Broadcasting, *C. B. Aiken*, page 333.
Noise in Vacuum Tubes, Fluctuation, *G. L. Pearson*, page 634.
Noise, Microphonic, The Measurement and Reduction of in Vacuum Tubes, *D. B. Penick*, page 614.

P

Pearson, G. L., Fluctuation Noise in Vacuum Tubes, page 634.
Penick, D. B., The Measurement and Reduction of Microphonic Noise in Vacuum Tubes, page 614.

Peterson, E., J. G. Kreer and L. A. Ware, Regeneration Theory and Experiment, page 680.
Peterson, Liss C. and S. A. Levin, An Extension of the Theory of Three-Electrode Vacuum Tube Circuits, page 523.
Physical Factors (of Wire Transmission of Symphonic Music and Its Reproduction in Auditory Perspective), *J. C. Steinberg and W. B. Snow,* page 245.
Physics, XXVII, Contemporary Advances in—The Nucleus, Second Part, *Karl K. Darrow,* page 102.
Physics, XXVIII, Contemporary Advances in—The Nucleus, Third Part, *Karl K. Darrow,* pages 391 and 580.

Q

Quartz Crystal Circuit Elements, Some Improvements in, *F. R. Lack, G. W. Willard and I. E. Fair,* page 453.
Quartz Crystals in Elements, Electrical Wave Filters Employing, *W. P. Mason,* page 405.

R

Radio: Line Filter for Program System, *A. W. Clement,* page 382.
Radio: The Effect of Background Noise in Shared Channel Broadcasting, *C. B. Aiken,* page 333.
Radio: Wide-Band Open-Wire Program System, *H. S. Hamilton,* page 351.
Radio Telephony, The Compandor—An Aid Against Static in, *R. C. Mathes and S. B. Wright,* page 315.
Regeneration Theory and Experiment, *E. Peterson, J. G. Kreer, and L. A. Ware,* page 680.

S

Schelkunoff, S. A., The Electromagnetic Theory of Coaxial Transmission Lines and Cylindrical Shields, page 532.
Scriven, E. O., Amplifiers, page 278.
Snow, W. B. and J. C. Steinberg, Physical Factors (of Wire Transmission of Symphonic Music and Its Reproduction in Auditory Perspective), page 245.
Steinberg, J. C. and W. B. Snow, Physical Factors (of Wire Transmission of Symphonic Music and Its Reproduction in Auditory Perspective), page 245.
Strieby, M. E. and L. Espenschied, Systems for Wide-Band Transmission Over Coaxial Lines, page 654.

T

Telephotography and Television, A Theory of Scanning and Its Relation to the Characteristics of the Transmitted Signal in, *Pierre Mertz and Frank Gray,* page 464.
Television, A Theory of Scanning and Its Relation to the Characteristics of the Transmitted Signal in Telephotography and, *Pierre Mertz and Frank Gray,* page 464.
Thuras, A. L. and E. C. Wente, Loud Speakers and Microphones, page 259.
Transmission Lines, *H. A. Affel, R. W. Chesnut and R. H. Mills,* page 285.

V

Vacuum Tube Circuits, Three-Electrode, An Extension of the Theory of, *S. A. Levin and Liss C. Peterson,* page 523.
Vacuum Tube Electronics at Ultra-high Frequencies, *F. B. Llewellyn,* page 59.
Vacuum Tubes, Fluctuation Noise in, *G. L. Pearson,* page 634.
Vacuum Tubes, The Measurement and Reduction of Microphonic Noise in, *D. B. Penick,* page 614.

W

Ware, L. A., E. Peterson and J. G. Kreer, Regeneration Theory and Experiment, page 680.
Wente, E. C. and A. L. Thuras, Loud Speakers and Microphones, page 259.
Wide-Band Open-Wire Program System, *H. S. Hamilton*, page 351.
Wide-Band Transmission Over Coaxial Lines, Systems for, *L. Espenschied and M. E. Strieby*, page 654.
Wide Band: The Electromagnetic Theory of Coaxial Transmission Lines and Cylindrical Shields, *S. A. Schelkunoff*, page 532.
Wide Band: Symposium on Wire Transmission of Symphonic Music and Its Reproduction in Auditory Perspective, pages 239–301.
Wide Band: Stabilized Feedback Amplifiers, *H. S. Black*, page 1.
Willard, G. W., I. E. Fair and F. R. Lack, Some Improvements in Quartz Crystal Circuit Elements, page 453.
Wright, S. B. and R. C. Mathes, The Compandor—An Aid Against Static in Radio Telephony, page 315.

The Bell System Technical Journal

January, 1934

Stabilized Feedback Amplifiers*

By H. S. BLACK

This paper describes and explains the theory of the feedback principle and then demonstrates how stability of amplification and reduction of modulation products, as well as certain other advantages, follow when stabilized feedback is applied to an amplifier. The underlying principle of design by means of which singing is avoided is next set forth. The paper concludes with some examples of results obtained on amplifiers which have been built employing this new principle.

The carrier-in-cable system dealt with in a companion paper [1] involves many amplifiers in tandem with many telephone channels passing through each amplifier and constitutes, therefore, an ideal field for application of this feedback principle. A field trial of this system was made at Morristown, New Jersey, in which seventy of these amplifiers were operated in tandem. The results of this trial were highly satisfactory and demonstrated conclusively the correctness of the theory and the practicability of its commercial application.

Introduction

DUE TO advances in vacuum tube development and amplifier technique, it is now possible to secure any desired amplification of the electrical waves used in the communication field. When many amplifiers are worked in tandem, however, it becomes difficult to keep the overall circuit efficiency constant, variations in battery potentials and currents, small when considered individually, adding up to produce serious transmission changes for the overall circuit. Furthermore, although it has remarkably linear properties, when the modern vacuum tube amplifier is used to handle a number of carrier telephone channels, extraneous frequencies are generated which cause interference between the channels. To keep this interference within proper bounds involves serious sacrifice of effective amplifier capacity or the use of a push-pull arrangement which, while giving some increase in capacity, adds to maintenance difficulty.

However, by building an amplifier whose gain is deliberately made, say 40 decibels higher than necessary (10,000 fold excess on energy basis), and then feeding the output back on the input in such a way

* Presented at Winter Convention of A. I. E. E., New York City, Jan. 23–26, 1934. Published in *Electrical Engineering*, January, 1934.

[1] "Carrier in Cable" by A. B. Clark and B. W. Kendall, presented at the A. I. E. E. Summer Convention, Chicago, Ill., June, 1933; published in *Electrical Engineering*, July, 1933, and in *Bell Sys. Tech. Jour.*, July, 1933.

as to throw away the excess gain, it has been found possible to effect extraordinary improvement in constancy of amplification and freedom from non-linearity. By employing this feedback principle, amplifiers have been built and used whose gain varied less than 0.01 db with a change in plate voltage from 240 to 260 volts and whose modulation products were 75 db below the signal output at full load. For an amplifier of conventional design and comparable size this change in plate voltage would have produced about 0.7 db variation while the modulation products would have been only 35 db down; in other words, 40 db reduction in modulation products was effected. (On an energy basis the reduction was 10,000 fold.)

Stabilized feedback possesses other advantages including reduced delay and delay distortion, reduced noise disturbance from the power supply circuits and various other features best appreciated by practical designers of amplifiers.

It is far from a simple proposition to employ feedback in this way because of the very special control required of phase shifts in the amplifier and feedback circuits, not only throughout the useful frequency band but also for a wide range of frequencies above and below this band. Unless these relations are maintained, singing will occur, usually at frequencies outside the useful range. Once having achieved a design, however, in which proper phase relations are secured, experience has demonstrated that the performance obtained is perfectly reliable.

Circuit Arrangement

In the amplifier of Fig. 1, a portion of the output is returned to the input to produce feedback action. The upper branch, called the μ-circuit, is represented as containing active elements such as an amplifier while the lower branch, called the β-circuit, is shown as a passive network. The way a voltage is modified after once traversing each circuit is denoted μ and β respectively and the product, $\mu\beta$, represents how a voltage is modified after making a single journey around amplifier and feedback circuits. Both μ and β are complex quantities, functions of frequency, and in the generalized concept either or both may be greater or less in absolute value than unity.[2]

Figure 2 shows an arrangement convenient for some purposes where, by using balanced bridges in input and output circuits, interaction between input and output is avoided and feedback action and amplifier impedances are made independent of the properties of circuits connected to the amplifier.

[2] μ is not used in the sense that it is sometimes used, namely, to denote the amplification constant of a particular tube, but as the complex ratio of the output to the input voltage of the amplifier circuit.

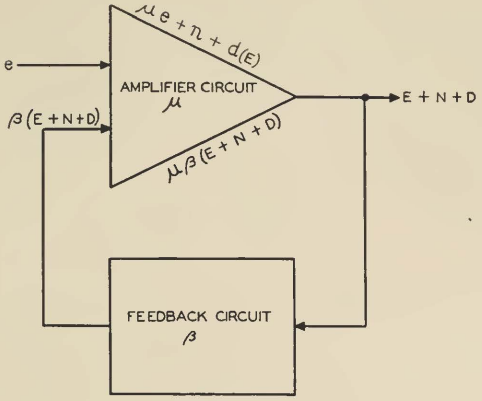


Fig. 1—Amplifier system with feedback.

e—Signal input voltage.
μ—Propagation of amplifier circuit.
μe—Signal output voltage without feedback.
n—Noise output voltage without feedback.
$d(E)$—Distortion output voltage without feedback.
β—Propagation of feedback circuit.
E—Signal output voltage with feedback.
N—Noise output voltage with feedback.
D—Distortion output voltage with feedback.

The output voltage with feedback is $E + N + D$ and is the sum of $\mu e + n + d(E)$, the value without feedback plus $\mu\beta[E + N + D]$ due to feedback.

$$E + N + D = \mu e + n + d(E) + \mu\beta[E + N + D]$$

$$[E + N + D](1 - \mu\beta) = \mu e + n + d(E)$$

$$E + N + D = \frac{\mu e}{1 - \mu\beta} + \frac{n}{1 - \mu\beta} + \frac{d(E)}{1 - \mu\beta}$$

If $|\mu\beta| \gg 1$, $E \doteq -\frac{e}{\beta}$. Under this condition the amplification is independent of μ but does depend upon β. Consequently the over-all characteristic will be controlled by the feedback circuit which may include equalizers or other corrective networks.

General Equation

In Fig. 1, β is zero without feedback and a signal voltage, e_0, applied to the input of the μ-circuit produces an output voltage. This is made up of what is wanted, the amplified signal, E_0, and components that are not wanted, namely, noise and distortion designated N_0 and D_0 and assumed to be generated within the amplifier. It is further assumed that the noise is independent of the signal and the distortion generator or modulation a function *only of the signal output.* Using the notation of Fig. 1, the output without feedback may be written as:

$$E_0 + N_0 + D_0 = \mu e_0 + n + d(E_0), \tag{1}$$

where zero subscripts refer to conditions without feedback.

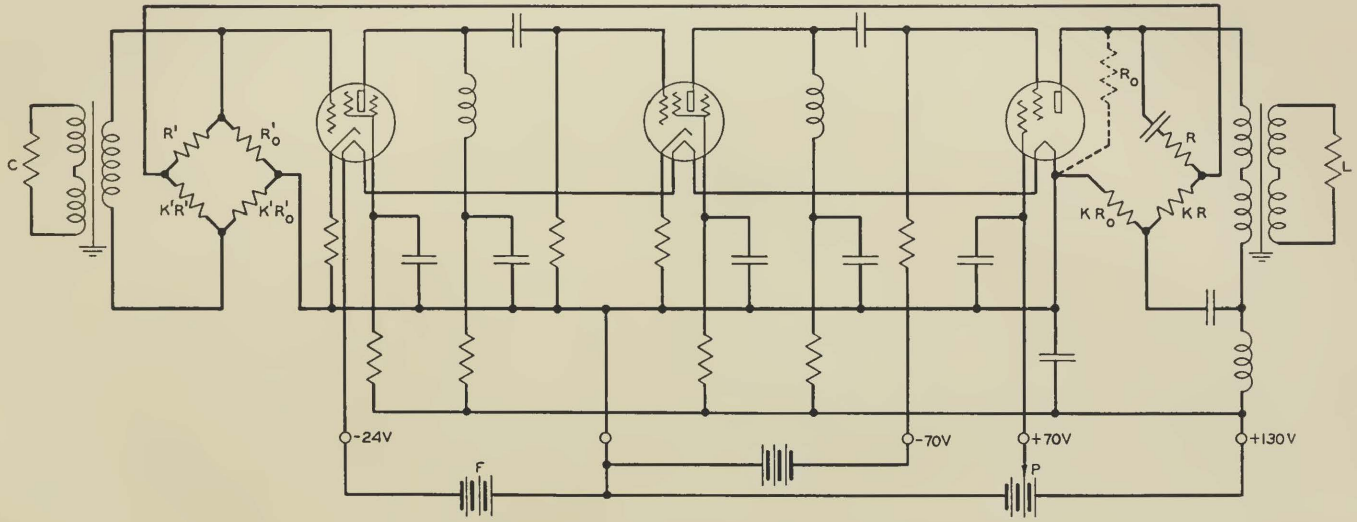


Fig. 2—Circuit of a negative feedback amplifier.

With feedback, β is not zero and the input to the μ-circuit becomes $e_0 + \beta(E + N + D)$. The output is $E + N + D$ and is equal to $\mu[e_0 + \beta(E + N + D)] + n + d(E)$ or:

$$E + N + D = \frac{\mu e_0}{1 - \mu\beta} + \frac{n}{1 - \mu\beta} + \frac{d(E)}{1 - \mu\beta}. \tag{2}$$

In the output, signal, noise and modulation are divided by $(1 - \mu\beta)$, and assuming $|1 - \mu\beta| > 1$, all are reduced.

Change in Gain Due to Feedback

From equation (2), the amplification with feedback equals the amplification without feedback divided by $(1 - \mu\beta)$. The effect of adding feedback, therefore, usually is to change the gain of the amplifier and this change will be expressed as:

$$G_{CF} = 20 \log_{10} \left| \frac{1}{1 - \mu\beta} \right|, \tag{3}$$

where G_{CF} is *db change in gain due to feedback.* $1/(1 - \mu\beta)$ will be used as a quantitative measure of the effect of feedback and the feedback referred to as *positive feedback* or *negative feedback* according as the absolute value of $1/(1 - \mu\beta)$ is greater or less than unity. Positive feedback increases the gain of the amplifier; negative feedback reduces it. The term *feedback* is not limited merely to those cases where the absolute value of $1/(1 - \mu\beta)$ is other than unity.

From $\mu\beta = |\mu\beta| \underline{|\Phi}$ and (3), it may be shown that:

$$10^{-G_{CF}/10} = 1 - 2|\mu\beta| \cos \Phi + |\mu\beta|^2, \tag{4}$$

which is the equation for a family of concentric circles of radii $10^{-G_{CF}/10}$ about the point 1, 0. Figure 3 is a polar diagram of the vector field of $\mu\beta = |\mu\beta| \underline{|\Phi}$. Using rectangular instead of polar coordinates, Fig. 4 corresponds to Fig. 3 and may be regarded as a diagram of the field of $\mu\beta$ where the parameter is db change in gain due to feedback. From these diagrams all of the essential properties of feedback action can be obtained such as change in amplification, effect on linearity, change in stability due to variations in various parts of the system, reduction of noise, etc. Certain significant boundaries have been designated similarly on both figures.

For example, boundary A is the locus of zero change in gain due to feedback. Along this parametric contour line where the absolute magnitude of amplification is not changed by feedback action, values of $|\mu\beta|$ range from zero to 2 and the phase shift, Φ, around the amplifier

and feedback circuits equals $\cos^{-1} |\mu\beta|/2$ and, therefore, lies between $-90°$ and $+90°$. For all conditions inside or above this boundary, the gain with feedback is increased; outside or below, the gain is *decreased.*

Stability

From equation (2), $\mu e_0/(1 - \mu\beta)$ is the amplified signal with feedback and, therefore, $\mu/(1 - \mu\beta)$ is an index of the amplification. It is of course a complex ratio. It will be designated A_F and referred to as the amplification with feedback.

To consider the effect of feedback upon stability of amplification, the stability will be viewed as the ratio of a change, δA_F, to A_F where δA_F is due to a change in either μ or β and the effects may be derived by assuming the variations are small.

$$A_F = \frac{\mu}{1 - \mu\beta}, \tag{5}$$

$$\left[\frac{\delta A_F}{A_F}\right]_\mu \doteq \frac{\left[\frac{\delta\mu}{\mu}\right]}{1 - \mu\beta}, \tag{6}$$

$$\left[\frac{\delta A_F}{A_F}\right]_\beta \doteq \frac{\mu\beta}{1 - \mu\beta}\left[\frac{\delta\beta}{\beta}\right]. \tag{7}$$

If $\mu\beta \gg 1$, it is seen that μ or the μ-circuit is stabilized by an amount corresponding to the reduction in amplification and the effect of introducing a gain or loss in the μ-circuit is to produce no material change in the overall amplification of the system; the stability of amplification as affected by β or the β-circuit is neither appreciably improved nor degraded since increasing the loss in the β-circuit raises the gain of the amplifier by an amount almost corresponding to the loss introduced and vice-versa. If μ and β are both varied and the variations sufficiently small, the effect is the same as if each were changed separately and the two results then combined.

In certain practical applications of amplifiers it is the change in gain or ammeter or voltmeter reading at the output that is a measure of the stability rather than the complex ratio previously treated. The conditions surrounding gain stability may be examined by considering the absolute value of A_F. This is shown as follows: Let (db) represent the gain in decibels corresponding to A_F. Then

$$(db) = 20 \log_{10} |A_F|,$$

$$\delta(db) \doteq 8.686 \left[\frac{\delta|A_F|}{|A_F|}\right]. \tag{8}$$

To get the absolute value of the amplification: Let

$$\mu\beta = |\mu\beta| \, \underline{\Phi}, \tag{9}$$

$$|A_F| = \frac{|\mu|}{\sqrt{1 - 2|\mu\beta| \cos \Phi + |\mu\beta|^2}}. \tag{10}$$

The stability of amplification which is proportional to the gain stability is given by:

$$\left[\frac{\delta|A_F|}{|A_F|}\right]_{|\mu|} \doteq \frac{1 - |\mu\beta| \cos \Phi}{|1 - \mu\beta|^2} \left[\frac{\delta|\mu|}{|\mu|}\right], \tag{11}$$

$$\left[\frac{\delta|A_F|}{|A_F|}\right]_{|\beta|} \doteq \left|\frac{\mu\beta}{1 - \mu\beta}\right| \left[\frac{\cos \Phi - |\mu\beta|}{|1 - \mu\beta|}\right] \left[\frac{\delta|\beta|}{|\beta|}\right], \tag{12}$$

$$\left[\frac{\delta|A_F|}{|A_F|}\right]_{\Phi} \doteq - \left|\frac{\mu\beta}{1 - \mu\beta}\right| \left[\frac{\sin \Phi}{|1 - \mu\beta|}\right] [\delta\Phi]. \tag{13}$$

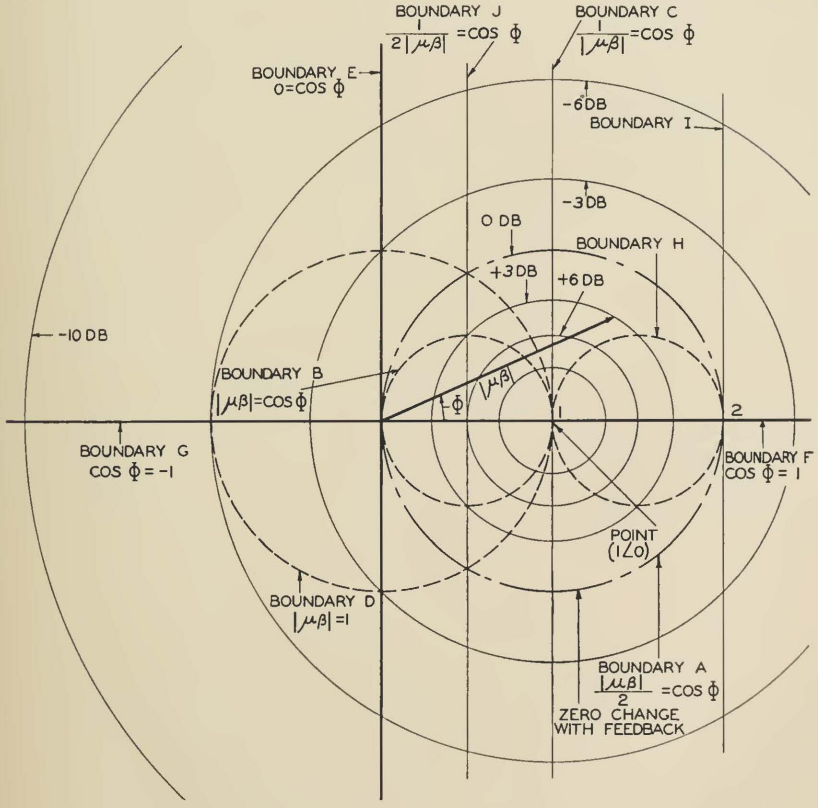


Fig. 3—The vector field of $\mu\beta$. See caption for Fig. 4.

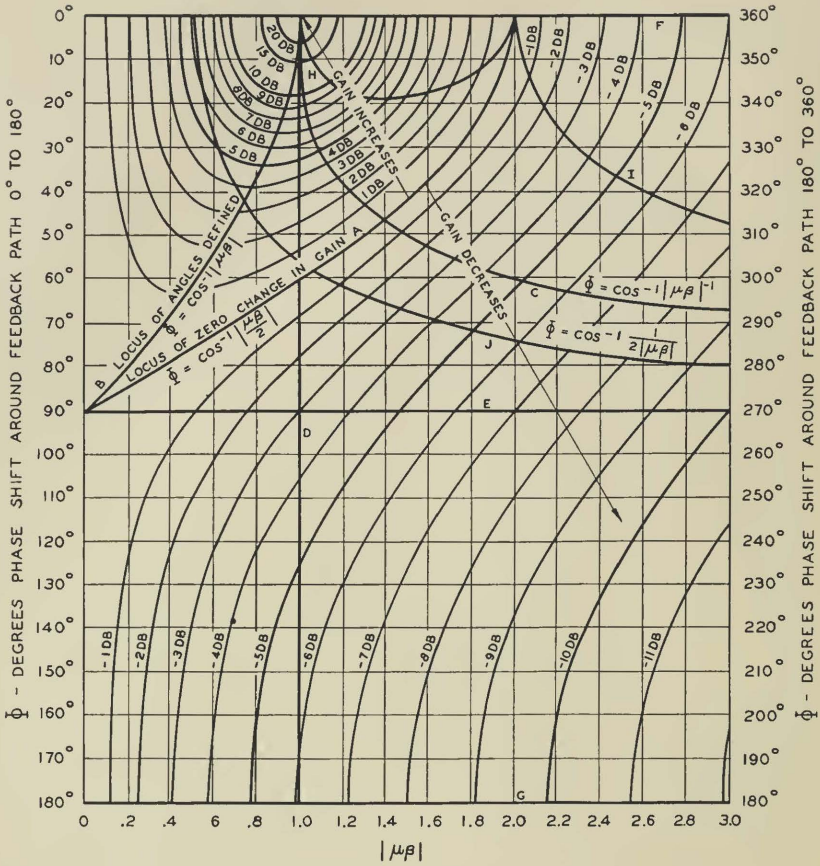


Fig. 4—Phase shift around the feedback path plotted as a function of $|\mu\beta|$, the absolute value of $\mu\beta$.

$\mu\beta$ is a complex quantity which represents the ratio by which the amplifier and feedback (or more generally μ and β) modify a voltage in a single trip around the closed path.

First, there is a set of boundary curves indicated as A, B, C, D, E, F, G, H, I, and J which gives either limiting or significent values of $|\mu\beta|$ and Φ.

Secondly, there is a family of curves in which db change in gain due to feedback is the parameter.

Boundaries

A. Conditions in which gain and modulation are unaffected by feedback.

B. Constant amplification ratio against small variations in $|\beta|$.

Constant change in gain, $\frac{1}{|1-\mu\beta|}$, against variations in $|\mu|$ and $|\beta|$.

Stable phase shift through the amplifier against variations in Φ_β.

The boundary on which the stability of amplification is unaffected by feedback.

C. Constant amplification ratio against small variations in $|\mu|$.

Constant phase shift through the amplifier against variations in Φ_μ.

The absolute magnitude of the voltage fed back $\frac{|\mu\beta|}{|1-\mu\beta|}$ is constant against variations in $|\mu|$ and $|\beta|$.

A curious fact to be noted from (11) is that it is possible to choose a value of $\mu\beta$ (namely, $|\mu\beta| = \sec \Phi$) so that the numerator of the right hand side vanishes. This means that the gain stability is perfect, assuming differential variations in $|\mu|$. Referring to Figs. 3 and 4, contour C is the locus of $|\mu\beta| = \sec \Phi$ and it includes all amplifiers whose gain is unaffected by small variations in $|\mu|$. In this way it is even possible to stabilize an amplifier whose feedback is positive, i.e., feedback may be utilized to raise the gain of an amplifier and, at the same time, the gain stability with feedback *need* not be degraded but on the contrary improved. If a similar procedure is followed with an amplifier whose feedback is negative, the gain stability will be theoretically perfect and independent of the reductions in gain due to feedback. Over too wide a frequency band practical difficulties will limit the improvements possible by these methods.

With negative feedback, gain stability is always improved by an amount at least as great as corresponds to the reduction in gain and generally more; with positive feedback, gain stability is never degraded by more than would correspond to the increase in gain and under appropriate conditions, assuming the variations are not too great, is as good as or much better than without feedback. With positive feedback, the variations in μ or β must not be permitted to become sufficiently great to cause the amplifier to sing or give rise to instability as defined in a following section on "Avoiding Singing."

Modulation

To determine the effect of feedback action upon modulation produced in the amplifier circuit, it is convenient to assume that the output of undistorted signal is made the same with and without feedback and that a comparison is then made of the difference in modulation with and without feedback. Therefore, with feedback, the input is changed to $e = e_0(1 - \mu\beta)$ and, referring to equation (2), the output voltage is μe_0, and the generated modulation, $d(E)$, assumes its value without feedback, $d(E_0)$, and $d(E)/(1-\mu\beta)$ becomes $d(E_0)/(1-\mu\beta)$ which is $D_0/(1 - \mu\beta)$. This relationship is approximate because the

---

D. $|\mu\beta| = 1$.

E. $\Phi = 90°$. Improvement in gain stability corresponds to twice db reduction in gain.

F and *G*. Constant amplification ratio against variations in Φ.
Constant phase shift through the amplifier against variations in $|\mu|$ and $|\beta|$.

H. Same properties as *B*.

I. Same properties as *E*.

J. Conditions in which $\frac{|\mu|}{|1 - \mu\beta|} = \frac{-1}{|\beta|}$ the overall gain is the exact negative inverse of the transmission through the β-circuit.

voltage at the input without feedback is free from distortion and with feedback it is not and, hence, the assumption that the generated modulation is a function *only of the signal output* used in deriving equation (2) is not necessarily justified.

From the relationship $D = D_0/(1 - \mu\beta)$, it is to be concluded that modulation with feedback will be reduced db for db as the effect of feedback action causes an arbitrary db reduction in the gain of the amplifier, i.e., when the feedback is negative. With positive feedback the opposite is true, the modulation being increased by an amount corresponding to the increase in amplification.

If modulation in the β-circuit is a factor, it can be shown that usually in its effect on the output, the modulation level at the output due to non-linearity of the β-circuit is approximately $\mu\beta/(1 - \mu\beta)$ multiplied by the modulation generated in the β-circuit acting alone and without feedback.

Additional Effects

Noise

A criterion of the worth of a reduction in noise is the reduction in signal-to-noise ratio at the output of an amplifier. Assuming that the amount of noise introduced is the same in two systems, for example with and without feedback respectively, and that the signal outputs are the same, a comparison of the signal-to-noise ratios will be affected by the amplification between the place at which the noise enters and the output. Denoting this amplification by a and a_0 respectively, it can be shown that the relation between the two noise ratios is $(a_0/a)(1 - \mu\beta)$. This is called the *noise index.*

If noise is introduced in the power supply circuits of the last tube, $a_0/a = 1$ and the noise index is $(1 - \mu\beta)$. As a result of this relation less expensive power supply filters are possible in the last stage.

Phase Shift, Envelope Delay, Delay Distortion

In the expression $A_F = [\mu/(1 - \mu\beta)] \angle \theta$, θ is the overall phase shift with feedback, and it can be shown that the *phase shift through the amplifier with feedback may be made to approach the phase shift through the β-circuit plus 180 degrees.* The effect of phase shift in the β-circuit is not correspondingly reduced. It will be recalled that in reducing the change in phase shift with frequency, envelope delay, which is the slope of the phase shift with respect to the angular velocity, $\omega = 2\pi f$, also is reduced. The *delay* distortion likewise is reduced because a measure of delay distortion at a particular frequency is the difference between the envelope delay at that frequency and the least envelope delay in the band.

β-Circuit Equalization

Referring to equation (2), the output voltage, E, approaches $-e_0/\beta$ as $1 - \mu\beta \doteq -\mu\beta$ and equals it in absolute value if $\cos\Phi = \dfrac{1}{2|\mu\beta|}$ where $\mu\beta = |\mu\beta| \angle\Phi$. Under these circumstances increasing the loss in the β-circuit one db raises the gain of the amplifier one db and vice-versa, thus giving any gain-frequency characteristic for which a like loss-frequency characteristic can be inserted in the β-circuit. This procedure has been termed β-circuit equalization. It possesses other advantages which cannot be dwelt upon here.

Avoid Singing

Having considered the theory up to this point, experimental evidence was readily acquired to demonstrate that $\mu\beta$ might assume large values,

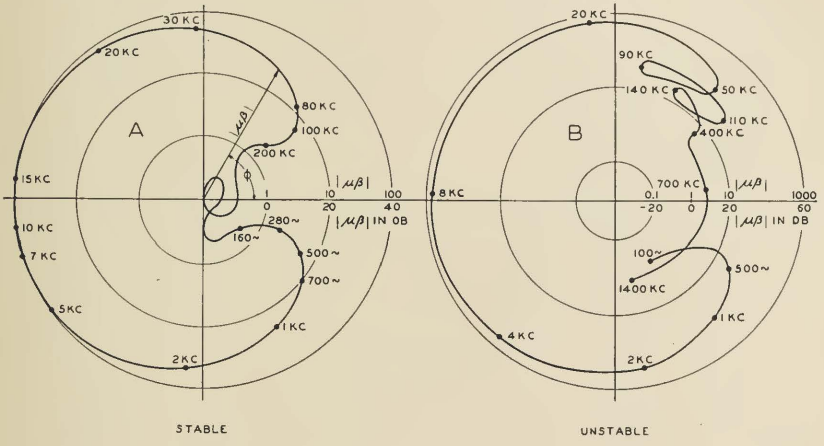


Fig. 5—Measured $\mu\beta$ characteristics of two amplifiers.

for example 10 or 10,000, provided Φ was not at the same time zero. However, one noticeable feature about the field of $\mu\beta$ (Figs. 3 and 4) is that it implies that even though the phase shift is zero and the absolute value of $\mu\beta$ exceeds unity, self-oscillations or singing will not result. This may or may not be true. When the author first thought about this matter he suspected that owing to practical non-linearity, singing would result whenever the gain around the closed loop equalled or exceeded the loss and simultaneously the phase shift was zero, i.e., $\mu\beta = |\mu\beta| + jo \geq 1$. Results of experiments, however, seemed to indicate something more was involved and these matters were described to Mr. H. Nyquist, who developed a more general criterion

for freedom from instability [3] applicable to an amplifier having linear positive constants.

To use this criterion, plot $\mu\beta$ (the modulus and argument vary with frequency) and its complex conjugate in polar coordinates for all values of frequency from 0 to $+\infty$. If the resulting loop or loops do not enclose the point (1, 0) the system will be stable, otherwise not.[3] The envelope of the transient response of a stable amplifier

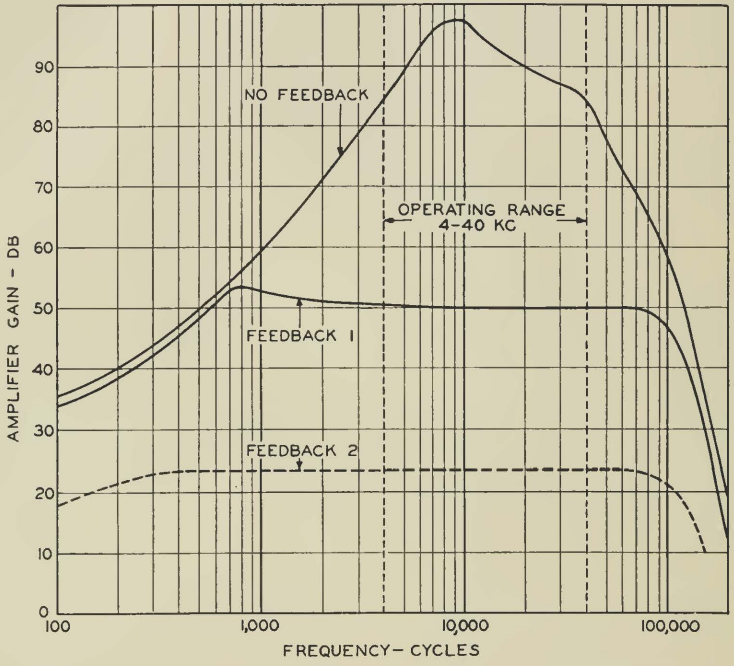


Fig. 6—Gain frequency characteristics with and without feedback of amplifier of Fig. 2.

always dies away exponentially with time; that of an unstable amplifier in all physically realizable cases increases with time. Characteristics A and B in Fig. 5 are results of measurements on two different amplifiers; the amplifier having $\mu\beta$-characteristic denoted A was stable; the other unstable.

The number of stages of amplification that can be used in a single amplifier is not significant except insofar as it affects the question of avoiding singing. Amplifiers with considerable negative feedback

[3] For a complete description of the criterion for stability and instability and exactly what is meant by enclosing the point (1, 0), reference should be made to "Regeneration Theory"—H. Nyquist, *Bell System Technical Journal*, Vol. XI, pp. 126–147, July, 1932.

have been tested where the number of stages ranged from one to five inclusive. In every case the feedback path was from the output of the last tube to the input of the first tube.

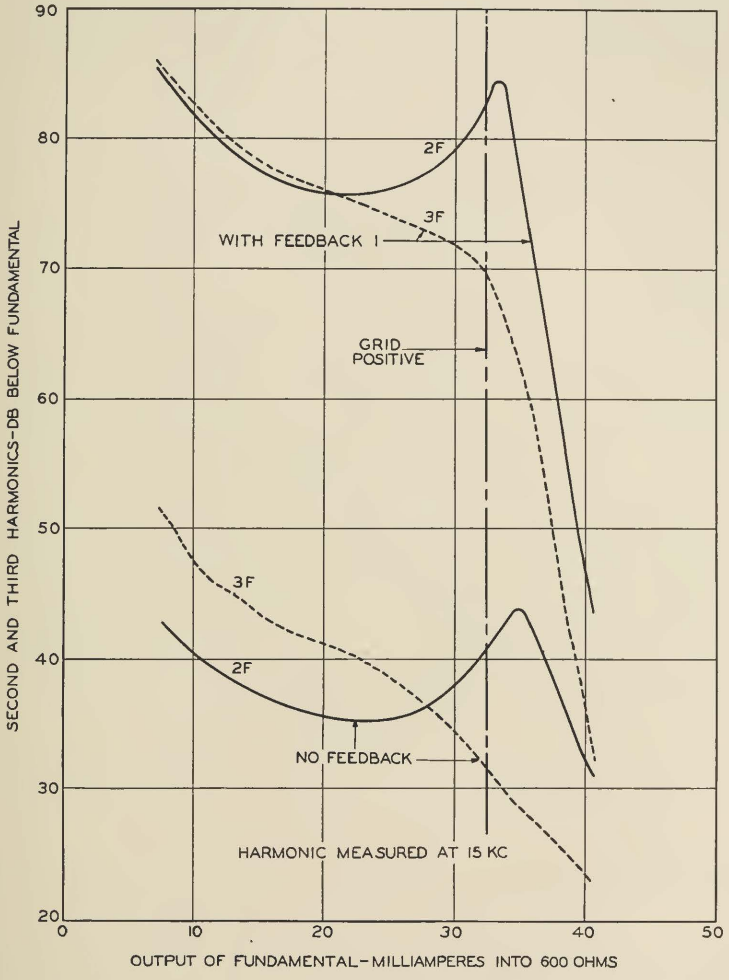


Fig. 7—Modulation characteristics with and without feedback for the amplifier of Fig. 2.

Experimental Results

Figures 6 and 7 show how the gain-frequency and modulation characteristics of the three-stage impedance coupled amplifier of Fig. 2 are improved by negative feedback. In Fig. 7, the improvement in harmonics is not exactly equal to the db reduction in gain. Figure 8

shows measurements on a different amplifier in which harmonics are reduced as negative feedback is increased, db for db over a 65 db range.

That the gain with frequency is practically independent of small variations in $|\mu|$ is shown by Fig. 9. This is a characteristic of the Morristown amplifier described in the paper by Messrs. Clark and Kendall[1] which meets the severe requirements imposed upon a repeater amplifier for use in cable carrier systems. Designed to amplify frequencies from 4 kc

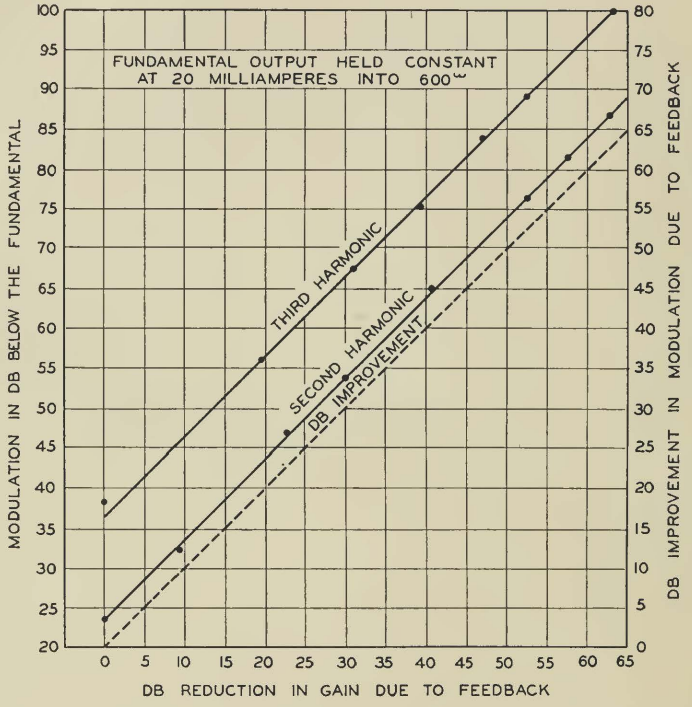


Fig. 8—Improvement of harmonics with feedback. One example of another amplifier in which with 60 db feedback, harmonic currents in the output are only one-thousandth and their energy one-millionth of the values without feedback.

to 40 kc the maximum change in gain due to variations in plate voltage does not exceed 7/10000 db per volt and at 20 kc the change is only 1/20000 db per volt. This illustrates that for small changes in $|\mu|$, the ratio of the stability without feedback to the stability with feedback, called the *stability index*, approaches $|1 - \mu\beta|^2/(1 - |\mu\beta| \cos \Phi)$ and gain stability is improved at least as much as the gain is reduced and usually more and is theoretically perfect if $\cos \Phi = 1/|\mu\beta|$.

[1] Loc. cit.

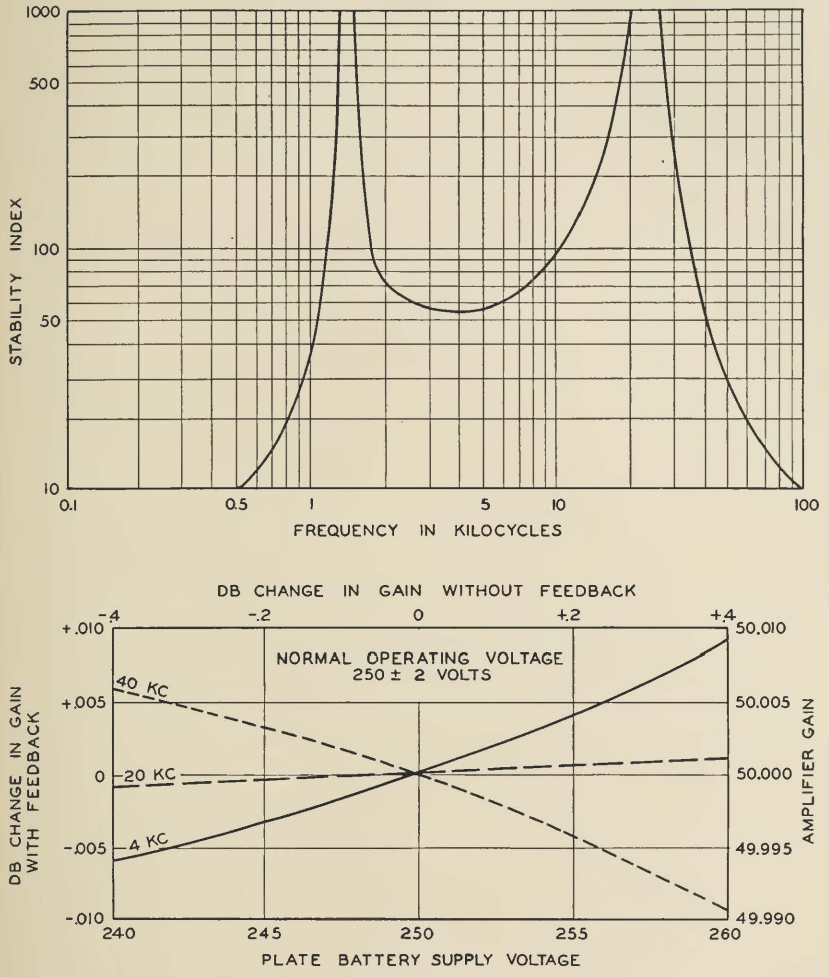


Fig. 9—Representative gain stability of a single amplifier as determined by measuring 69 feedback amplifiers in tandem at Morristown, N. J.

The upper figure shows the absolute value of the stability index. It can be seen that between 20 and 25 kc the improvement in stability is more than 1000 to 1 yet the reduction in gain was less than 35 db.

The lower figure shows change in gain of the feedback amplifier with changes in the plate battery voltage and the corresponding changes in gain without feedback. At some frequencies the change in gain is of the same sign as without feedback and at others it is of opposite sign and it can be seen that near 23 kc the stability must be perfect.

Figure 10 indicates the effectiveness with which the gain of a feedback amplifier can be made independent of variations in input amplitude up to practically the overload point of the amplifier. These measurements were made on a three-stage amplifier designed to work from 3.3 kc to 50 kc.

Figure 11 shows that negative feedback may be used to improve phase shift and reduce delay and delay distortion. These measurements

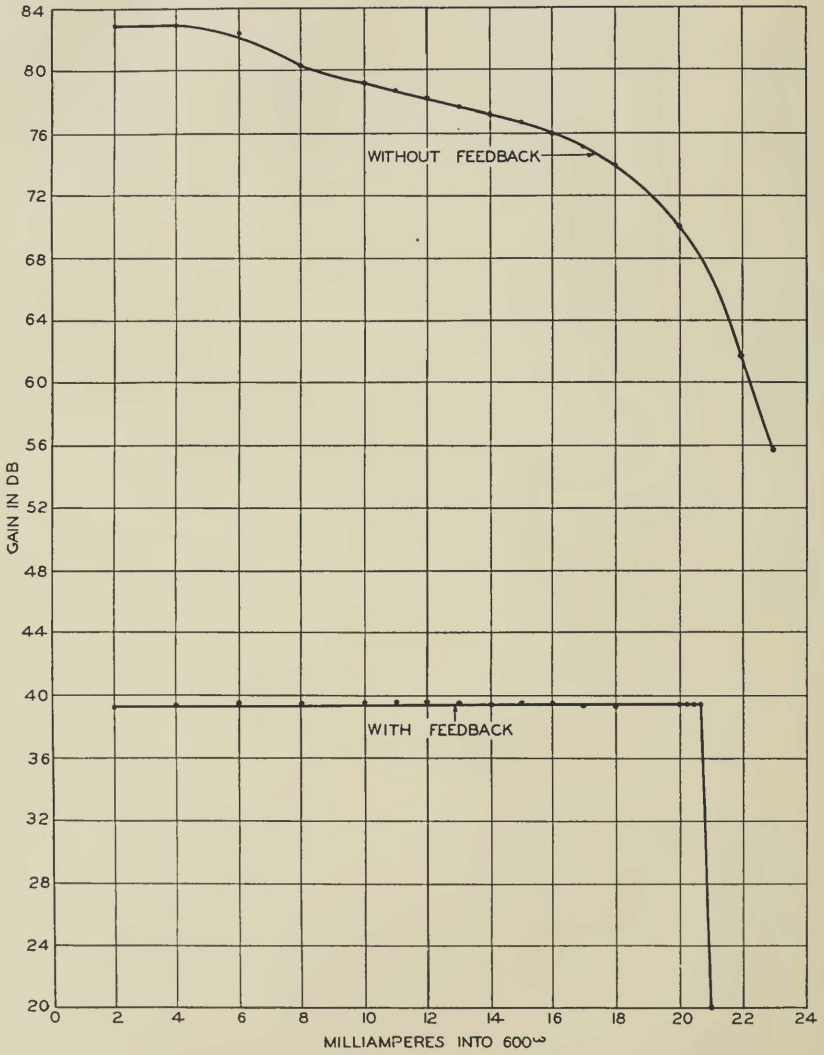


Fig. 10—Gain-load characteristic with and without feedback for a low level amplifier designed to amplify frequencies from 3.5 to 50 kc.

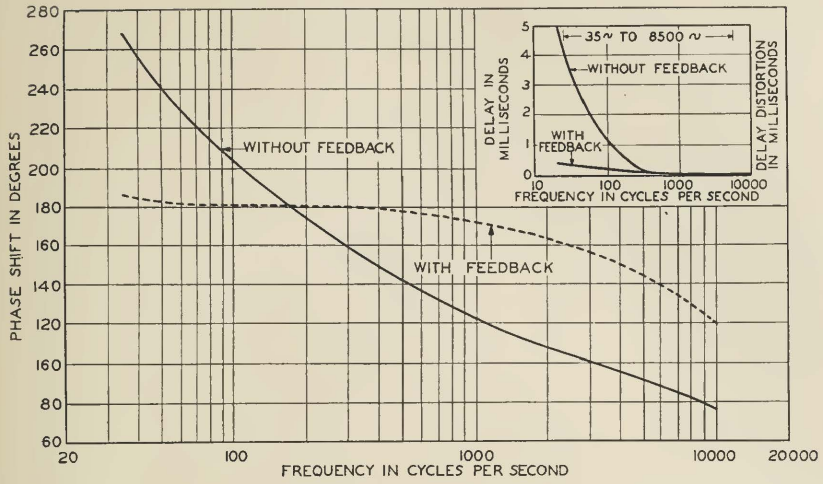


Fig. 11—Phase shift, delay, and delay distortion with and without feedback for a single tube voice frequency amplifier.

were made on an experimental one-tube amplifier, 35–8500 cycles, feeding back around the low side windings of the input and output transformers.

Figure 12 gives the gain-frequency characteristic of an amplifier with and without feedback when in the β-circuit there was an equalizer

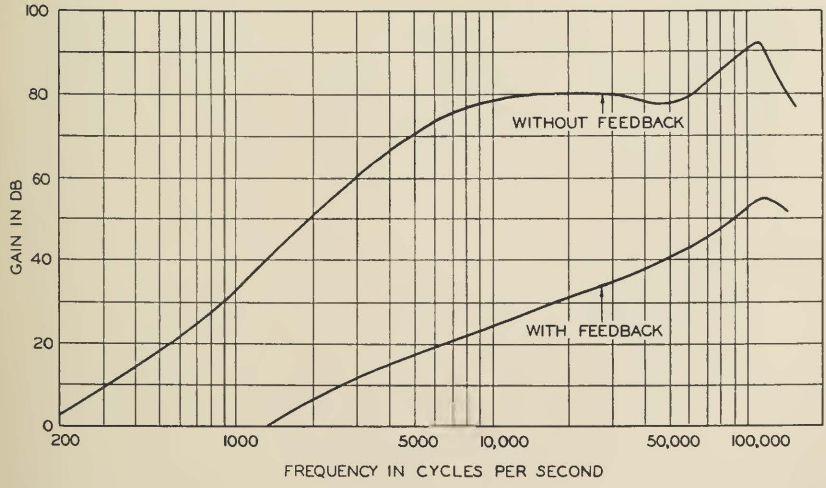


Fig. 12—Gain-frequency characteristic of an amplifier with an equalizer in the β-circuit. This was designed to have a gain frequency characteristic with feedback of the same shape as the loss frequency characteristic of a non-loaded telephone cable.

designed to make the gain-frequency characteristic of the amplifier with feedback of the same shape as the loss-frequency characteristic of a non-loaded telephone cable.

Conclusion

The feedback amplifier dealt with in this paper was developed primarily with requirements in mind for a cable carrier telephone system, involving many amplifiers in tandem with many telephone channels passing through each amplifier. Most of the examples of feedback amplifier performance have naturally been drawn from amplifiers designed for this field of operation. In this field, vacuum tube amplifiers normally possessing good characteristics with respect to stability and freedom from distortion are made to possess superlatively good characteristics by application of the feedback principle.

However, certain types of amplifiers in which economy has been secured by sacrificing performance characteristics, particularly as regards distortion, can be made to possess improved characteristics by the application of feedback. Discussion of these amplifiers is beyond the scope of this paper.

Open-Wire Crosstalk *

By A. G. CHAPMAN

Introduction

THE tendency of communication circuits to crosstalk from one to another was greatly increased by the advent of telephone repeaters and carrier current methods. Telephone repeaters multiplied circuit lengths many times, increased the power applied to the wires, and at the same time made the circuits much more efficient in transmitting crosstalk currents as well as the wanted currents. Carrier current methods added higher ranges of frequency with consequently increased crosstalk coupling. Program transmission service added to the difficulties since circuits for transmitting programs to broadcasting stations must accommodate frequency and volume ranges greater than those required for message telephone circuits.

As these new types of circuits were developed, their application to existing open-wire lines was attended with considerable difficulty from the crosstalk standpoint. Severe restrictions had to be placed on the allocation of pairs of wires for different services in order to keep the crosstalk within tolerable bounds. In many cases the existing lines were retransposed but, nevertheless, there were still important restrictions. While great reduction in crosstalk was obtained by the transposition arrangements the crosstalk reduction was finally limited by unavoidable irregularities in the spacing of the transposition poles and in the spacing of the wires, including differences in wire sag. To further improve matters it was, therefore, necessary to alter the wire configurations so as to reduce the coupling per unit length between the various circuits.

Recently this study of wire configurations has resulted in extensive use of new configurations of open-wire lines in which the two wires of a pair are placed eight inches apart instead of 12 inches, the horizontal separation between wires of different pairs being correspondingly increased. With these eight-inch pairs it has usually been found desirable to discard the time-honored phantoming method of obtaining

* This paper gives a comprehensive discussion of the fundamental principles of crosstalk between open-wire circuits and their application to the transposition design theory and technique which have been developed over a period of years. In this issue of the *Technical Journal* the first half of the paper is published. In the April 1934 issue will be the concluding part, together with an appendix entitled "Calculation of Crosstalk Coefficients."

additional circuits so as to make it possible to obtain a greater number of circuits by more intensive application of carrier current methods.

It is the object of this paper to outline the fundamental principles concerning crosstalk between open-wire circuits and recent developments in transposition design theory and technique which have led to the latest pole line configurations and transposition designs.

To those generally interested in electrical matters it is hoped that this paper will give an insight into the problem of keeping crosstalk in open-wire lines within proper bounds. To those interested in crosstalk it is hoped that the paper will give a useful review of the whole matter and perhaps an insight into the importance of some phenomena which do not seem to be generally appreciated.

The principles set forth in this paper will also be found of considerable interest in connection with problems of control of cable crosstalk, particularly for the high frequencies involved in carrier transmission. It will also be recognized that use is made here of the same general principles as are used in the calculation of effects of impedance irregularities and echoes on repeater operation. These general principles have also been found useful in the development of combinations or arrays of radio antennas of the long horizontal wire type.

The art of crosstalk control in open-wire lines has grown up as a result of the efforts of many workers. The individual contributions are so numerous that it has not been considered practicable in this paper to make individual mention of them except in a few special cases.

General

In the evolution of a satisfactory transposition design technique, complicated electrical actions must be considered and it has been convenient to divide the total crosstalk coupling into various types, all of which may contribute in producing crosstalk between any two circuits in proximity. The first portion of this paper is therefore devoted largely to an examination of the underlying principles and the definition of some of the special terms employed, such as *transverse crosstalk*, *interaction crosstalk*, *reflection crosstalk*, etc. The paper then considers the general effect of transpositions in reducing crosstalk and how this effect depends on the attenuation and phase change accompanying the transmission of communication currents. Consideration is next given to the practical significance of and methods for determining the *crosstalk coefficients* which are used in calculating the crosstalk in a short part of a parallel between two currents. The matter of *type unbalances* inherent in different arrangements of trans-

positions and used in working from short lengths to long lengths is discussed at length. The next section of the paper is devoted to the effect of constructional irregularities caused by pole spacing, wire sag, "drop bracket" transpositions, etc. Various "non-inductive" wire arrangements are considered. The paper closes with a general discussion of practical transposition design methods based on the principles previously disclosed.

Underlying Principles

The discussion under this heading will cover the general causes of crosstalk coupling between open-wire circuits and the general types into which it is convenient to divide the crosstalk effect. The usual measures of crosstalk coupling will also be discussed.

Causes and Types of Crosstalk

The crosstalk coupling between open-wire pairs is due almost entirely to the external electric and magnetic fields of the disturbing circuit. If these fields were in some way annulled there would remain the possibility of resistance coupling between the pairs because of leakage from one circuit to the other by way of the crossarms and insulators, tree branches, etc. This leakage effect is minor in a well-maintained line. It enters as a factor in the design of open-wire transpositions only in so far as the attenuation of the circuits is affected which indirectly affects the crosstalk.

Figure 1 indicates cross-sections of two pairs of wires designated as 1–2 and 3–4. If pair 1–2 existed alone and if the two wires were similar, a voltage impressed at one end of the circuit would result in

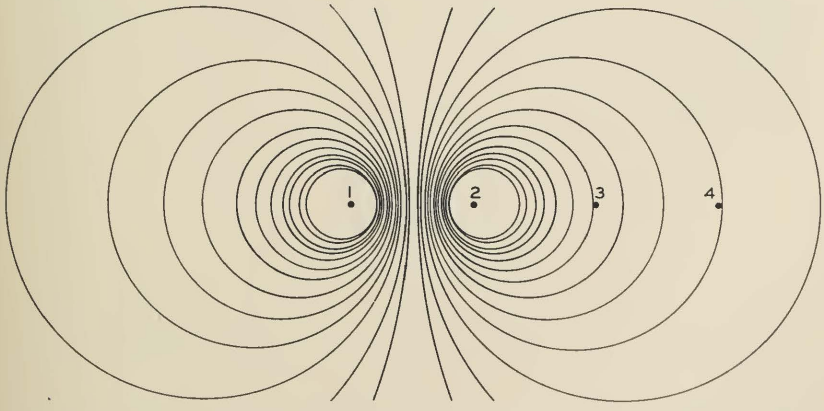


Fig. 1—Magnetic field produced by equal and opposite currents in wires 1 and 2.

equal and opposite currents at any point. These currents would produce a magnetic field as indicated on the figure. If circuit 3–4 parallels 1–2 a certain amount of this magnetic flux would thread between wires 3 and 4 and induce a voltage in circuit 3–4 which would result in a crosstalk current in this circuit. This induced voltage is, of course, due to the difference between the two magnetic fields set up by the opposite directional currents in wires 1 and 2. Since wires 1 and 2 are not very far apart, the resultant field is much weaker than if transmission over wire 1 with ground return were attempted. It is important, therefore, that the wires of a circuit be placed as close together as practicable and that these wires be similar in material and gauge in order to keep the currents practically equal and opposite.

Equal and opposite charges accompany the equal and opposite currents in wires 1 and 2. The equipotential lines of the resultant electric field set up by the two charges are also indicated by Fig. 1. This field will cause different potentials at the surfaces of wires 3 and 4 and this potential difference will cause a crosstalk current in circuit 3–4. As in the case of magnetic induction this current may be minimized by close spacing and electrical similarity between the two wires of a pair.

Calculations of crosstalk coupling must, in general, consider both the electric and magnetic components of the electromagnetic field of the disturbing circuit.

The exact computation of crosstalk coupling between communication circuits is very complex.[1] Approximate computations are sufficient for transposition design. In such computations, it is convenient to divide the total coupling into components of several general types. In calculations of coupling of these types it is assumed that the two wires of a circuit are similar in material and gauge. If there is any slight dissimilarity, such as extra resistance in one wire due to a poor joint, the effect on the crosstalk may be computed separately. The general types of crosstalk coupling are:

1. Transverse crosstalk coupling.
   1*a*. Direct.
   1*b*. Indirect.
2. Interaction crosstalk coupling.

A multi-wire pole line involves many circuits all mutually coupled. In explaining the above terms, it is convenient to start with the simple conception of but two paralleling coupled circuits; Fig. 2A

[1] The general mathematical theory is given in the Carson-Hoyt paper listed under "Bibliography."

indicates such a parallel. In calculating the crosstalk coupling between a terminal of circuit a and a terminal of circuit b, the parallel may be divided into a series of thin transverse slices. One such slice of thickness d is indicated on the figure. The coupling in each slice is calculated and, then, the total coupling between circuit terminals due to all the slices.

In Fig. 2A circuit a is considered to be the disturber and to be energized at the left-hand end. In the single slice indicated, a transmission current will be propagated along circuit a and will cause crosstalk currents in circuit b at both ends of the slice. In this slice, therefore, the left-hand end of circuit a may be considered to be coupled to the two ends of circuit b through the transmission paths n_{ab} and f_{ab}. The path n_{ab} is called the near-end crosstalk coupling and the path f_{ab} is called the far-end crosstalk coupling.

The presence of a tertiary circuit, such as c of Fig. 2B, changes both the near-end and the far-end coupling between a and b in the transverse slice. In addition to the direct couplings n_{ab} and f_{ab} there are indirect couplings n_{acb} and f_{acb} by way of circuit c.

The *transverse crosstalk coupling* between a terminal of a disturbing circuit and a terminal of a disturbed circuit is defined as the coupling between these points due to all the small couplings in all the thin transverse slices including indirect couplings in each slice by way of other circuits. (There are also indirect couplings involving more than one slice and these are not included in the transverse crosstalk coupling.)

In computations of transverse crosstalk coupling it is convenient to distinguish between the direct and indirect components. The direct component considers only the currents and charges in the disturbing circuit while the indirect component takes account of certain charges in tertiary circuits resulting from transmission over the disturbing circuit. The tertiary circuits may be circuits used for transmission purposes or any other circuits which can be made up of combinations of wires on the line or of these wires and ground. If there are only two pairs on the line as in Fig. 2A there are still tertiary circuits, namely, the "phantom" circuit consisting of pair a as one side of the circuit and pair b as the return and the "ghost" circuit consisting of all four wires with ground return. In a multi-wire line many of the tertiary circuits involve the wires of the disturbing circuit. If these tertiary circuits did not exist the currents at any point in the two wires of the disturbing circuit would be equal and opposite. The presence of the tertiary circuits makes these currents unequal and it is convenient to divide the actual currents into two components, i.e.,

equal and opposite or "balanced" currents in the two wires of the disturbing circuit and equal currents in phase in the two wires. The latter may be called "tertiary circuit" currents. The charges on the two wires of the disturbing circuit may be similarly divided into components.

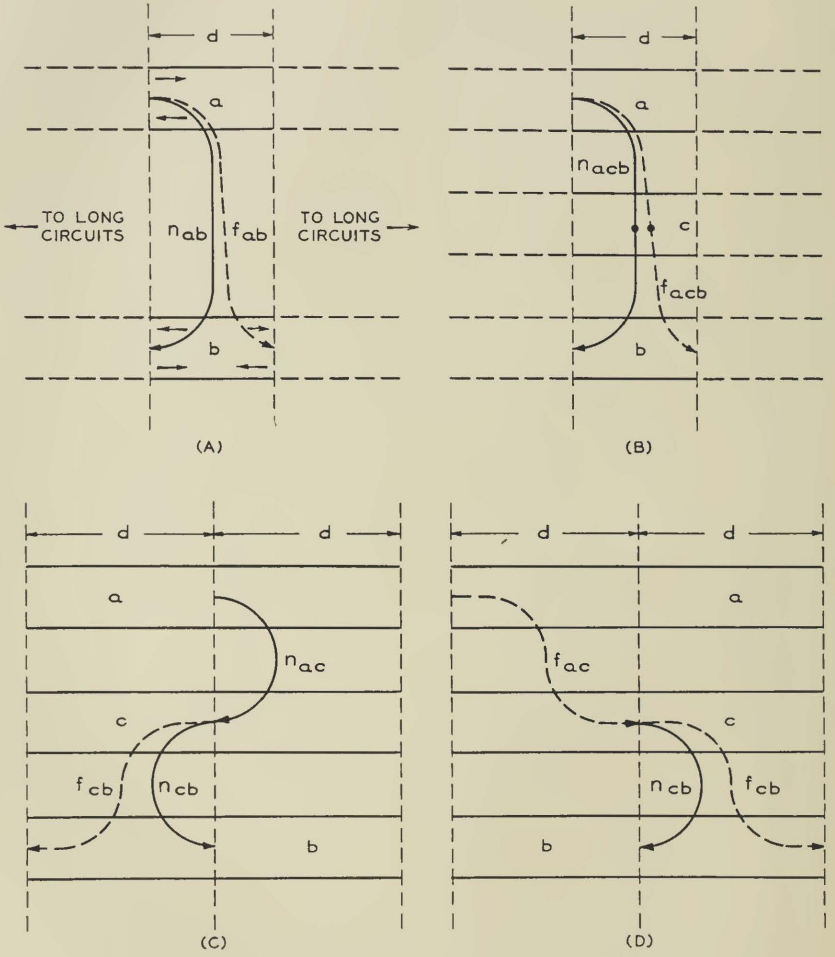


Fig. 2—Transverse and interaction crosstalk.

The *direct* component of the transverse crosstalk coupling is defined as that part which is due to balanced charges and currents in the disturbing circuit.[2] The *indirect* component is defined as that part

[2] "Direct" is here used in a different sense from that used in connection with the term "direct capacity unbalance" which was originated by Dr. G. A. Campbell and has been much used in discussions of cable crosstalk.

which is due to *charges* on tertiary circuits which arise within any thin transverse slice due to coupling with the disturbing circuit in that same slice. This coupling, in any slice, causes currents as well as charges in the tertiary circuit in that slice, but, as discussed in detail in Appendix A, the effect of these currents in producing crosstalk currents in the disturbed circuit is small compared with the effect of the charges. The currents and charges in the tertiary circuits in any thin slice due to the coupling with the disturbing circuit in that same slice may be but a small part of the total currents and charges in the slice. The total values are due to couplings of the tertiary circuits with the disturbing circuit in all the slices. When the total values are considered currents as well as charges in the tertiary circuit may be important in causing crosstalk currents in the disturbed circuit. To consider the total currents and charges in the tertiary circuits it is necessary to take account of both the interaction crosstalk coupling and the transverse crosstalk coupling between disturbing and disturbed circuits.

The nature of interaction crosstalk coupling is indicated by Figs. 2C and 2D which indicate two successive thin transverse slices of width d in a parallel between two circuits a and b and the typical tertiary circuit c. Assuming transmission from left to right on circuit a in Fig. 2C this circuit is coupled with c in the right-hand slice by the near-end crosstalk coupling indicated by n_{ac}. This coupling causes transmission of crosstalk current (and charge) into the left-hand part of circuit c which has both near-end and far-end crosstalk coupling to circuit b. Consideration of these two successive transverse slices, therefore, introduces the two compound couplings $n_{ac}n_{cb}$ and $n_{ac}f_{cb}$. There are two more of these compound couplings as indicated by Fig. 2D. There is a far-end crosstalk coupling between circuits a and c in the left-hand slice which combines with both near-end and far-end couplings in the right-hand slice. The compound types of crosstalk of Fig. 2C and 2D are called interaction crosstalk since the various slices interact on each other in producing indirect couplings. The *interaction crosstalk coupling* between a terminal of a disturbing circuit and a terminal of a disturbed circuit is defined as the coupling between these points due to the indirect couplings involving *all possible combinations* of different thin transverse slices.

The distinction between indirect transverse crosstalk and interaction crosstalk is that the former takes account of the effect of indirect crosstalk from disturbing to tertiary to disturbed circuit in a single thin transverse slice while the latter involves indirect crosstalk from primary circuit to tertiary circuit in one slice, transmission along the

tertiary circuit into another slice and then crosstalk from tertiary circuit to disturbed circuit.

The notion that there is only transverse crosstalk within any one "thin slice" implies that the slice thickness corresponds to a distance along the line of only infinitesimal length. If this distance were finite it would correspond to a series of "thin slices" having interaction crosstalk between them. Practically, however, if the distance along the line corresponds to a line angle of five degrees or less, the interaction crosstalk in this length is small compared with the transverse crosstalk. A five degree line angle corresponds to a length of about .1 mile at 25 kilocycles, .05 mile at 50 kilocycles, etc. A transposed line is divided into short lengths or *segments* by the transposition poles and the line angle of these segments is ordinarily less than five degrees at the highest frequency for which the transposition system is suitable. Therefore, the crosstalk coupling between such transposed circuits may be computed on the basis of transverse crosstalk within any segment and interaction crosstalk between any two segments.

As shown by Fig. 2 the interaction effect involves the four compound couplings:

$$n_{ac}n_{cb}, \qquad n_{ac}f_{cb}, \qquad f_{ac}n_{cb}, \qquad f_{ac}f_{cb}.$$

The near-end crosstalk couplings n_{ac} and n_{cb} of Fig. 2 are usually much larger than the far-end couplings f_{ac} and f_{cb}. The reason for this, as discussed in Appendix A, is that the electric and magnetic fields of the disturbing circuit tend to aid each other in producing near-end crosstalk coupling such as n_{ac}, and to oppose each other in the case of far-end coupling such as f_{ac}. For this reason the compound coupling $n_{ac}n_{cb}$ is the most important and is usually the only compound coupling which requires consideration in transposition design. Since the path $n_{ac}n_{cb}$ results in a crosstalk current at the far end of the disturbed circuit, it is in connection with far-end crosstalk between long circuits that this matter of interaction crosstalk is important. Far-end rather than near-end crosstalk coupling is controlling in connection with open-wire carrier frequency systems for the reasons explained below.

Figure 3A indicates very schematically two one-way carrier frequency channels routed over two long paralleling open-wire pairs. The boxes at the end indicate the repeaters or terminal apparatus and the arrows on these boxes the direction of transmission of this apparatus. Transmission from the left on pair a results in near-end and far-end crosstalk into pair b, as indicated by the couplings n_{ab} and f_{ab}. The near-end crosstalk current cannot pass to the input of the terminal

apparatus since the latter is a one-way device. In practice, to obtain two-way circuits each of these one-way channels is associated with another one-way channel transmitting in the opposite direction over the same pair of wires. These return channels utilize a different band of carrier frequencies and the near-end crosstalk current is largely excluded from this frequency band by selective filters. The far-end crosstalk is, therefore, the sole consideration with such a carrier system. Use is not made of the same carrier frequencies in both directions on a toll line largely because of difficulties in controlling the near-end crosstalk.

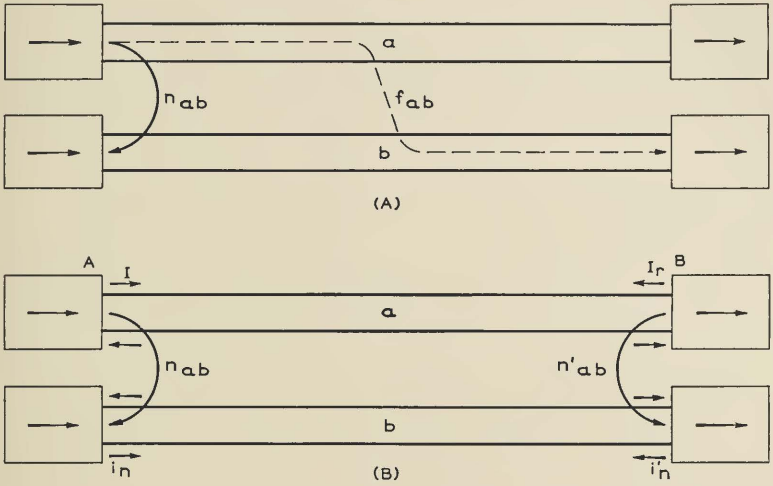


Fig. 3—Crosstalk between two one-way carrier frequency channels.

In connection with the arrangement of Fig. 3A, there is a type of crosstalk of considerable practical importance known as "*reflection crosstalk.*" The theory of this is indicated by Fig. 3B which shows the same two one-way carrier channels. Transmission from left to right on circuit a is assumed. When the transmission current I arrives at point B, a certain portion of it will be reflected if there is any deviation of the input impedance of the terminal apparatus from the characteristic impedance of circuit a. This reflected current I_r causes a near-end crosstalk current i_n' at point B in the disturbed circuit. Similarly, a part of the near-end crosstalk current i_n at point A in the disturbed circuit may be reflected and transmitted to point B. Therefore, two additional crosstalk currents may result from these two reflections and such currents can enter the terminal apparatus at B and pass through to the output of this apparatus.

For like circuits, like impedance mismatches and like near-end crosstalk couplings at the two ends of the line, these two additional far-end crosstalk currents are of equal importance. Similar reflection effects will occur at any intermediate points in the lines having impedance irregularities. Since the far-end crosstalk coupling can be much more readily reduced by transpositions than the near-end crosstalk coupling this reflection crosstalk effect is important in practice. It is, therefore, necessary to carefully design the terminal and intermediate apparatus and cables to minimize impedance mismatches as far as practicable.

In calculation of crosstalk coupling it is ordinarily assumed that the two wires of a circuit are electrically similar or "balanced" (except as regards crosstalk from other wires). This is substantially true in practice except for accidental deviations, such as resistance differences due to poor joints and leakage differences due to cracked insulators, foliage, etc. Resistance differences may be of considerable practical importance and are said to cause *resistance unbalance crosstalk*. The following discussion indicates the general nature of this effect.

As discussed in connection with Fig. 1, the external field of the disturbing circuit is minimized by the opposing effects of substantially equal and opposite currents or charges in the two wires of the circuit. The two wires may be considered as two separate circuits, each having its return in the ground. At any point in the line these two wires would normally have practically equal and opposite voltages with respect to ground. These voltages would normally cause almost equal and opposite currents in the two wires. If the resistance of one wire is increased due to a bad joint, the current in that wire is reduced and the currents in the two wires are no longer equal and opposite. The external field of the two wires and the resulting voltage induced in the disturbed circuit are, therefore, altered. If this voltage had previously been practically cancelled out by means of transpositions, the alteration in the field would increase the crosstalk current at the terminal of the disturbed circuit.

A resistance unbalance in the disturbed circuit will have a similar effect as indicated by Fig. 4A. This figure shows a short length d of two long paralleling circuits. Equal and opposite transmission currents in the disturbing circuit 1–2 are indicated by I. Equal crosstalk currents in the two wires of the disturbed circuit 3–4 at one end of the short length are indicated by i. It is assumed that these crosstalk currents have been made substantially equal by transpositions in other parts of the line. Since the currents in wires 3 and 4 are equal and in the same direction, there will be no current in a receiver connected

at the terminal of the line between these wires. If, however, one wire has a bad joint, the two crosstalk currents become unequal and there will be a current in such a receiver.

Resistance unbalance crosstalk is of particular importance if two pairs are used to create a phantom circuit in order to obtain three transmission circuits from the four wires. The distribution of the phantom transmission current I_p in a short length of the two pairs is indicated by Fig. 4B. Ideally, half the phantom current flows in each of the four wires. The two currents in wires 1 and 2 are then equal and in the same direction and there will be no current in terminal apparatus connected between wires 1 and 2. In other words, transmission over the phantom circuit results in no crosstalk in the side circuit 1–2. The same may be said of side circuit 3–4. A bad joint in any wire, such as 3, makes the two currents in wires 3 and 4 unequal and results in a current in the side circuit 3–4.

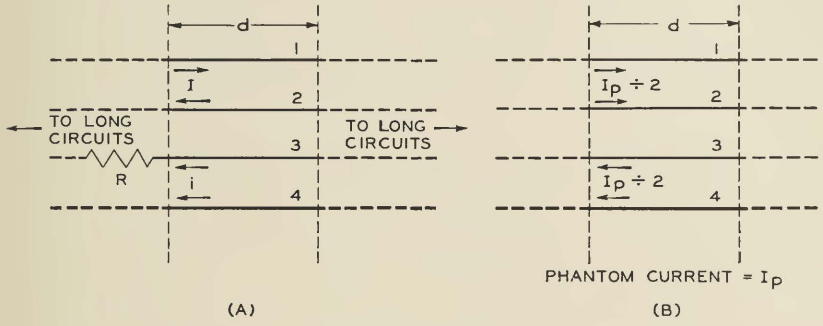


Fig. 4—Effect of resistance unbalance on crosstalk.

The phantom-to-side crosstalk effect of resistance unbalance is much more severe than the effect on crosstalk between two side circuits or two non-phantomed circuits. The reason for this is evident from Figs. 4B and 4A. In Fig. 4B the entire transmission current of the disturbing phantom circuit normally flows in the two wires of the disturbed side circuit and if a resistance unbalance causes a small percentage difference in the currents in these two wires objectionable crosstalk results. In Fig. 4A only crosstalk currents flow in wires 3 and 4 and a much larger percentage difference between these small currents can be tolerated.

In designing and operating phantom circuits, it is necessary to exercise great care to minimize any dissimilarity between the two wires of a side circuit, in order to avoid crosstalk from a phantom to its side circuit or vice versa. Otherwise, the problem of crosstalk between

a phantom circuit and some other circuit is generally similar to the problem of crosstalk between two pairs. In other words, the discussion of transverse, interaction and reflection crosstalk is applicable.

Measures of Crosstalk Coupling

In designing transposition systems, the usual measure of the coupling effect between two open-wire circuits is the ratio of current at the output terminal of the disturbed circuit to current at the input terminal of the disturbing circuit. For circuits of different characteristic impedances this current ratio must be corrected for the difference in impedance. The corrected current ratio is the square root of the corresponding power ratio.

The current ratio is ordinarily very small and for convenience is multiplied by 1,000,000 and called the crosstalk coupling or, in brief, the crosstalk. This usage will be followed from this point in this paper. For example, crosstalk of 1000 units means a current ratio of .001. Crosstalk may also be expressed as the transmission loss in db corresponding to the current ratio. A ratio of .001 means a transmission loss of 60 db corresponding to 1000 crosstalk units.

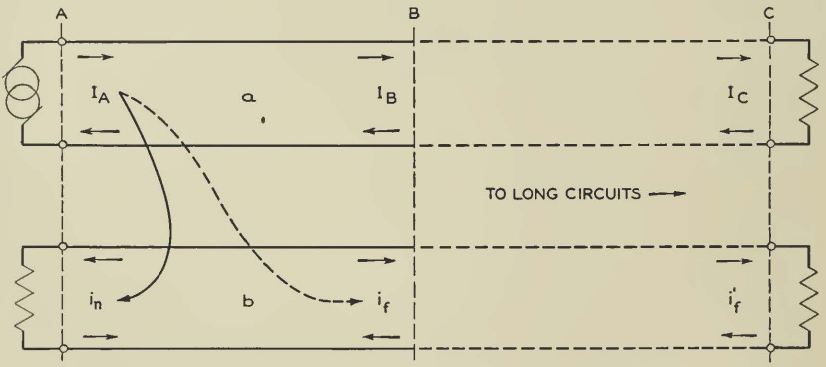


Fig. 5—Schematic of near-end and far-end crosstalk.

Figure 5 indicates two paralleling communication circuits a and b with an e.m.f. impressed at one end of circuit a. The crosstalk currents i_n and i_f in circuit b are due to the crosstalk coupling in length AB. The *near-end crosstalk* in the length AB is the ratio $10^6 i_n/I_A$, while the *far-end crosstalk* is $10^6 i_f/I_A$. The ratio $10^6 i_f/I_B$ has been called the "output-to-output" or "measured" crosstalk. This ratio is a convenient measure of far-end crosstalk between parts of similar circuits because it is related in a simple way to the far-end crosstalk between the terminals of the complete circuits. The following discussion explains this relation.

Both of the currents I_B and i_f will be propagated to point C. They will be attenuated or amplified alike if the circuits are similar and their ratio will be unchanged. The output-to-output crosstalk at C due to the length AB will, therefore, be the same as that determined for point B. In other words $10^6 i_f'/I_C$ will equal $10^6 i_f'/I_B$. The far-end crosstalk between the terminals A and C, due to length AB, will be $10^6 i_f'/I_A$. This differs from the output-to-output crosstalk at C in that the reference current is I_A instead of I_C. The part of the far-end crosstalk between A and C due to AB is, therefore, obtained from the output-to-output crosstalk at B by simply multiplying by the attenuation ratio I_C/I_A. If the output-to-output crosstalk is expressed as a loss in decibels, the far-end crosstalk is obtained by adding the net loss of the complete circuit between A and C.

Effects of Transpositions

The effects of transpositions on both the transmission currents and the crosstalk currents will now be discussed in a general way. The general method of computing the crosstalk between circuits without constructional irregularities and transposed in any manner will also be outlined.

General Principles

If there is only one circuit on a pole line, and this is balanced and free from irregularities, the communication currents will be propagated along this circuit according to the simple exponential law. If a current is propagated from the start of the circuit to some other point at a distance L, the magnitude of the current will be reduced by the attenuation factor $\epsilon^{-\alpha L}$ and the phase of the current will be retarded by the angle βL where α is the attenuation constant and β is the phase change constant.

If there are a number of circuits on a pole line this simple law of propagation may be altered due to crosstalk into surrounding circuits. This is illustrated by the curves, Fig. 6, which indicate the relation between observed output-to-input current ratio and frequency for two different circuits, each about 300 miles long and having 165-mil copper wires. The number of decibels corresponding to the current ratio is plotted rather than the ratio itself. For the simple law of propagation such curves would show the number of decibels increasing smoothly with frequency due to increasing losses in the line wires and insulators. The upper curve is for a circuit too infrequently transposed for the frequency range covered and the current ratio is abnormally small at particular frequencies. The corresponding number of decibels is abnormally large. The lower curve is for a circuit much more fre-

quently transposed and its current ratios practically follow the simple propagation law mentioned above over the frequency range shown.

Even though a circuit is very frequently transposed, its propagation constant is slightly affected by the presence of other circuits on the line. This may be explained by consideration of Figs. 2B and 7. As previously explained, Fig. 2B indicates the indirect transverse crosstalk by way of a tertiary circuit in one thin transverse slice of a parallel

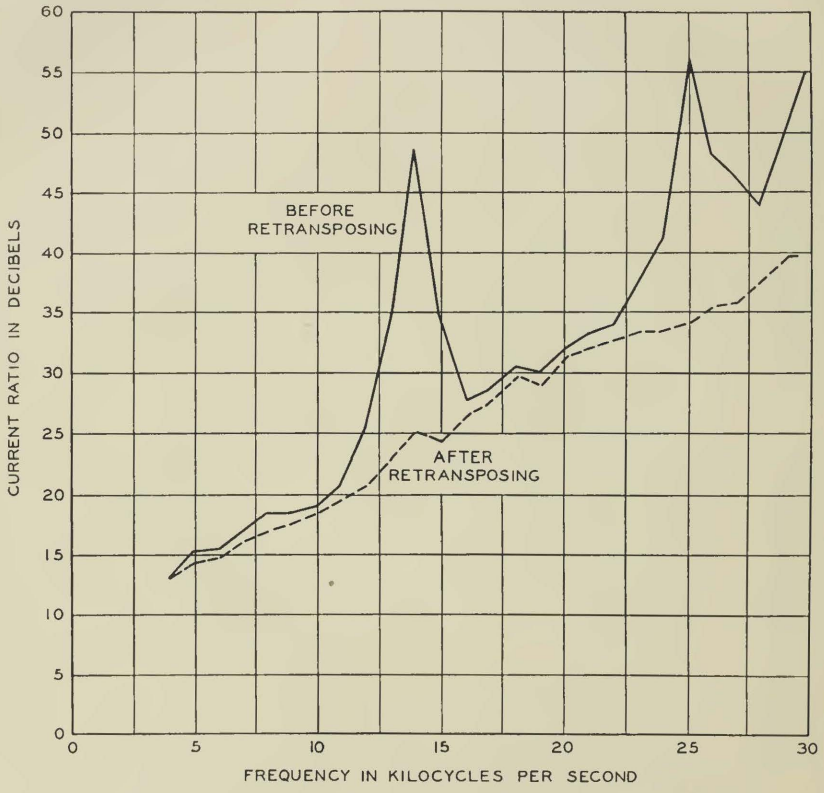


Fig. 6—Effect of transpositions on attenuation of an open-wire pair.

between two long circuits *a* and *b*. The circuit *c* has currents and charges due to crosstalk from the disturbing circuit *a*. These currents and charges not only alter the crosstalk currents in circuit *b* but also react to change the transmission current in circuit *a*. Since circuits *a* and *c* are loosely coupled, this reaction effect could usually be estimated with sufficient accuracy by calculating the crosstalk from *a* to *c* and back again and neglecting the further reactions of the change in the current in *a* on the current in *c*, etc.

Figure 7 shows the crosstalk paths from a to c and back again. In this figure, circuit a is indicated as two separate circuits for comparison with Fig. 2B. It is assumed that circuit a in Fig. 7 is energized at point A, the currents I_A and I_B being the currents which would exist at the input and output of the short length d if there were no tertiary circuits. The near-end crosstalk path indicated by n will cause a small crosstalk current i_n at point A in circuit a. There will be a crosstalk path similar to n in each thin slice of the parallel between a and c. Each of these paths will transmit a small crosstalk current to point A in circuit a. The sum of all these crosstalk currents will *increase* the input current I_A and, therefore, the impedance of circuit

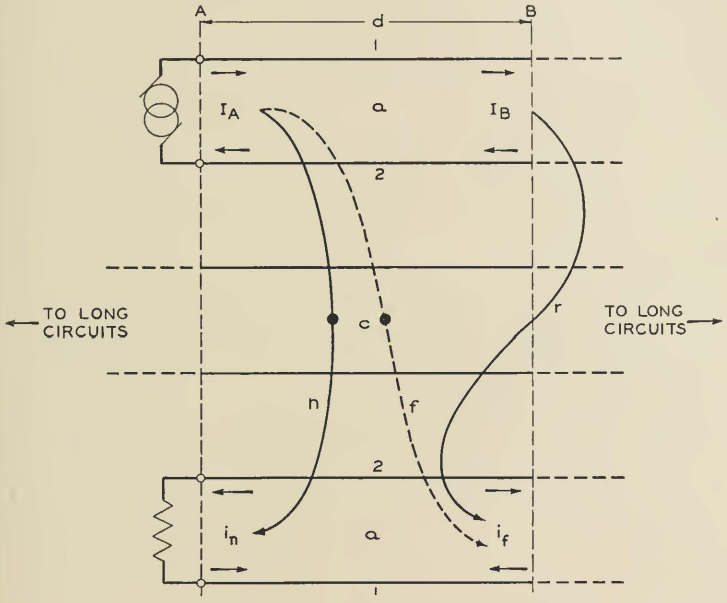


Fig. 7—Effect of circuit c on propagation in circuit a.

a is lowered. Thin slices remote from the sending end will contribute little to this effect, since the crosstalk currents from such slices will be attenuated to negligible proportions. A long circuit on a multi-wire line will, therefore, have a definite sending-end impedance slightly lower than that for one circuit alone on the line.

Figure 7 also indicates a far-end crosstalk path f which produces a crosstalk current i_f at point B in circuit a. This *reduces* the transmission current I_B at this point and, therefore, increases the attenuation constant of the circuit. For calculations of both the circuit

impedance and attenuation, the effect of surrounding circuits is taken care of in practice by using a capacity per unit length slightly higher than the value which would exist with only one circuit on the line. The proper capacity to use is determined in practice by measurements on a short length of a multi-wire line.

The effect on the propagation constant of the transverse crosstalk paths indicated by n and f of Fig. 7 cannot be suppressed by transpositions. As explained later, if the two circuits marked a were actually different circuits, the effect could be largely suppressed by transposing one circuit at certain points and leaving the other circuit untransposed at these points. Since the disturbing and disturbed circuits indicated by Fig. 7 are actually the same circuit, they must be transposed at the same points and, therefore, the transverse crosstalk effect cannot be suppressed by frequent transpositions.

Figure 7 also shows a crosstalk path marked r. This is one of the possible interaction crosstalk paths. The effect of such paths on the impedance and attenuation of the circuit may be largely suppressed by suitable transpositions. The difference between the two curves of Fig. 6 is due to lack of this suppression in the case of the upper curve.

Such an extreme effect of crosstalk reacting back into the primary or initiating transmission circuit and thus affecting direct transmission is seldom important in practical transposition design. A marked reaction on the primary circuit would necessitate such large crosstalk currents in neighboring communication circuits as to make them unfit for communication service at the frequency transmitted over the primary circuit. Therefore, it is only when the neighboring circuits are not to be used at this frequency that transposition design to control simply the direct transmission becomes of practical importance. When many circuits on a line are used for carrier operation, the crosstalk currents must be made so weak (by transpositions, physical separation of circuits, etc.) that their reactions back into the primary circuits are very small.

The effect of transpositions on crosstalk from one circuit into another *different* circuit will now be considered. The discussion of the control of this effect is the main object of this paper.

Figure 8A shows a short segment of a parallel between two long circuits and a near-end crosstalk coupling marked n. The segment could be divided into a series of thin slices and theoretically there would be interaction crosstalk between different slices. The segment length is, however, assumed to be short enough to neglect interaction crosstalk. The coupling n is, therefore, due either to direct or indirect transverse crosstalk in the short segment or to both of these types of

crosstalk. If circuit a is energized from the left, a near-end crosstalk current i_n results at point A in circuit b.

If two successive short segments are considered, as indicated by Fig. 8B, there will be a near-end crosstalk coupling n in each segment and each of these couplings will result in a crosstalk current at point A

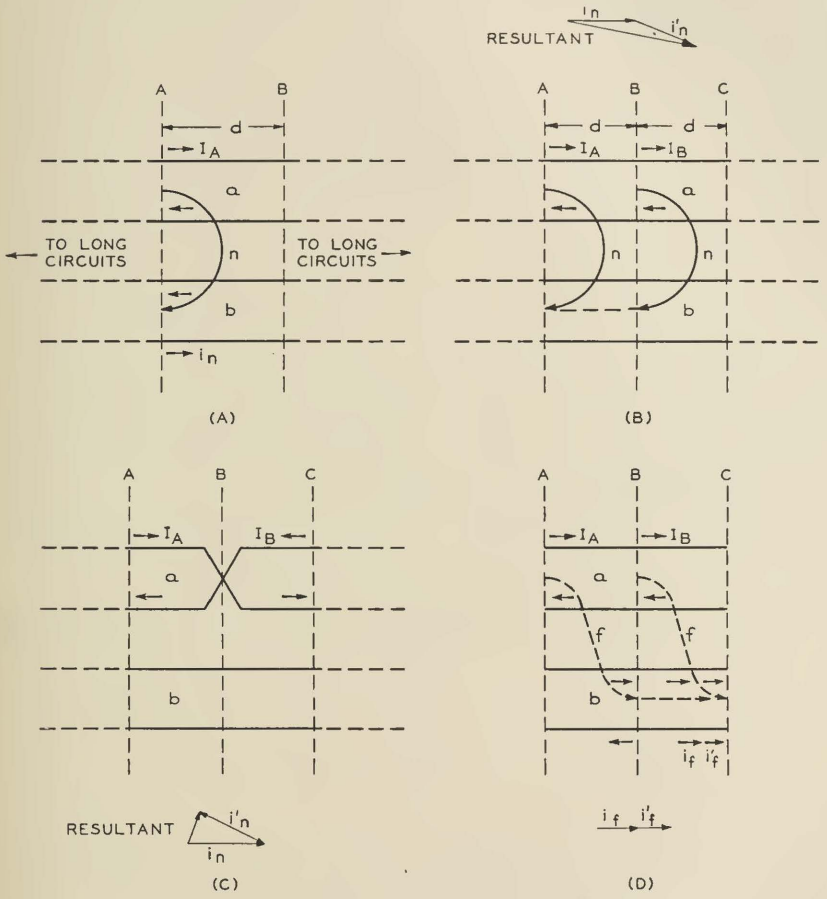


Fig. 8—Effect of transpositions on transverse crosstalk.

of circuit b. This is indicated by the vector diagram over the figure, where i_n indicates the crosstalk current due to the segment AB and i_n' indicates the crosstalk current at A due to BC. The latter current is slightly smaller and slightly retarded in phase with respect to i_n because in order for i_n' to appear at point A, the transmission current I_A must be propagated a distance d and the resulting crosstalk current

at B must also be propagated a distance d in order to reach A. As indicated by this vector diagram the total crosstalk current due to the two short segments is a little less than the arithmetic sum of the individual crosstalk currents.

Figure 8C is like Fig. 8B except that a transposition is inserted in the middle of circuit a at point B. This reverses the phase of the transmission current at the right of B and also reverses any crosstalk current due to current in circuit a between B and C. As a result the crosstalk current i_n' of Fig. 8B is reversed and the resultant of the two crosstalk currents is very much reduced as indicated by the vector diagram of Fig. 8C. The angle between i_n and i_n' is proportional to the length $2d$ which equals AC. The tendency for the two currents to cancel may, therefore, be increased by reducing the length AC which, in a long line, would mean increasing the number of transpositions.

Figure 8D is like Fig. 8B except that the far-end transverse crosstalk coupling f in each of the two short segments is considered. The coupling in the left-hand segment results in a crosstalk current at point B of circuit b, which is propagated to point C as indicated by i_f. The far-end crosstalk coupling in the right-hand segment produces a crosstalk current i_f' at point C. Since the total propagation distance is from A to C for both of these crosstalk currents, they must be equal in magnitude and in phase if circuits a and b are similar. This is indicated by the vector diagram of Fig. 8D. A transposition at point B in either circuit would reverse one of these crosstalk currents and, therefore, the resultant crosstalk current would be nil.

From consideration of Figs. 8C and 8D, it may be seen that if both circuits were transposed at point B, the sum of the crosstalk currents for the two segments would be the same as if neither circuit were transposed. Transposing one circuit reverses the phase of one of the component crosstalk currents, but if the second circuit is also transposed the original phase relations between the two currents are restored.

The foregoing discussion applies only to transverse crosstalk as discussed in connection with Fig. 2. When interaction crosstalk must be considered, a different principle is involved.

In connection with Fig. 8D, it was shown that the transverse far-end crosstalk between similar circuits could be readily annulled by transposing one of the circuits at the center of their paralleling length. Far-end crosstalk of the interaction type is not so readily annulled. The effect of transpositions on this type of crosstalk is indicated by Fig. 9.

This figure shows four short segments in a parallel between two

circuits *a* and *b*, there being an interposed tertiary circuit *c*. Interaction crosstalk involving two near-end crosstalk couplings is considered since this is usually the controlling type. There is an interaction crosstalk path designated *r* between the first two segments as indicated by Fig. 9A. There is a similar path between the third and fourth segments. Each of these paths would produce a far-end crosstalk current in circuit *b* at point *E*. For similar circuits these currents would be equal in magnitude and would add directly. The two currents can be made to cancel by transposing one of the circuits at *C*, the midpoint of the parallel. Such a transposition also cancels the transverse far-end crosstalk in length *AC* against that in length *CE*. There remains, however, the interaction crosstalk between length *CE* and length *AC*.

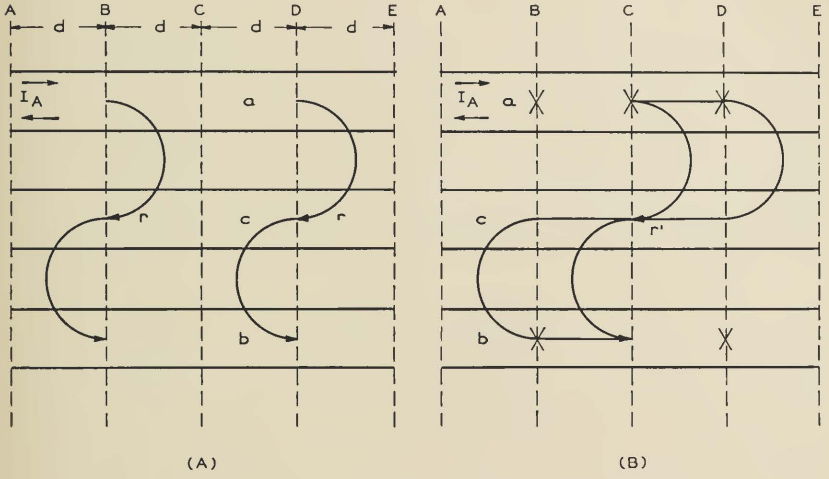
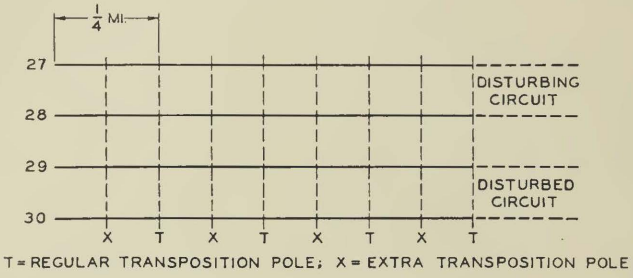


Fig. 9—Effect of transpositions on interaction crosstalk.

Figure 9B shows a transposition at *C* in circuit *a* and also other transpositions whose purpose is to minimize the interaction crosstalk between length *CE* and length *AC*. This crosstalk coupling, designated by *r'*, is a compound effect, depending on the near-end crosstalk between circuit *a* and circuit *c* in length *CE* and the near-end crosstalk between *c* and *b* in length *AC*. The near-end crosstalk coupling between *a* and *c* in length *CE* can be greatly reduced by a transposition in circuit *a* at point *D*, while the crosstalk coupling between *c* and *b* in length *AC* can likewise be reduced by a transposition at point *B* in circuit *b*. The latter two transpositions would not, however, minimize the interaction crosstalk between *CE* and *AC* with circuit *b* as the

disturbing circuit and it is necessary, therefore, to transpose both circuits at points B and D. The addition of these four transpositions does not affect the cancellation of far-end crosstalk in length AC against that in length CE by means of the transposition at C. After the four transpositions are added, length AC is still similar to length CE and the far-end crosstalk currents at E, due to these two lengths, are equal. Therefore, they will cancel when one of them is reversed in phase by the transposition at C.

It may be concluded that, while transposing both circuits at the same points has no effect on transverse crosstalk, it has a large effect on the interaction crosstalk. An experimental illustration is given in Fig. 10. This figure shows frequency plotted against output-to-output



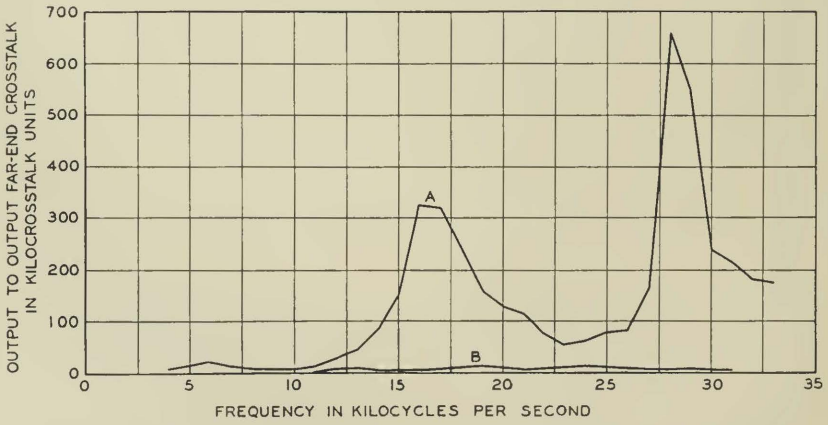


Fig. 10—Effect on far-end crosstalk of extra transpositions in both circuits.

far-end crosstalk between the two side circuits of a phantom group on a 140-mile length of line. The curve marked A is for the two circuits transposed for voice-frequency operation. Curve B is for the two circuits transposed in the same manner except that four transpositions

per mile were added to both circuits at the same points which are indicated by x on Fig. 10. The large effect of these transpositions shows the practical importance of the interaction type of far-end crosstalk.

In connection with Fig. 9B, there arises the question of how far apart the transpositions can be placed without serious crosstalk, in other words, how long is it permissible to make the segment d. If this length is increased the transpositions at B and D become less effective in suppressing the near-end crosstalk between a and c in length CE and between c and b in length AC. The degree to which the interaction crosstalk path r' must be suppressed is, therefore, important in determining the maximum permissible length of d. If d is increased the transposition at C becomes less effective in controlling the near-end crosstalk between a and b and, therefore, the length d also depends on the permissible near-end crosstalk.

It may be noted that transpositions at B and D in but one of the circuits a or b will help to suppress r', but the suppression is less effective than if both circuits are transposed at these points. If a is transposed at B and D the near-end crosstalk between a and c in length CE is reduced but the near-end crosstalk between c and b in length AC is not reduced. The product of these two near-end crosstalk values is greater, therefore, than if they had both been reduced by transposing both circuits at B and D.

Crosstalk Coefficients

The crosstalk between any two long open-wire circuits may be calculated by dividing the parallel into a succession of thin transverse slices and summing up the crosstalk for all these slices. To calculate the crosstalk in any slice it is necessary to know certain "crosstalk coefficients." The discussion below defines these coefficients and describes briefly how they are measured or computed.

Figures 2A and 2B indicate both near-end and far-end crosstalk coupling of both the direct and indirect transverse types in a thin transverse slice. Any of these couplings may be expressed in crosstalk units and the value of the coupling in a short length divided by the length in miles is called the crosstalk per mile. Since, as shown in the previous section, the crosstalk may not increase directly as length, strictly speaking, the crosstalk per mile is the limit of the ratio of coupling to length as the length approaches zero. The crosstalk per mile includes both the direct and indirect types of transverse crosstalk coupling. In the frequency range of interest (i.e., above a few hundred cycles for near-end crosstalk and above a few thousand cycles for

far-end crosstalk) this total transverse coupling varies about directly with the frequency and the crosstalk coefficient commonly used is the *crosstalk per mile per kilocycle.*

If many wires are involved, it is impracticable to determine these coefficients with good accuracy by computation and they are, therefore, derived from measurements. Examples of near-end and far-end coefficients, plotted against frequency, are shown in Fig. 11. The coefficients are for pairs designated 1–2 and 3–4 on the pole head diagram shown on the figure. These coefficients were derived from measurements of the near-end and far-end crosstalk over a range of frequencies. The length of line was about .2 mile and, for the range of frequencies covered, this length is sufficiently short so that interaction crosstalk is negligible and the transverse crosstalk is directly proportional to the length. The coefficients plotted are, therefore, nearly equal to the measured values of crosstalk divided by the length and by the frequency. (A small correction was made at the higher frequencies to allow for deviation of near-end crosstalk from simple proportionality to length and the curves were "smoothed" through the actual points calculated from the measurements.)

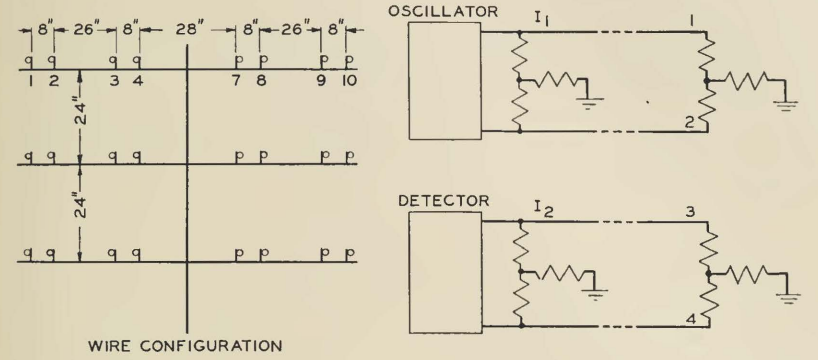
In order to obtain the crosstalk coefficients applicable to a short part of a long line, all the wires on the line were terminated in such a manner as to roughly simulate their extension for long distances in both directions, but without crosstalk coupling between the test pairs in such extensions. This is done by terminating each pair at each end with a resistance approximating its characteristic impedance and connecting the midpoint of each resistance to ground through a second resistance. These latter resistances terminate any phantom of two pairs as indicated on Fig. 11 for pairs 1–2 and 3–4. Any circuit with ground return is also terminated by these resistances.

Both of the test pairs are transposed at the midpoint of the line during the measurement. This minimizes the currents reaching the ends of the tertiary circuits and makes even the above approximate termination of the tertiary circuits of little importance.

Figure 11 shows near-end and far-end crosstalk coefficients for three conditions, A, B, and C. The two curves marked A show the measured values with all wires terminated and the test pairs transposed as described above.

For curves B, only the transposed test pairs were terminated as described above and the other wires were opened at the middle, at the quarter points and at both ends. Since no section of any of these wires connected points of substantially different potential in the field of the disturbing circuit there were practically no currents or charges

in these wires and the crosstalk coefficients for the two test pairs were practically the same as if the other wires had been removed from the line. It will be seen that the crosstalk coefficients for curves *B* are



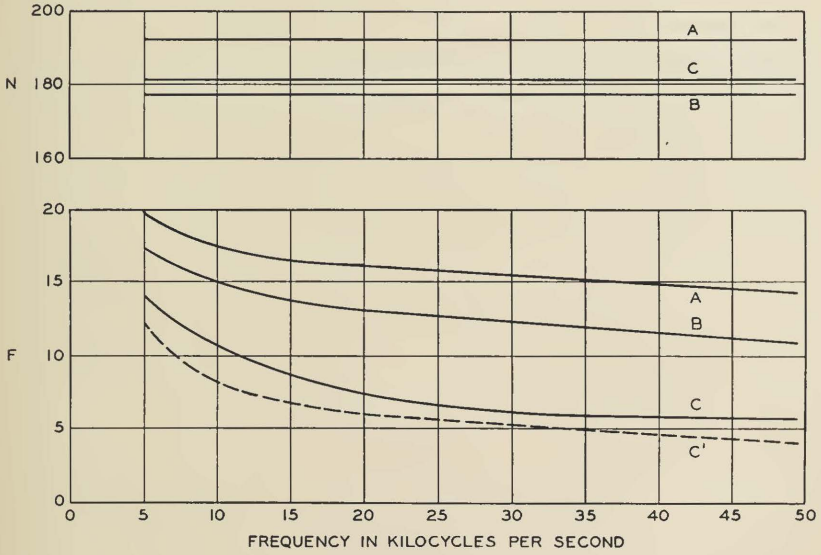


Fig. 11—Near-end and far-end crosstalk coefficients between pairs 1–2 and 3–4.

less than those for curves *A*. The coefficients of curves *B* involve tertiary circuits, however, since there could be crosstalk currents in the phantom of the two test pairs and also in the ghost circuit involving wires 1 to 4 with ground return.

Curves *C* show the coefficients with the test pairs *without* transpositions and terminated at both ends as accurately as practicable, but without the midpoints of these terminations connected to ground to terminate the phantom and ghost circuits. These tertiary circuits were, with this arrangement, prevented from connecting points of substantially different potential and the coefficients of curves *C*, therefore, approach the direct crosstalk coefficients. It is extremely difficult to experimentally determine the direct far-end coefficient. It may be computed, however, and the computed value which assumes perfect terminations and the effect of the phantom completely removed is shown by curve *C'*.

It may be noted that the near-end crosstalk coefficients are about independent of frequency. This is ordinarily true above a few hundred cycles. The total far-end coefficient (curve *A*) is about independent of frequency in the important carrier frequency range. The direct far-end coefficient of curve *C'* decreases considerably with frequency for reasons discussed in Appendix A. Since transpositions are ordinarily designed for the condition of a number of wires on a line, the total crosstalk coefficient is the one usually used in practice.

Curves *C* of Fig. 11 also indicate that the direct near-end coefficient is much larger than the direct far-end coefficient. This is usually true and, as discussed in detail in Appendix A, the explanation is that the crosstalk currents caused by the electric and magnetic fields add almost directly in the case of direct near-end crosstalk but tend to cancel in the case of direct far-end crosstalk. As discussed in the appendix, the indirect (vector difference of curves *A* and *C*) crosstalk in a very short length is due almost entirely to the electric field of the tertiary circuits and is the same for both near-end and far-end crosstalk. In Fig. 11, the total near-end coefficient (curve *A*) is increased by the indirect crosstalk since curve *C* is lower than curve *A*. The reverse is usually true, however. In the case of far-end crosstalk the total coefficient is usually increased by the indirect crosstalk.

Crosstalk coefficients are vector quantities and may be measured in magnitude and phase. If it is desired to compute the crosstalk between two long pairs of wires which do not change their pin positions, it is only necessary to know the magnitude of the crosstalk coefficient, since the problem is to determine the ratio of the crosstalk for many elementary lengths to the crosstalk for one such length. However, if it is desired to know the crosstalk between long circuits which do change their pin positions, several crosstalk coefficients must be known, one for each combination of pin positions. In order to determine the total crosstalk for several segments of a line involving different pin

positions, it is necessary to know both the phase and magnitude of the crosstalk coefficients. For practical purposes, however, the coefficients may, in most cases, be regarded as algebraic quantities having sign but not angle.

The direct component of the total crosstalk coefficient may be readily computed as discussed in Appendix A. If more than a very few wires are involved, an exact calculation of the indirect component is impracticable but a fair approximation may be obtained by the method discussed in Appendix A. This method is used when a wire configuration is under consideration but is not available for measurement.

As pointed out in 1907 by Dr. G. A. Campbell, an accurate calculation of the total crosstalk coefficient would involve determination of the "direct capacitances" between wires of the test pairs. Since these capacitances are functions of the distances between all combinations of wires on the lead and between wires and ground, their calculation is usually impracticable. In the past, the crosstalk coefficients were computed by a method proposed by Dr. Campbell which involved measurement of the direct capacitances.[3]

The part of the coefficient due to the electric field was computed from the "direct capacitance unbalance." The part due to the magnetic field was computed as discussed in Appendix A. When loaded open-wire circuits were in vogue it was necessary to be able to separate the electric and magnetic components of the coefficients. After loading was abandoned this separation was unnecessary and it was found more convenient to measure the total coefficients than to measure the direct capacitances or differences between pairs of these capacitances.

As previously discussed, in designing transpositions it is necessary to compute the interaction type of crosstalk indicated by Fig. 2C, and it is, therefore, necessary to have some coupling factor for use in this computation. Such a coupling factor could, theoretically, be determined as indicated schematically by Fig. 12. The interaction crosstalk between two short lengths of line would be measured by transmitting on one pair and receiving on the other pair at the junction of the two short lengths as indicated by the figure.

If there were but a single tertiary circuit such as c of the figure, the crosstalk measured would be that due to the compound crosstalk path $n_{ac}n_{cb}$. In this product, n_{ac} is the near-end crosstalk between a and c in the right-hand short length d and n_{cb} is the near-end crosstalk between c and b in the left-hand short length. Since n_{ac} and n_{cb} when

[3] See papers by Dr. Campbell and Dr. Osborne listed under "Bibliography."

expressed in crosstalk units are current ratios times a million, their product $n_{ac}n_{cb}$ is a current ratio times a million squared. The crosstalk measured would be this current ratio times a million or $n_{ac}n_{cb}10^{-6}$.

For small values of d, n_{ac} and n_{cb} vary directly as the frequency and as the length d. Therefore:

$$n_{ac} = N_{ac}Kd,$$

$$n_{cb} = N_{cb}Kd,$$

$$n_{ac}n_{cb}10^{-6} = N_{ac}N_{cb}K^2d^2 10^{-6},$$

where N_{ac} and N_{cb} are the near-end crosstalk coefficients, K is the frequency in kilocycles and d is expressed in miles. The measured

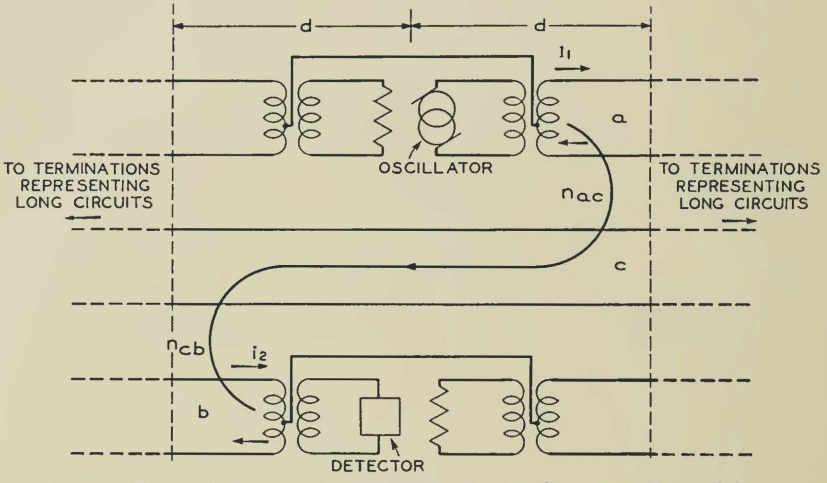


Fig. 12—Theoretical method of measuring interaction crosstalk coefficient.

crosstalk $n_{ac}n_{cb}10^{-6}$ divided by K^2d^2 gives the quantity $N_{ac}N_{cb}10^{-6}$ which may be designated as I_{ab} and called *the interaction crosstalk coefficient.* Values of I_{ab} determined from crosstalk measurements on multi-wire lines would include the effect of numerous tertiary circuits instead of that of a single tertiary circuit as indicated by Fig. 12.

While the interaction crosstalk coefficient I_{ab} could theoretically be measured as outlined above, it is simpler to deduce an approximate value from the measured value of the far-end crosstalk coefficient F_{ab}. The indirect component of F_{ab} is due to the tertiary circuits and must, therefore, be related to I_{ab} which is also due to these circuits. As discussed in detail in Appendix A:

$$I_{ab} = -\frac{2\gamma_c F_{ab}}{K} \text{ approximately.}$$

In this expression K is the frequency in kilocycles and $\gamma_c = \alpha_c + j\beta_c$ is the propagation constant of the tertiary circuit c. On a multi-wire line there would be numerous tertiary circuits with various values of γ. With practicable wire sizes the attenuation constants indicated by α are small compared with the phase change constants indicated by β. Measurements of crosstalk indicate that the values of β are all in the neighborhood of the value given by the expression $\pi K/90$. This corresponds to a speed of propagation of 180,000 miles per second which is about the average for the present carrier frequency range. Neglecting the attenuation constants:

$$\gamma_c = j\beta = j\frac{\pi K}{90},$$

$$I_{ab} = -j\frac{2\pi F_{ab}}{90}.$$

This relation is much used in transposition design. As noted above, the indirect component of F_{ab} should, strictly speaking, be used to obtain I_{ab}. In most cases, however, the total value of F_{ab} may be used since this total is determined largely by the indirect component.

Type Unbalance

A conception important in transposition design is that of "type unbalance." This conception will now be explained and the general method of computation will be discussed.

As we have seen, any two open-wire circuits tend to crosstalk into each other due to coupling between them. By transposing the circuits, the coupling in any short length of line is nearly balanced in another short length by a second coupling of about the same size but about opposite in phase. This balancing is never perfect and there is always a residual unbalanced coupling due to (1) attenuation and change in phase of the disturbing transmission current and resulting crosstalk currents as they are propagated along the circuits and (2) irregularities in the spacing of the transpositions and irregularities in the spacings between the various wires. The term "type unbalance" has been chosen to indicate the residual unbalance caused by propagation effects. It is expressed as an "equivalent untransposed length," that is, the type unbalance times the crosstalk per mile gives the residual crosstalk due to propagation effects assuming no constructional irregularities.

The method of computing the type unbalance for near-end crosstalk will now be discussed. The part of the near-end crosstalk due to interaction between all the different thin slices of line may be ignored

since, as discussed in connection with Figs. 2C and 2D, the interaction crosstalk involves the product of a near-end crosstalk path and a far-end crosstalk path. This product is small since the coupling through the far-end path is inherently small. Therefore, the interaction crosstalk coefficient is much smaller for near-end crosstalk than for far-end crosstalk, while for the transverse crosstalk coefficients the reverse is true.

As was indicated by the discussion of Fig. 8B, the transverse near-end crosstalk between two long circuits may be computed by dividing the parallel into short segments, each having the same transverse crosstalk coupling. The coupling between circuit terminals for any segment will be different from that at the segment terminals due to propagation effects as explained in connection with Fig. 8B. Therefore, the coupling at the circuit terminal for each segment must be determined and, finally, the sum of the coupling values for all the segments.

The simplest case is that of two non-transposed circuits. The problem is indicated by Fig. 13 which is like Fig. 8B except that more segments are shown.

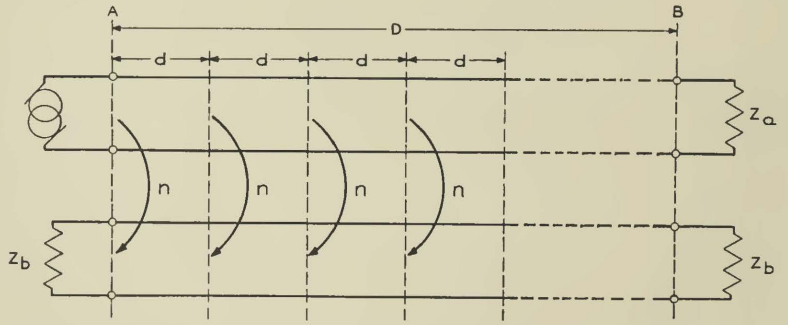


Fig. 13—Method of computing near-end crosstalk between untransposed circuits in length D.

The near-end crosstalk coupling n at point A due to the first segment is NKd, where N is the crosstalk coefficient and K is the frequency in kilocycles. The crosstalk current from the second segment relative to that from the first segment is attenuated by the factor $\epsilon^{-(\alpha_1+\alpha_2)d}$, and also retarded in phase by the angle $\epsilon^{-j(\beta_1+\beta_2)d}$. In other words, the crosstalk current from the second segment is equal to the crosstalk current from the first segment times the factor $\epsilon^{-(\gamma_1+\gamma_2)d}$, where γ_1 and γ_2 are the propagation constants for the two circuits and γ equals $\alpha + j\beta$. Letting γ be the average propagation constant, the coupling

at point A for the second segment is equal to that for the first segment times $\epsilon^{-2\gamma d}$ or $NKd\epsilon^{-2\gamma d}$. The coupling at point A for the third segment is $NKd\epsilon^{-4\gamma d}$. The sum of the crosstalk couplings at point A at all the segments is, therefore:

$$NKd(1 + \epsilon^{-2\gamma d} + \epsilon^{-4\gamma d} + \epsilon^{-6\gamma d} + \text{etc.}).$$

This expression may be summed up for the number of segments corresponding to the total length D. It is simpler, however, to let d be an infinitesimal length and to integrate over the length D, i.e., from point A to point B of Fig. 13. This gives for the total near-end crosstalk for non-transposed circuits:

$$NK\frac{1 - \epsilon^{-2\gamma D}}{2\gamma}.$$

In the special case when D is only the usual short segment between transposition poles, the above expression is practically equal to NKD.

The near-end crosstalk between circuits having transposition poles spaced a considerable distance D apart may now be computed. Figure 14 shows a length $2D$ in a parallel between two long circuits, there being a transposition in one circuit at the center of $2D$. The near-end crosstalk for the length AB is given by the above expression. The near-end crosstalk at point A for the length BC will be the same expression multiplied by the propagation factor $\epsilon^{-2\gamma D}$ and reversed in sign due to the effect of the transposition. The near-end crosstalk at point A for the length $2D$ will, therefore, be the sum of the values for lengths AB and BC. This sum is:

$$NK\frac{1 - \epsilon^{-2\gamma D}}{2\gamma}(1 - \epsilon^{-2\gamma D}).$$

This quantity divided by NK is the type unbalance for the length $2D$ of Fig. 14. If D is only the length of a short segment the above expression is about equal to $NKD(2\gamma D)$.

Similarly the near-end crosstalk at point A for a length $3D$ will be:

$$NK\frac{1 - \epsilon^{-2\gamma D}}{2\gamma}(1 - \epsilon^{-2\gamma D} \mp \epsilon^{-4\gamma D})$$

and the type unbalance is this quantity divided by NK. For a length $4D$ the quantity in the parentheses becomes $(1 - \epsilon^{-2\gamma D} \mp \epsilon^{-4\gamma D} \mp \epsilon^{-6\gamma D})$, etc. The sign of each term in the parentheses is determined by the arrangement of "relative" transpositions, i.e., those at points where only one of the two circuits is transposed. Each term corre-

sponds to a length D. The transposition at the start of the second length (at point B of Fig. 14) reverses the sign of the term for the second length and also the signs for the following lengths until another transposition is reached which makes the next sign plus, etc.

A practical open-wire line is divided into a series of "transposition sections" of eight miles or less. In each section the crosstalk between any two circuits is approximately balanced out by means of transpositions. A main purpose of this division into sections is to provide suitable points for circuits to drop off the line. A circuit on the line for a part of a section may have more crosstalk to a through circuit than if the parallel extended for the whole section since coupling in

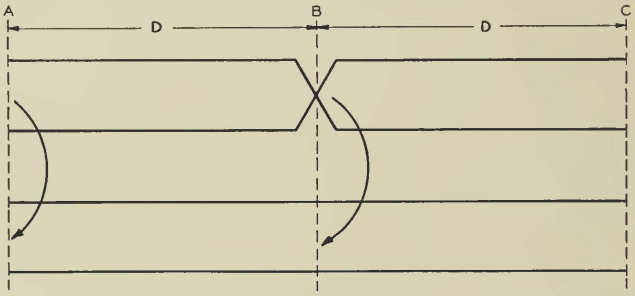


Fig. 14—Near-end crosstalk in length $2D$ between circuits a and b with circuit a transposed in the middle.

the last part of the section may tend to subtract from the coupling in the first part. The ends of sections are, therefore, the most suitable points for circuits to leave or enter the line. Ideally, the sections in a line should all be alike as regards length and transposition arrangements since this makes it practicable to so design the transpositions that residual crosstalk in one section tends to cancel that in another section. Practically, the sections vary in length and, therefore, in the transposition arrangements because the ends of some of the sections must fall at particular "points of discontinuity" determined by branching circuits and by requirements for balance against induction from power circuits.

In designing the transposition sections, type unbalances are computed for the section lengths of eight miles or less. For such lengths, the general method of computing type unbalances may be simplified. The general method involves the vector propagation constant γ. For a length as short as a single transposition section, attenuation can, ordinarily, be neglected. Therefore, in the type unbalance formulas γ can be replaced by $j\beta$ which greatly simplifies the computations.

Since attenuation can be neglected, the type unbalance for a transposition section depends only on the line angle βD. Since β increases practically directly with frequency, a plot of type unbalance against βD indicates the variation of type unbalance with frequency for a fixed length or the variation with length for a fixed frequency. It is convenient to plot the product of type unbalance and frequency (in kilocycles) since this product multiplied by the crosstalk coefficient gives the crosstalk. Two such plots for near-end type unbalance times frequency are shown on Fig. 15A. The plot marked P is for the condition of two circuits non-transposed or transposed alike. The plot marked O is for the same arrangement except for one relative transposition at the midpoint of the parallel.[4] The figure has a frequency scale corresponding to a length of eight miles as well as the general βD scale in degrees.

It will be seen that, for the case of no relative transpositions, the crosstalk varies directly with the frequency for only a short distance at the start of the curve. The effect of one relative transposition is to greatly reduce the crosstalk for small values of βL. For larger values the crosstalk is increased. It may be noted that the minimum values shown on the curves are somewhat in error since attenuation was neglected.

The minimum values in the P curve are due to "natural transpositions" in the non-transposed circuits. When the line angle is 180 degrees the crosstalk at the near-end of the disturbed circuit due to the second half of the line is just 180 degrees out of phase with the crosstalk due to the first half. This reversal in phase is due to the phase change accompanying the propagation of current to the mid-point and back. The total crosstalk due to both halves of the line lengths is the same as if the crosstalk coupling in the second half were translated to the near-end and the parallel without phase change but one circuit was transposed at the mid-point. When the line angle is 360 degrees the "natural transpositions" are at the quarter points, etc.

The near-end crosstalk between any two circuits in a transposition section may be estimated by multiplying the crosstalk coefficient by values of type unbalance times frequency similar to those of Fig. 15A. The total crosstalk in a succession of similar transposition sections is calculated at any particular frequency by working out a factor similar to the type unbalance in order to obtain the relation between the crosstalk in many transposition sections and that in one section. In calculating this factor, attenuation cannot be neglected since long lengths of line are involved.

[4] Two circuits are relatively transposed by one transposition at a given point in the line. Transpositions in both circuits leave them relatively untransposed.

The method of computing type unbalances for far-end crosstalk will now be explained. As in the case of near-end crosstalk, the type unbalance is defined by expressing the far-end crosstalk between two long circuits as the product of the crosstalk coefficient, the frequency in kilocycles and the type unbalance.

Figure 15B indicates the periodic variation with frequency of the far-end crosstalk when type unbalance is controlling.

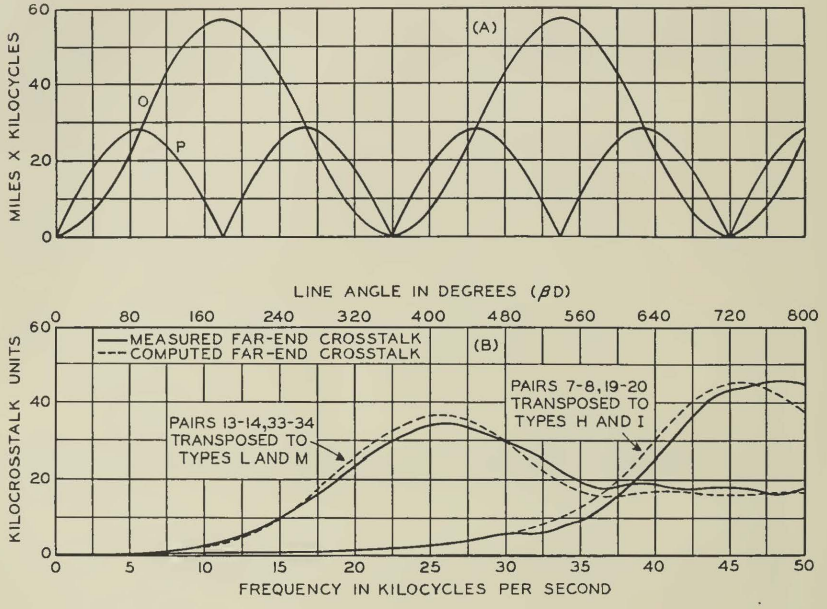


Fig. 15—Type unbalance and crosstalk vs. frequency and line angle in degrees. For Part (*B*), see Fig. 27*A* for wire configuration and Fig. 28 for transposition types.

In the case of near-end crosstalk, the method of computing the type unbalance neglected interaction crosstalk since, ordinarily, the transpositions needed to control transverse crosstalk make the interaction effect negligible. In the case of far-end crosstalk, the most important type of interaction crosstalk is included in calculations of type unbalances but another type of interaction crosstalk and the direct transverse crosstalk are neglected. The transpositions needed to properly suppress the important type of interaction crosstalk and the indirect transverse crosstalk ordinarily make the neglected types of crosstalk very small and the application of a more precise method of computing type unbalances for far-end crosstalk is not justified in practice.

The far-end type unbalance for a non-transposed part of a long parallel between two circuits will be computed first. Such a part of a parallel is indicated by length D of Fig. 16. For purposes of computation this length is divided up into a number of short segments each of length d. Considering the far-end crosstalk for two such segments at the start of the length D it will be seen from the discussion of crosstalk coefficients that transverse crosstalk in the length $2d$ will be

$$2FKd = 2(F_d + F_i)Kd.$$

In the above expression F is the far-end crosstalk coefficient, F_d being that part due to direct crosstalk and F_i that part due to indirect crosstalk.

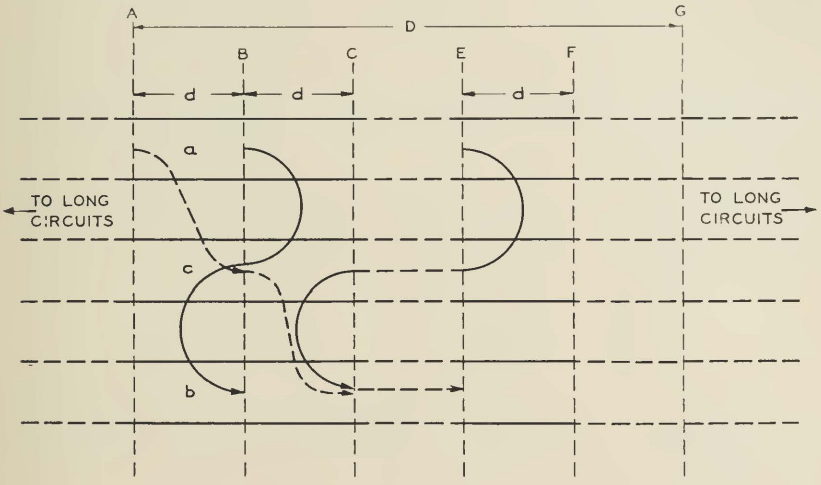


Fig. 16—Far-end crosstalk between untransposed circuits in length D.

The above expression relates to the output-to-output crosstalk. The input-to-output crosstalk is obtained by multiplying by the propagation factor $\epsilon^{-2\gamma d}$ to allow for propagation from A to C. This correction is usually made only when it is desired to obtain the input-to-output crosstalk between complete circuits and it is usually satisfactory to correct by using the attenuation factor and ignoring change in phase.

The total transverse output-to-output crosstalk in the length D is:

$$(F_d + F_i)KD.$$

This is about equal to F_iKD since F_d is ordinarily small compared to F_i.

Figure 16 indicates with a solid line the important type of interaction crosstalk between the first two segments by way of a representative tertiary circuit c. As discussed in the section on crosstalk coefficients and in Appendix A, the far-end crosstalk (output-to-output) of this interaction type will be

$$N_{ac}N_{cb}K^2d^2 10^{-6} = -2\gamma F_i K d^2 \text{ approximately.}$$

The interaction crosstalk as well as the transverse crosstalk is about proportional to the indirect coefficient F_i.

Each segment of the disturbing circuit will have a similar interaction crosstalk coupling with each preceding segment of the disturbed circuit. The interaction crosstalk between segment EF and segment BC is indicated on Fig. 16. The expression for this differs from the above expression in that the additional propagation distance from E to C and back must be allowed for. To get the total output-to-output far-end crosstalk it is necessary to sum up all these interaction crosstalk couplings between segments and to this sum add the total transverse crosstalk in length D.

This clumsy summation process may be avoided by letting d be an infinitesimal length and integrating between points A and G. This results in the following approximate expression for the output-to-output far-end crosstalk in the length D.

$$F_d KD + F_i KD + F_i K\left[\frac{1 - \epsilon^{-2\gamma D}}{2\gamma} - D\right].$$

This assumes the same propagation constant for the disturbing, disturbed and tertiary circuits. This approximation is justified for short lengths of, say, 10 miles or less.

The last term represents the interaction crosstalk and this term is negligible for small values of D. For larger values of D interaction crosstalk must be considered and it is convenient to rewrite the expression as follows:

$$F_d KD + F_i K\frac{1 - \epsilon^{-2\gamma D}}{2\gamma}.$$

The first term representing the direct crosstalk is negligible for values of D corresponding to a line angle of 90 degrees or less since F_d is ordinarily small compared with F_i and D is not large compared with $(1 - \epsilon^{-2\gamma D}/2\gamma)$. Therefore, direct crosstalk ordinarily may be neglected in computing far-end type unbalance. Another reason for neglecting direct crosstalk is that it is readily cancelled by a few relative transpositions while the remaining far-end crosstalk depends

upon the transpositions in a complicated way, because the various interaction crosstalk couplings involve a variety of propagation distances and, therefore, have a variety of phase angles. If both circuits are transposed frequently but alike the direct crosstalk is not affected by the transpositions but it is ordinarily small compared with the indirect transverse crosstalk.

Figure 16 indicates by a dashed line another type of interaction crosstalk involving the product of two far-end crosstalk couplings. This effect can be neglected with practical arrangement of transpositions but may be important in the case of circuits having few transpositions or none at all.

In computing type unbalance the far-end crosstalk in an untransposed segment of line of length D may, therefore, be written as:

$$F_i K \frac{1 - \epsilon^{-2\gamma D}}{2\gamma} = FK \frac{1 - \epsilon^{-2\gamma D}}{2\gamma} \text{ approx.}$$

Since the magnitude of F_i is ordinarily about equal to that of F, the measured coefficient, it is usually satisfactory to use the latter value.

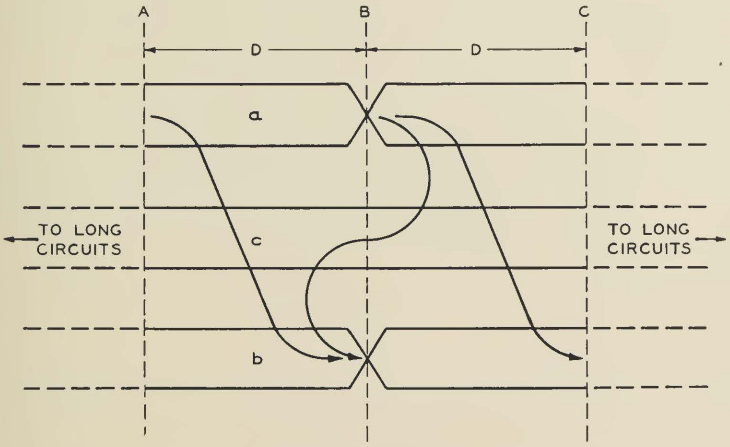


Fig. 17—Far-end crosstalk in length $2D$ between circuits a and b with each circuit transposed at the middle.

Having derived the above expression it is now possible to derive the far-end type unbalance for two transposed circuits. Figure 17 indicates a parallel between two long circuits. The type unbalance will be computed for a length $2D$ in which both circuits are transposed at the center. In the length $2D$ three far-end crosstalk paths must be considered, that is, the far-end crosstalk in length AB, that in length

BC and the important type of interaction crosstalk between length BC and length AB. The output-to-output crosstalk values only will be written for all these paths or, in other words, the effect of the propagation distance AC will not be considered in the expressions.

The far-end crosstalk for either length AB or BC is given by the above expression. Since *both* circuits are transposed at point B the far-end crosstalk values in the two lengths will add directly and their sum will be

$$2FK\frac{1-\epsilon^{-2\gamma D}}{2\gamma}.$$

Transmission from A to C through the crosstalk path in length AB is reversed in sign due to the transposition in circuit b at B. The output current of circuit a is also reversed in sign. In general, the output-to-output current ratio may or may not be reversed in sign depending on the transposition arrangement. It is convenient, however, to consider the first path as a reference and assign a plus sign to the crosstalk. Other paths are then assigned the proper relative phase angles.

As discussed in connection with Fig. 16, if the length D is very short the interaction crosstalk between the two segments may be written:

$$-2\gamma F_i K D^2 = -2\gamma FKD^2 \text{ approx.}$$

In practice the length D may be too long for this approximate expression in which case it is necessary to substitute for D in the above expression the value derived in connection with the discussion of the near-end crosstalk in a length D. In other words, D of the above expression should be replaced by

$$\frac{1-\epsilon^{-2\gamma D}}{2\gamma}.$$

With this substitution the interaction crosstalk between the two lengths becomes

$$-FK\frac{(1-\epsilon^{-2\gamma D})^2}{2\gamma}.$$

Transmission from A to C through this crosstalk path involves two transpositions and therefore the sign of the above expression is not reversed. Relative to the reference path through the crosstalk in length AB the sign should be reversed, however, and become plus. The total crosstalk in the length $2D$ is, therefore,

$$2FK\frac{1-\epsilon^{-2\gamma D}}{2\gamma}+FK\frac{(1-\epsilon^{-2\gamma D})^2}{2\gamma}=FK\frac{3-4\epsilon^{-2\gamma D}+\epsilon^{-4\gamma D}}{2\gamma}.$$

The latter expression divided by F is the frequency times the far-end type unbalance for the length $2D$. If one of the circuits were transposed at point B the crosstalk in length AB would be cancelled by that in length BC. The sign of the interaction crosstalk between the two lengths would be reversed and the expression would become

$$-FK\frac{(1-\epsilon^{-2\gamma D})^2}{2\gamma}.$$

If neither circuit were transposed at B, the far-end crosstalk would be that for a non-transposed length of $2D$ or:

$$FK\frac{1-\epsilon^{-4\gamma D}}{2\gamma}.$$

The frequency times the type unbalance values for the cases of one transposition and no transpositions are the same (in magnitude) as those derived for near-end crosstalk which were plotted (neglecting attenuation) as curves O and P on Fig. 15A.

If both circuits are transposed at B the near-end type unbalance remains the same as if there were no transpositions. The far-end type unbalance is radically altered, however. This is evident if the above equation is compared with that for the case of both circuits transposed.

This process of computing type unbalances may be extended from two equal lengths to any number of equal lengths. It is necessary to consider the interaction crosstalk between each length of the disturbing circuit and each preceding length of the disturbed circuit. The relative propagation distances through the various interaction crosstalk couplings must be taken account of.

Computations of far-end type unbalances are greatly simplified by assuming the same propagation constants for the disturbing, disturbed and tertiary circuits and by neglecting attenuation within a transposition section as in the case of near-end crosstalk. Since the tertiary circuit may be composed of any combination of wires on the line or of these wires and ground return, the propagation constant for a tertiary circuit may be somewhat different from that for the disturbing and disturbed circuits. This is particularly true of earth-return circuits, but these are of little practical importance due to their relatively high attenuation. All circuits not involving the earth have somewhere near the same speed of propagation but the tertiary circuits may differ greatly in attenuation constants.

For practical reasons a fair balance against crosstalk must be obtained in each transposition section (eight miles or less) and, as in the case of near-end crosstalk, type unbalances are calculated for the

transposition arrangements which may exist in a single transposition section. Since the attenuation in a transposition section is not great, these calculations need not take into account differences in the attenuation constants of the various tertiary circuits. A long line has a series of transposition sections of various types and the total far-end crosstalk for any two circuits is a summation of the crosstalk values obtained from the type unbalances for the various sections plus interaction crosstalk between the various combinations of sections. With practical methods of transposition design, the transposition arrangements are so chosen that the interaction crosstalk between two sections is usually small compared with the far-end crosstalk in one section. A long line for the most part consists of a succession of similar sections with occasional sections of other types. Interaction crosstalk between dissimilar sections does not ordinarily contribute appreciably to the total far-end crosstalk. For the important case of a succession of similar sections interaction crosstalk between sections must be carefully considered since it may build up systematically and the total may be large compared with the summation for the far-end crosstalk values for the individual sections.

Serious interaction crosstalk between similar sections is guarded against by computing factors relating the far-end type unbalance in one section to that in various numbers of successive sections with various transposition arrangements at the junctions of sections. The factors actually computed are somewhat in error since they involve long distances and assume the same attenuation constants for all circuits. The errors are not sufficient, however, to prevent the factors from being a proper guide in avoiding systematic building up of interaction crosstalk between sections.

The above discussion assumes that the tertiary circuits are indefinitely extended or terminated to simulate their characteristic impedance. The tertiary circuits may not be terminated at the ends of a line since many of them are not used for transmission of speech or signals. Complete reflections of the crosstalk current in the tertiary circuits will, therefore, occur at their ends and these reflections somewhat modify the crosstalk currents in other circuits. This effect is important in a very short line since the reflected wave is again reflected at the distant end and at particular frequencies large changes in the tertiary crosstalk currents may occur due to multiple reflections. In a long line such multiple reflections are damped out and, in general, tertiary circuit reflection effects are not important.

If all the pairs on a line are transposed for the same maximum useful frequency, the transposed pairs will usually be relatively

unimportant as tertiary circuits, that is, two pairs having small crosstalk between them usually contribute but little to the crosstalk between one of these pairs and any third pair. In some cases, however, this effect is important. On some lines certain pairs may be transposed for carrier operation and other circuits on the line for voice frequencies only. A combination of the two kinds of circuits may have large crosstalk between them at carrier frequencies and may contribute appreciably to the carrier frequency crosstalk between the pair transposed for carrier operation and some other pair also so transposed.

Far-end type unbalances which take account of transpositions in a tertiary circuit must, therefore, be calculated. This can be done by following the same general method discussed in connection with Fig. 17. From the discussion of coefficients it follows that the far-end coefficient for use in computing such a type unbalance will be:

$$\frac{-N_{ac}N_{cb}K}{2\gamma}\,10^{-6},$$

where N_{ac} and N_{cb} are the near-end crosstalk coefficients for the combination of disturbing circuit and tertiary circuit and the combination of tertiary circuit and disturbed circuit. Since these circuit combinations involve recognized transmission circuits, their near-end coefficients will be available since they must be measured or computed in order to compute the near-end crosstalk.

If a parallel between two circuits is divided into a large number of segments by transposition poles there is a wide variety of transposition arrangements which may be installed at these poles. It is, therefore, a complicated problem to devise charts and tables in reasonable numbers which will cover all the possible type unbalance values for the various transposition arrangements over a wide range of frequencies. This is particularly true in the case of far-end type unbalances since the type unbalance is altered by transposing both circuits at the same points and it is necessary to work out a type unbalance for each combination of transposition arrangements which may be used in two circuits. In the case of near-end crosstalk a number of different transposition arrangements will have the same type unbalance since only the relative transpositions need be considered.

The circuit capacity of a line may be increased by the use of phantom circuits (generally when carrier-frequency systems are not involved) which must, of course, be transposed to avoid noise and crosstalk. The crosstalk between phantom circuits may be calculated in a manner similar to that for pairs. The calculation of crosstalk between side

circuits of the phantoms or between a side circuit and a phantom circuit is complicated by the fact that the phantom transpositions cause the side circuits to change pin positions. Near-end and far-end type unbalances have been computed, however, which take account of this "pin shift" effect of the phantom circuits. In general, the use of phantom circuits seriously limits the crosstalk reduction which may be obtained by transpositions. Phantom circuits are often uneconomic since they seriously restrict the number of carrier frequency channels which may be operated over a given pole line.

As indicated by Fig. 15A the values of type unbalance times frequency have marked maximum and minimum values when they are plotted against frequency or length. The maximum values are usually reduced by increasing the number of transpositions in a given length. When there are a number of circuits on the line it is usually necessary that the propagation of current between successive transposition poles does not change the phase by more than about five degrees. Since the phase change is about two degrees per mile per kilocycle the maximum transposition interval in miles is about $2.5/F$ where F is the frequency in kilocycles. This means .25 mile or 1300 feet at 10 kilocycles and .06 mile or 300 feet at 40 kilocycles.

It does not follow, however, that the least maximum value of type unbalance for a range of frequencies is obtained by using the greatest number of transpositions for a given number of transposition poles. This is illustrated by Fig. 15A which shows that the least maximum value is obtained with no transpositions rather than with one transposition. The total crosstalk current at a terminal is composed of numerous elements of various magnitudes and phase relations. The vector sum of these elements tends to be small at particular frequencies with no transpositions at all and it is important to preserve this tendency as much as possible when choosing an arrangement of transpositions. The vector sum of the elements can never be made zero since this would require that the circuits have no attenuation and infinite speed of propagation. This sum and, therefore, the type unbalances can be made very small, however, by choosing a suitable transposition arrangement and making the interval between transposition poles very small. In practice, the values of type unbalance times frequency for adjacent circuits are restricted to values much less than those of Fig. 15A.

Vacuum Tube Electronics at Ultra-high Frequencies *

By F. B. LLEWELLYN

Vacuum tube electronics are analyzed when the time of flight of the electrons is taken into account. The analysis starts with a known current, which in general consists of direct-current value plus a number of alternating-current components. The velocities of the electrons are associated with corresponding current components, and from these velocities the potential differences are computed, so that the final result may be expressed in the form of an impedance.

Applications of the general analysis are made to diodes, triodes with negative grid, and to triodes with positive grid and either negative or positive plate which constitute the Barkhausen type of ultra-high-frequency oscillator. A wave-length range extending from infinity down to only a few centimeters is considered, and it is shown that even in the low-frequency range certain slight modifications should be made in our usual analysis of the negative grid triode.

Oscillation conditions for positive grid triodes are indicated, and a brief discussion of the general assumptions made in the theory is appended.

I. Foreword

THE art of producing, detecting, and modulating ultra-high-frequency electric oscillations has reached the same state of development which was attained in early work on lower frequency oscillations when experiment had outstripped theory. The experimenters were able to produce oscillations by using vacuum tubes, but were not able to explain why. They were able to make improvements by the long and tedious process of cut and try, but did not have the powerful tools of theoretical analysis at their command. In particular, the advantage of the theoretical attack may be illustrated by the rapid advance in technique which followed the theoretical concept of the internal cathode-plate impedance of three-element vacuum tubes. The work of van der Bijl and Nichols showed that for purposes of circuit analysis this path could be replaced by a fictitious generator of voltage, μe_g, having an internal impedance whose magnitude is given by the reciprocal of the slope of the static $V_p - I_p$ characteristic. Development of commercially reliable vacuum tube circuits began forthwith. In a similar, yet less complicated manner, the internal network of two-element tubes may be replaced by an equivalent resistance when relatively low frequencies only are considered.

In these concepts where the vacuum tube is replaced by its equiva-

* Presented in brief summary before U. R. S. I., Washington, D. C., April, 1932. *Proc. I. R. E.*, Vol. 21, No. 11, November, 1933.

lent network impedance, one outstanding feature is exemplified: namely, the separation of the alternating- and direct-current components. The equivalent networks are applicable to the alternating-current fundamental component of the current and differ widely from the direct-current characteristics. A complete realization of the importance of this separation will be of advantage in the later steps where extension of the classical theory to the case of ultra-high-frequency currents is described.

For a short time after the original introduction of the equivalent network of the tube, affairs progressed smoothly. Soon, however, frequencies were increased and a new complication arose. The difficulty was attributable to the interelectrode capacities existing between the various elements of the vacuum tube. The original attempts to take this into account were based on the viewpoint that the tube network should be complete in itself and separate from the external circuit network to which it was attached. Correct results, of course, were obtained by this method but later developments showed the advantage of considering the equivalent network of the complete circuit, including both tube and external impedances in a single network. For instance, by grouping the combination of grid-cathode capacity with whatever external impedance was connected between these two electrodes, a great simplification occurred. This step also has its analogy in the development of ultra-high-frequency relations.

As time went on, higher and higher frequencies were desired, and they were produced by the same kind of vacuum tubes operating in the same kind of circuits, although refinements in circuit and tube design allowed the technique to be improved to the point where oscillations of the order of 70 to 80 megacycles were obtainable with fair efficiency. When the frequency was increased still further, it was found that extension of the same kind of refinements was unavailing in maintaining the efficiency and mode of operation of the higher frequency oscillations at the level which had previously been secured. Ultimately, the three-electrode tube regenerative oscillator ceases to function as a power generator in the neighborhood of 100 megacycles for the more usual types of transmitting tubes. When this point was reached, the external circuit had not yet shrunk up to zero proportions and neither had its losses become sufficiently high to account altogether for the failure of the tube to produce oscillations. From this point on, the old-time cut-and-try methods were employed and marked improvements were secured. In fact, low power tubes have been made which operate at wave-lengths of the order of 50 to 100 centimeters with fair stability, although quite low efficiency.

In the meantime, the production of ultra-high-frequency oscillations had been progressing in a somewhat different direction. The discovery, about 1920, by Barkhausen that oscillations of less than 100 centimeters wave-length could be secured in a tube having a symmetrical structure, when the grid was operated at a fairly high positive potential, while the plate was approximately at the cathode potential, started experiments on what was thought to be an altogether different mode of oscillation. Workers by the score have extended both the experimental technique and the theory of production of this newer type of oscillation. However, one of the results which an analysis of ultra-high-frequency electronics illustrates is that the electron type of oscillator is merely another example of the same kind of oscillation which was produced in the old-time so-called regenerative circuits.

For the purpose of extending the theory of electronics within vacuum tubes to frequencies where the time of transit of the electrons becomes comparable with the oscillation period, it is important at the outset to select an idealized picture which is simple enough to allow exact mathematical relations to be written. At the same time, the picture must be capable of adaptation to practical circuits without undue violence to the mathematics. An example of this kind of adaptation is illustrated by the classical calculation of the amplification factor μ, which was accomplished by consideration of the force of the electrostatic field existing near the cathode in the absence of space charge even though tubes were never operated under this condition. In a like manner, such violations of the ideal must, of necessity, be made in ultra-high-frequency analysis but their practical validity lies in so choosing them that the quantitative error introduced is less than the expected precision of measurement. It becomes, therefore, of the utmost importance to state clearly the transitions which occur between results obtained for the idealized case to which the mathematics is strictly applicable and the practical circuits where the assumptions and approximations are made to conform with operating conditions.

A start has already been made on the problem of developing such a generally valid system of electronics. This was done by Benham [1] who considers a special case comprising two parallel-plane electrodes, one of which is an emitter and the other a collector, when conditions at the emitter are restricted by the assumption that the electrons are emitted with neither initial velocity nor acceleration. This work of Benham's has the utmost importance in a general electronic theory

[1] W. E. Benham, "Theory of the Internal Action of Thermionic Systems at Moderately High Frequencies," Part I, *Phil. Mag.*, p. 641; March (1928); Part II, *Phil. Mag.*, Vol. 11, p. 457; February (1931).

and, in fact, the means of extending his theory exists primarily in the selection of much more general boundary conditions than were assumed by him. It will, therefore, result that some repetition of Benham's work will appear in the following pages. However, in view of the new state of the theory and the importance of accurate foundations for it, this repetition is advantageous rather than otherwise.

With these preliminary remarks in mind, the next step is the selection of the idealized starting point for a mathematical analysis. Exactly as was done by Benham we take two parallel planes of infinite extent, one of which is held at a positive potential V with respect to the other, and between the two electrons are free to move under the influence of the existing fields. The next step in the idealization constitutes the separation of alternating- and direct-current components not only of current and potential, but also of electron velocity, charge density, and electric intensity. With this separation, the restriction that the direct-current component of the electron velocity and acceleration is zero at the negative plane may be made while leaving us free to select much more general boundary conditions for the alternating-current component. It is true that the more general conditions now proposed will not fit the original physical picture where the negative plane consists of a thermionic emitter. Nevertheless the extension is of importance since it allows application to be made to the wide number of physical cases where "virtual cathodes" are formed. One such example is the convergence of electrons toward a plate maintained at cathode potential while a grid operating at a high positive potential with respect to both is interposed between them. In a stricter mathematical sense, the broader boundary conditions come about because of the fact that the general equations containing all components are separable into a system of equations, one for each component, and that the boundary conditions for the different equations of the system are independent of each other.

The concept of an alternating-current velocity component requires a few words of explanation. In the absence of all alternating-current components, electrons leave the cathode with zero velocity and acceleration and move across to the anode with constantly increasing velocity under the well-known classical laws. This velocity constitutes the direct-current velocity component. When the alternating-current components are introduced, there will be a fluctuation in velocity superposed on the direct-current value, and the alternating-current component need not be zero at a virtual cathode. This separation of components will come about naturally in the course of the mathematical analysis which follows, but since the interpretation of the

equations is of paramount importance, a few words of explanation and repetition will be necessary.

II. Fundamental Relations

For the development of the fundamental relations existing between the two parallel planes, we have the classical equations of the electromagnetic theory which may be set down in the following form:

$$\left.\begin{aligned} E &= -\frac{\partial V}{\partial x}, \\ \frac{\partial E}{\partial x} &= 4\pi P, \\ J &= PU + \frac{1}{4\pi}\frac{\partial E}{\partial t}, \end{aligned}\right\} \tag{1}$$

where E is the electric intensity, V the potential, P the charge density, J the total current density consisting of conduction and displacement components, and U is the charge velocity. These equations apply to frequencies such that the time which would be taken by an electromagnetic wave in traveling between the two planes is inappreciable when compared with the period of any alternating-current frequency considered. Ordinarily this limitation will become of importance only at frequencies higher even than those in the centimeter wave-length range where the time of electron transit is of great importance, although the time of passage of an electromagnetic wave is still negligibly small.

An electron situated between the two parallel plates will be acted upon by a force which determines its acceleration. The resulting velocity is a function both of the distance, x, from the cathode and the time, t, so that in terms of partial derivatives, the equation expressing the relation between the force and acceleration is

$$\frac{\partial U}{\partial t} + U\frac{\partial U}{\partial x} = \frac{e}{m}E. \tag{2}$$

From (1) and (2) may readily be obtained

$$\left(\frac{\partial}{\partial t} + U\frac{\partial}{\partial x}\right)^2 U = 4\pi\frac{e}{m}J. \tag{3}$$

In this equation we have a relation between the velocity and the total current density. The advantage of this form of equation for a starting point lies in the fact that the total current density J is not a function

of x. This comes about because of the plane shape and parallel disposition of the electrodes, and the fact that current always flows in closed paths. Thus, while the current between the two planes may be a function of time, it is not a function of x.

The separation of alternating- and direct-current components may now be made. We write

$$J = J_0 + J_1 + J_2 + \cdots \tag{4}$$

with corresponding

$$\left.\begin{aligned} U &= U_0 + U_1 + U_2 + \cdots, \\ V &= V_0 + V_1 + V_2 + \cdots, \end{aligned}\right\} \tag{5}$$

where the quantities with the zero subscript are dependent on x, only, those with subscript 1 are dependent to first order of small quantities upon time, those with subscript 2 are dependent to second order, and so forth. As a result of this separation in accord with the order of dependents upon time, (3) may be split up into a system of equations, the first of which expresses the relation between U_0, J_0, and x and does not involve time. This is the relation governing the direct-current components. The second equation of the system involves the relation between U_1, J_1, x, and time, and contains U_0 which was determined by the first equation. Likewise, the third equation contains U_2, U_1, J_2, x, and t. Since the series given by (4) and (5) are convergent so that, in general, the terms with higher order subscripts are smaller than those with lower subscripts, we may consider that, at least for small values of alternating-current components, the total fundamental frequency component is given by the terms with unity subscript.

The first two equations of the system are as follows:

$$U_0 \frac{\partial}{\partial x}\left(U_0 \frac{\partial U_0}{\partial x} \right) = 4\pi \frac{e}{m} J_0, \tag{6}$$

$$\left(\frac{\partial}{\partial t} + U_0 \frac{\partial}{\partial x} \right)\left(\frac{\partial U_1}{\partial t} + U_0 \frac{\partial U_1}{\partial x} + U_1 \frac{\partial U_0}{\partial x} \right) + U_1 \frac{\partial}{\partial x}\left(U_0 \frac{\partial U_0}{\partial x} \right) = 4\pi \frac{e}{m} J_1. \tag{7}$$

In the solution of (6), the boundary conditions are restricted so that when x is zero, the velocity and acceleration both are zero. These restrictions mean that initial velocities are neglected, and that complete space charge is assumed. Thus the solution for U_0 is

$$U_0 = \alpha x^{2/3}, \tag{8}$$

where

$$\alpha = \left(18\pi \frac{e}{m} J_0 \right)^{1/3}. \tag{9}$$

The solution of (7) is more complicated. We assign a particular value to J_1, namely, $J_1 = A \sin pt$ and find the corresponding value of U_1. To do this, it is convenient to change the variable x to a new variable ξ, which will be called the transit angle. This new variable is equal to the product of the angular frequency p and the time τ which it would take an electron moving with velocity U_0 to reach the point x and is given as follows:

$$\xi = p\tau = \frac{3p}{\alpha} x^{1/3}. \tag{10}$$

Upon changing the dependent variable from U_1 to ω, where $U_1 = \omega/\xi$, we find from (7)

$$\left(\frac{\partial}{\partial t} + p \frac{\partial}{\partial \xi} \right)^2 \omega = \xi\beta \sin pt, \tag{11}$$

where

$$\beta = 4\pi \frac{e}{m} A.$$

This has the solution

$$U_1 = -\frac{\beta}{p^2} \left[\sin pt + \frac{1}{\xi} \cos pt + F_1(\xi - pt) + \frac{1}{\xi} F_2(\xi - pt) \right]. \tag{12}$$

This equation contains two arbitrary functions of $(\xi - pt)$ which must be evaluated by the boundary conditions selected for U_1. Thus the boundary conditions for the alternating-current component make their first appearance.

From the form of (7) which is linear in U_1, it is evident that U_1 must be a sinusoidal function of time having an angular frequency p in order to correspond with the form of J_1. It follows, then, that the most general form which can be assumed for the steady state functions F_1 and F_2 is as follows:

$$\left.\begin{aligned} F_1(\xi - pt) &= a \sin (\xi - pt) + b \cos (\xi - pt) \\ F_2(\xi - pt) &= c \sin (\xi - pt) + d \cos (\xi - pt) \end{aligned}\right\}. \tag{13}$$

Now for the boundary conditions. As pointed out, there is no mathematical necessity for the boundary conditions imposed upon U_1 to correspond with those which were imposed upon U_0. At an actual cathode consisting of an electron emitting surface it would be appropriate to assume that the initial velocities are in no way dependent upon the current, but we shall have to deal not only with actual

cathodes, but also with virtual [2] cathodes where the assumption of zero alternating-current velocity and acceleration is unwarranted. Such a virtual cathode might occur, for instance, between a grid operated at a positive direct-current potential and a plate nearly at cathode potential. If enough electrons came through the mesh of the grid to depress the potential until it became practically zero at some point in the space between grid and plate, the direct-current boundary conditions of zero velocity and acceleration of electrons would be fulfilled at that point. The general equations for the alternating current will therefore apply when the origin is taken at the point of direct-current potential minimum which forms the virtual cathode, and when all of the electrons which are emitted by the actual cathode pass by the virtual cathode and reach the plate. In the event that some of the electrons are turned back at the virtual cathode and move again toward the grid, as indeed they all do when the plate is at a negative potential, a change in the form of the general equation is necessary, and will be described in the sections dealing particularly with positive grid triodes. This change, however, affects merely the form of the equations and not the physical arguments underlying the selection of boundary conditions, which are the same whether all the electrons reach the plate or whether some or all of them turn back toward the grid.

If the alternating-current velocity is determined by small variations in grid potential, let us say, it is evident that no additional assumptions save the requirement that the velocity must not become infinite may be made concerning its value at the virtual cathode. Consequently, a quite general set of boundary conditions will suffice to determine the quantities, a, b, c, d, which appear in (13) and thus completely determine U_1.

Since there are two arbitrary functions in (12), two boundary conditions will be needed. Further inspection shows that the stipulation that the alternating-current velocity be finite at the origin is sufficient to furnish one of these boundary conditions. For the other, a knowledge of the value of the alternating-current velocity at any point between the two reference planes is sufficient. Thus, if at a particular value of ξ, say ξ_1, we know that U_1 is equal to $M \sin pt + N \cos pt$, we have enough information to calculate its value at all other points between the two planes. For example, the two reference planes might be the grid and plate of a positive grid triode. In this event, the alternating-current velocity at the grid could be calculated at the grid plane by means of conditions between there and the cathode.

[2] E. W. B. Gill, "A Space-Charge Effect," *Phil. Mag.*, Vol. 49, p. 993 (1925).

In mathematical form the two boundary conditions may be set forth as follows:

when,

$$\xi = 0, \qquad U_1 \text{ must be finite,} \tag{14}$$

$$\xi = \xi_1, \qquad U_1 = M \sin pt + N \cos pt. \tag{15}$$

From (12) and (13) these result in the values:

$$c = 0, \qquad d = -2,$$

$$a = \frac{p^2}{\beta}(M \cos \xi_1 - N \sin \xi_1) + \cos \xi_1 - \frac{2}{\xi^1} \sin \xi_1, \tag{16}$$

$$b = \frac{2}{\xi_1}(1 - \cos \xi_1) - \sin \xi_1 - \frac{p^2}{\beta}(M \sin \xi_1 + N \cos \xi_1). \tag{17}$$

Thence from (12) we have for the alternating-current velocity, in general,

$$U_1 = (M + iN)(\cos \xi_1 + i \sin \xi_1)(\cos \xi - i \sin \xi) + \frac{\beta}{p^2}\left[\left\{\left(\cos \xi_1 - \frac{2}{\xi_1} \sin \xi_1\right) - i\left(\frac{2}{\xi_1} - \frac{2}{\xi_1} \cos \xi_1 - \sin \xi_1\right)\right\} (\cos \xi - i \sin \xi) - \left(1 - \frac{2}{\xi} \sin \xi\right) - i \frac{2}{\xi}(1 - \cos \xi)\right], \tag{18}$$

where, in accord with engineering practice, complex notation is employed, so that sin pt has been replaced by e^{ipt} and cos pt has been replaced by ie^{ipt}, where $i = \sqrt{-1}$.

The first step in the derivation of fundamental relations has now been achieved. The alternating-current velocity at any point between the two planes has been expressed in terms of the alternating-current velocity, $M + iN$, existing at a definite value of x, say x_1, corresponding to the transit angle ξ_1.

The next step is a determination of the potentials corresponding to the velocities U_0 and U_1, respectively. Thus from (1) and (2)

$$-\frac{e}{m}\frac{\partial V}{\partial x} = \frac{\partial U}{\partial t} + U\frac{\partial U}{\partial x} \tag{19}$$

and then with the separation of components as given by (5)

$$-\frac{e}{m}\frac{\partial V_0}{\partial x} = U_0 \frac{\partial U_0}{\partial x}, \tag{20}$$

$$-\frac{e}{m}\frac{\partial V_1}{\partial x} = \frac{\partial U_1}{\partial t} + \frac{\partial}{\partial x}(U_0 U_1). \tag{21}$$

The solution of (20) is

$$V_0 = -\frac{m}{2e} U_0^2 = -\frac{m}{2e} \alpha^2 x^{4/3}, \tag{22}$$

which is the well-known classical relation between the potential, the current, and the position between two parallel planes where complete space charge exists. The complete space-charge condition is postulated by the boundary conditions selected for U_0 and the implications involved are discussed by I. Langmuir and Karl T. Compton.[3]

The alternating-current component of the potential is obtained by integration of (21) as follows:

$$-\frac{e}{m} V_1 = \frac{\partial}{\partial t} \int U_1 \partial x + U_0 U_1 + f(t), \tag{23}$$

whence, from (18), and in complex notation

$$\begin{aligned} V_1 = & -\frac{2m}{e} \frac{\alpha^3}{9p^2} (M + iN)(\cos \xi_1 + i \sin \xi_1)[(\xi \sin \xi + \cos \xi) \\ & + i(\xi \cos \xi - \sin \xi)] \\ & - \frac{2m\alpha^3\beta}{e9p^4} \left[\left\{ \left(\cos \xi_1 - \frac{2}{\xi_1} \sin \xi_1 \right) - i \left(\frac{2}{\xi_1} - \frac{2}{\xi_1} \cos \xi_1 - \sin \xi_1 \right) \right\} \right. \\ & [(\xi \sin \xi + \cos \xi) + i(\xi \cos \xi - \sin \xi)] \\ & \left. - \cos \xi - i(\xi + \tfrac{1}{6}\xi^3 - \sin \xi) \right] + \text{constant}. \end{aligned} \tag{24}$$

With the attainment of (24), the fundamental relation between the alternating-current component J_1 and the alternating-current potential V_1 in the idealized parallel plate diode has been secured. In a more general sense the equation is applicable between any two fictitious parallel planes where one is located at an origin where the boundary conditions for U_0 are satisfied; namely, that the direct-current components of the velocity and acceleration are zero, and the value of the alternating-current velocity at a point, x_1, corresponding to the transit angle, ξ_1, is given by $M \sin pt + N \cos pt$, or by $M + iN$ in complex notation.

Equation (24) contains an additive constant which always appears in potential calculations. This constant disappears when the potential difference is computed. For instance, suppose the potential difference between planes where ξ has the values ξ and ξ', respectively, is desired.

[3] I. Langmuir and Karl T. Compton, "Electrical Discharges in Gases"—Part II, *Rev. Mod. Phys.*, Vol. 3, p. 191; April (1931).

We have

$$V_1 = f(\xi) + \text{constant},$$

$$V_1' = f(\xi') + \text{constant},$$

so that

$$V_1 - V_1' = f(\xi) - f(\xi'). \tag{24-a}$$

Since the potential difference is always required rather than the absolute potential, (24-a) gives the means for applying (24) to actual problems.

III. Application to Diodes

In the application of the fundamental relations to diodes where the thermionic emitter forms the plane located at the origin and the anode coincides with the other plane, the boundary condition is that U_1 shall be zero at the cathode. This means that both M and N are zero and that ξ_1 is also zero. The resulting forms taken by (18) and (24-a), respectively, are as follows:

$$U_1 = -\frac{\beta}{p^2}\left[\left(1 + \cos\xi - \frac{2}{\xi}\sin\xi\right) + i\left(\frac{2}{\xi} - \sin\xi - \frac{2}{\xi}\cos\xi\right)\right], \tag{25}$$

$$V_1 - V_1' = \frac{2m\alpha^3\beta}{e9p^4}\left[(2\cos\xi + \xi\sin\xi - 2) + i(\xi + \tfrac{1}{6}\xi^3 - 2\sin\xi + \xi\cos\xi)\right]. \tag{26}$$

These two equations are identical with those obtained by Benham,[1] and graphs are given in Figs. 1 and 2 showing their variation as a function of the transit angle ξ. In particular, the equivalent impedance between unit areas of the two parallel planes may be found from (26). It must be remembered that the current, A, was assumed positive when directed away from the origin. Hence, we may write

$$Z = -\frac{V_1 - V_1'}{A}. \tag{27}$$

Moreover, the coefficient outside the square brackets in the equation may be expressed more simply when it is realized that the low-frequency internal resistance of a diode is given by the expression

$$r_0 = -\frac{\partial V_0}{\partial J_0}, \tag{28}$$

the minus sign again appearing because of the assumed current direction. Consequently, under the condition of complete space charge,

we have from (22)

$$\frac{2m\alpha^3\beta}{e9p^4} = \frac{12r_0A}{\xi^4} . \tag{29}$$

In addition to the graphs in Figs. 1 and 2 showing the real and imaginary components of impedance and velocity, the graphs shown in Figs. 3 and 4 give their respective magnitudes and phase angles.

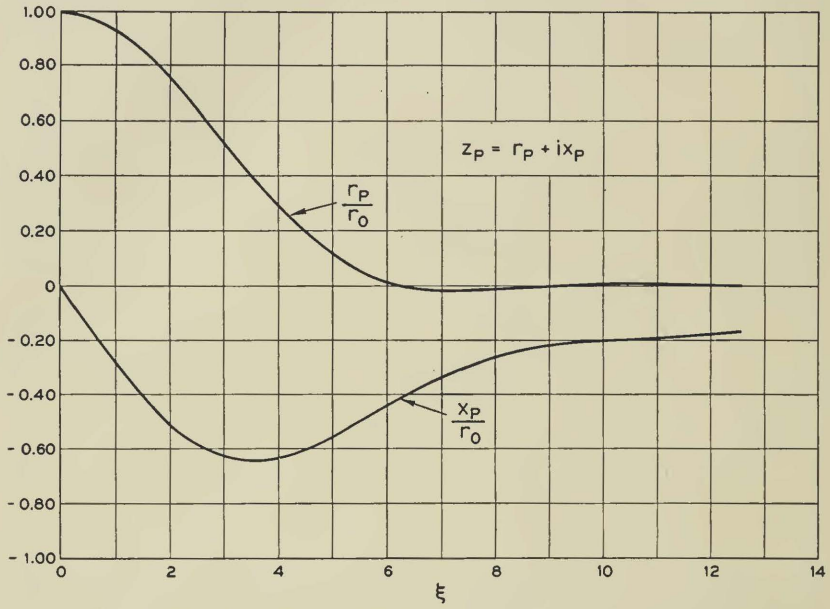


Fig. 1—Plate impedance of diodes or of negative grid triodes as a function of electron transit angle.

The impedance charts show a negative resistance for diodes in the neighborhood of a transit angle, ξ, of 7 radians. The possibility of securing oscillations in this region has been discussed by Benham, so that only a few additional remarks will be made here.

The magnitude of the ratio of reactance to resistance is about 15 when the transit angle is 7 radians. This means that oscillation conditions require an external circuit having a larger ratio of reactance to resistance. On account of the high value of reactance required, a tuned circuit or Lecher-wire system is needed, which would have to operate near an antiresonance point in order to supply the high reactance value. But the resistance component of the external circuit impedance is large at frequencies in the neighborhood of the tuning point, so that the ratio of reactance to resistance is small. Calcula-

tions show that the possibility of securing external circuits having low enough losses to meet the oscillation requirements of most of the diodes which are at present available is not very favorable. The large radio-frequency loss in the filamentary cathodes with which many tubes are supplied is an additional obstacle to be overcome before satisfactory ultra-high-frequency operation of diodes can be expected.

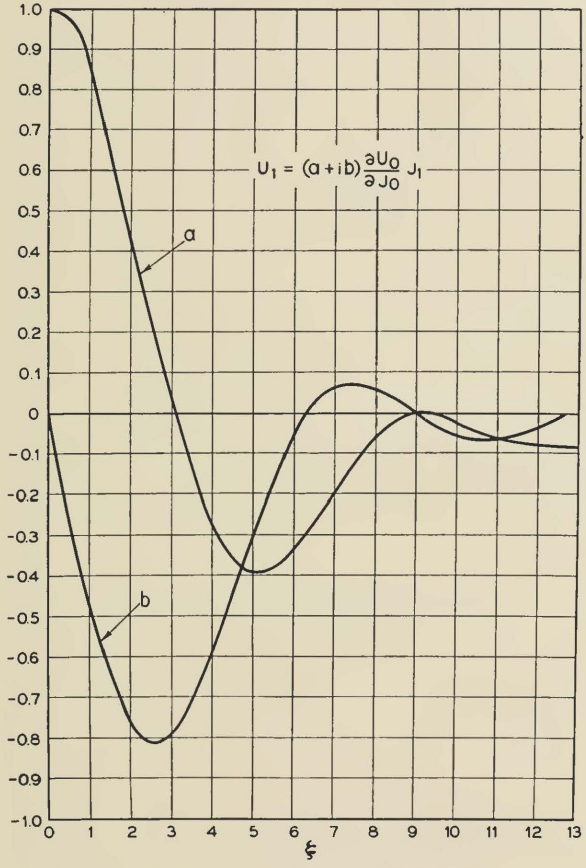


Fig. 2—Electron velocity fluctuation in diodes versus transit angle.

IV. Triodes with Negative Grid and Positive Plate

In the application of the fundamental relations to triodes operating with the grid at a negative potential, the problem becomes more complicated because of the several current paths which exist within the tube. Moreover, the direct-current potential distribution is disturbed in a radical way by the presence of the negative grid. In fact, the

negative grid triode in some respects offers greater theoretical difficulty than does the positive grid triode, which is treated in the next section. However, because of the greater ease in the interpretation of the results in terms which have become familiar through years of use, the negative grid triode is treated first.

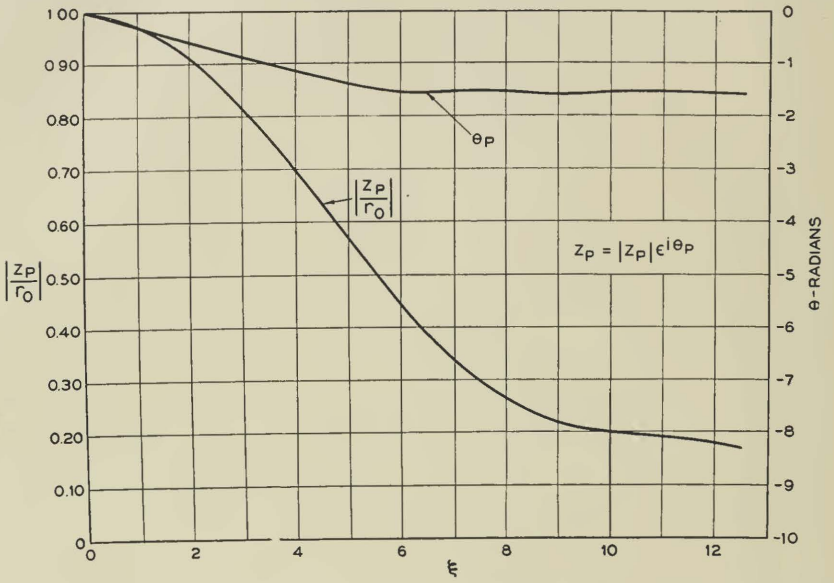


Fig. 3—Magnitude and phase angle of plate impedance of diodes or of negative grid triodes versus transit angle.

In the analysis recourse must be had to approximations and idealizations which allow the theory to fit the practical conditions. In the selection of these, the first thing to notice is that no electrons reach the grid, so that most of the electrostatic force from the grid acts on electrons quite near the cathode, where the charge density is very great. The most prominent effect of a change in grid potential will thus be a change in the velocity of electrons at a point quite near the cathode. It will thus be appropriate to assume as a starting point that the alternating-current velocity at a point x_1, located quite near the cathode is directly proportional to the alternating-current grid potential, V_g, so that we may write,

when $$\xi = \xi_1,$$

$$U_1 = (M + iN) = kV_g. \tag{30}$$

In any event, this relation may be justified if the factor of proportionality, k, be allowed to assume complex values, and ξ_1 is not taken too near the origin. Actually, the electron-free space surrounding the grid wires, and the fact that the electric intensity at a point midway between any two of the wires is directed perpendicularly to the plane of the grid, gives us more confidence in extending the approximation, so that k will be regarded as real, and ξ_1 will be taken very small.

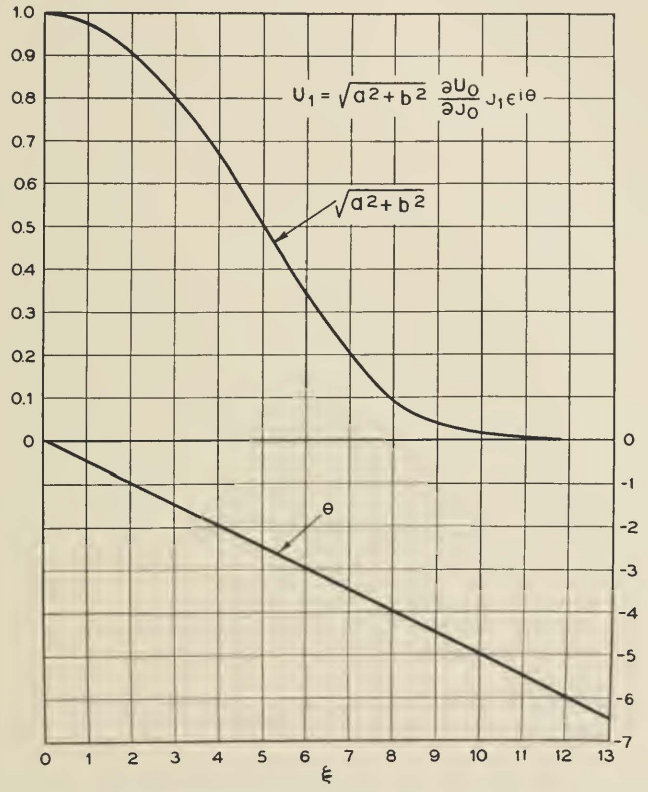


Fig. 4—Magnitude and phase angle of electron velocity fluctuation in diodes versus transit angle.

Equation (24) may, therefore, be applied under the conditions that $\xi_1 \to 0$, and gives the following for the potential difference between plate and cathode:

$$V_p = -\frac{12 r_0 A}{\xi^4}\left[(\xi \sin \xi + 2\cos\xi - 2) + i(\xi + \tfrac{1}{6}\xi^3 - 2\sin\xi + \xi\cos\xi)\right.$$
$$\left. -(M+iN)\frac{p^2}{\beta}\left[(\xi\sin\xi + \cos\xi - 1) - i(\sin\xi - \xi\cos\xi)\right]\right]. \quad (31)$$

This equation may be written in condensed form with the aid of (30)

$$V_p = J_1(r + ix) - V_g(\mu + i\nu), \tag{32}$$

where

$$\left.\begin{aligned} r &= -\frac{12r_0}{\xi^4}(\xi \sin \xi + 2\cos \xi - 2), \\ x &= -\frac{12r_0}{\xi^4}(\xi + \tfrac{1}{6}\xi^3 - 2\sin \xi + \xi \cos \xi), \\ \mu &= \frac{2\mu_0}{\xi^2}(\xi \sin \xi + \cos \xi - 1), \\ \nu &= \frac{2\mu_0}{\xi^2}(\xi \cos \xi - \sin \xi). \end{aligned}\right\} \tag{33}$$

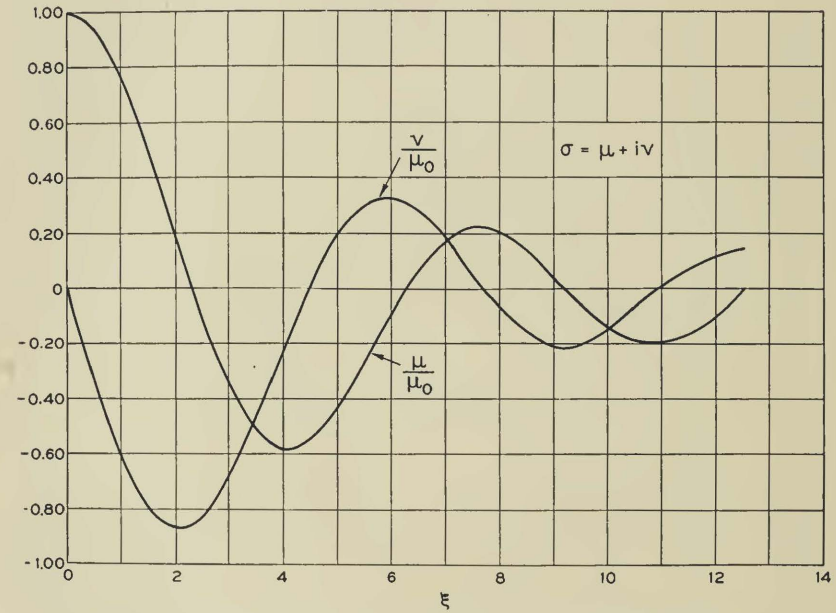


Fig. 5—Real and imaginary components of complex amplification factor of negative grid triodes versus transit angle.

The significance of (32) is at once apparent when it is compared with the classical form of the equation representing the alternating-current plate voltage, namely,

$$V_p = I_p r_0 - \mu V_g.$$

The plate resistance r_0 has now become complex as likewise has the amplification factor μ. Values of the plate impedance

$$z_p = r + ix$$

are the same as those obtained for the diode and are plotted in Figs. 1 and 3. Values of the amplification factor

$$\sigma = \mu + i\nu$$

are shown in Figs. 5 and 6.

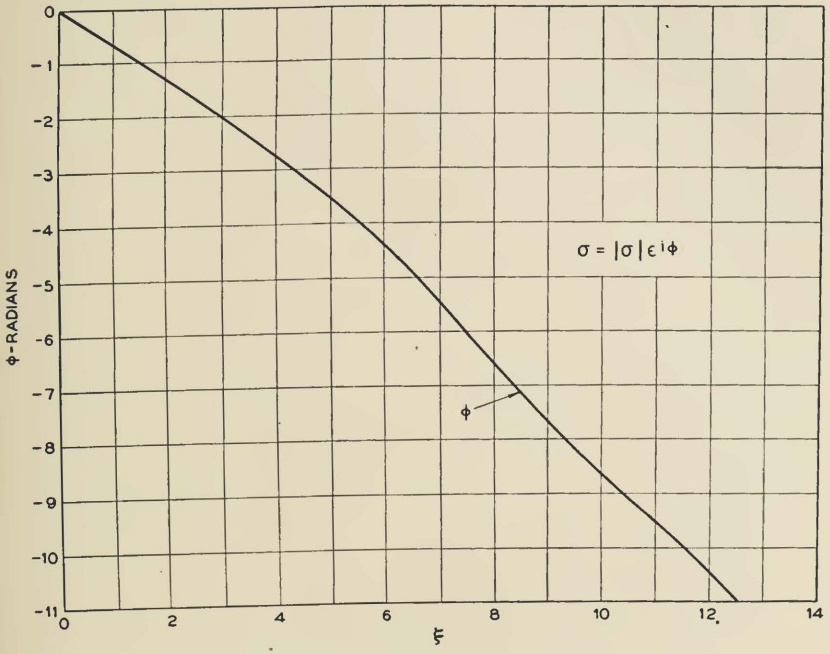


Fig. 6—Phase angle of amplification factor of negative grid triodes versus transit angle.

It is evident that radical changes in the phase angles existing between the grid voltage and plate current are present when the transit time becomes appreciable in comparison with the period of the applied electromotive force. The plate impedance decreases in magnitude as also does the magnitude of the amplification factor. However, the ratio of the two, namely, σ/z, maintains a fairly constant magnitude as shown in Fig. 7, whose phase angle nevertheless rotates continually in a negative direction becoming equal to 3 radians when ξ is 2π.

The interelectrode capacity between cathode and plate is included in the fundamental relations here employed. This inclusion exhibits one important difference between (32) and the classical case. At low frequencies, the equivalent circuit represented by (32) degenerates into that shown on Fig. 8. The capacity branch exists in parallel with the

resistive branch and they are both in series with the effective generator σe_g, whereas in the classical picture the capacity branch shunts the effective generator and plate resistance which are in series with each other. Practically the difference between the two equivalent circuits is negligible except at extremely high frequencies. The following physical viewpoint supports the newer picture.

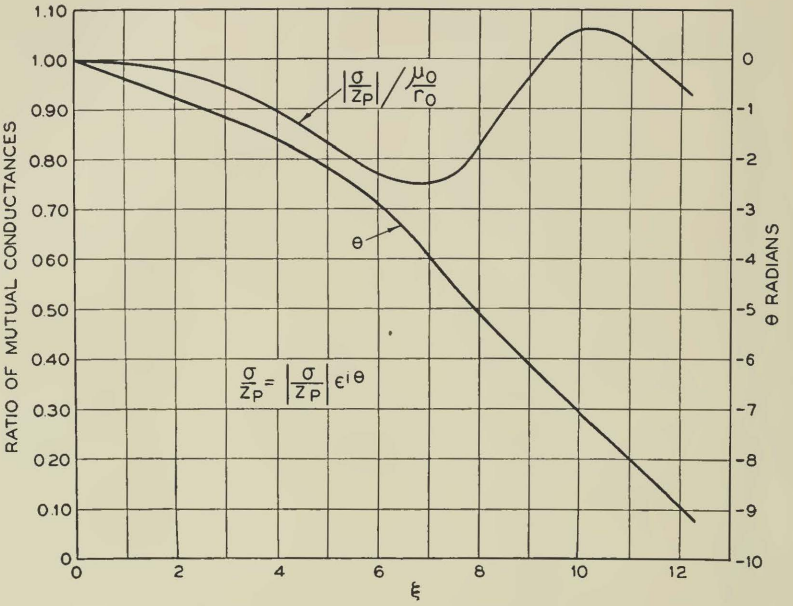


Fig. 7—Magnitude of complex mutual conductance of negative grid triodes versus transit angle.

As pointed out, the action of the grid is exerted mostly on the region of dense space charge existing very near the cathode and variations in the grid potential act on the velocities of the emerging electrons, thus producing the equivalent generator of the plate circuit. The plate current consists of conduction and displacement components whose sum is the same at all points in the cathode-plate path. Near the cathode, the conduction component comprises the whole current because of the high charge density and the effective generator acts in series with this current and hence in series with the path of the displacement current into which the character of the total current gradually changes as the plate is approached.

Strictly speaking, the equivalent circuit corresponding to (32) exists, not between the plate and cathode, but between the plate and the potential minimum near the cathode which is caused by the finite

velocities with which electrons are emitted from the cathode. Practically, the difference is negligible except at extremely high frequencies. Since the impedance between the cathode and potential minimum is small compared to the plate impedance, its effect is merely to add a loss to the system which increases with frequency since the plate impedance approaches a capacity as the frequency approaches infinity.

The grid cathode path presents less difficulty, although a somewhat less rigorous treatment is given here. As pointed out, the force from the grid acts on the high charge density region existing near the potential minimum. The impedance between cathode and grid, therefore, consists of two parts in series; namely, capacity between grid and potential minimum and impedance between potential minimum and cathode, the latter part of this impedance being common both to plate- and grid-current paths.

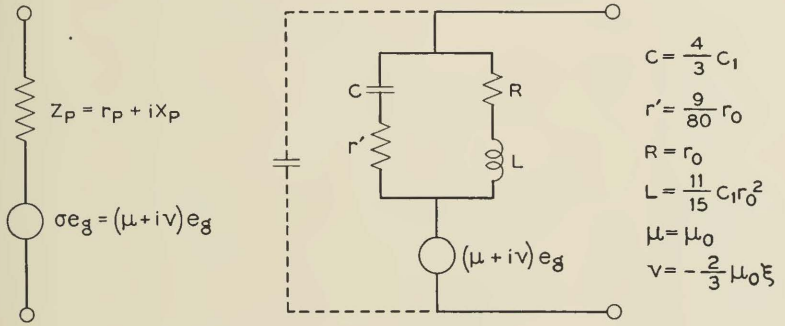


Fig. 8—Equivalent network of plate-cathode path of negative grid triodes for transit angles less than 0.3 radian.

If we were to connect the grid and cathode terminals of such a triode to a capacity bridge and measure the capacity existing there when the tube was cold and when the cathode was heated, we should find that the capacity would exhibit a slight increase in the latter case. The reason for this increase may best be explained by noting that in the cold condition the electrostatic force from the grid is exerted on the cathode itself, whereas in the heated state, the force acts on the electrons near the potential minimum, thus resulting in an increased capacity in series with a resistive component.

In some measurements of the losses in coils which were made at a frequency of 18 megacycles, J. G. Chaffee of the Bell Telephone Laboratories has found that a loss existed between grid and cathode of vacuum tubes which was much greater than can be accounted for by any of the dielectrics used and which was present only when the tube

filament was hot. This loss increased with frequency in the manner characteristic of that of the capacity-resistance combination between cathode and grid which was described above. Present indications are that, at least in part, the loss may be ascribed to the resistance existing between the cathode and the region of potential minimum.

Of the three current paths through the tube, one more still remains to be considered. This is the grid-plate path. The relations involved here are more readily seen by considering first a low-frequency example. Here the electron stream passes through the spaces between grid wires, afterward diverging as the plate is approached. Electrostatic force from the grid acts not only on the plate but also on the electrons in the space between. It is evident, then, that the path which, when the cathode was cold, constituted a pure capacity changes into an effective capacity different from the original in combination with a resistive component. The losses would be expected to increase with frequency just as they did in the grid-cathode type. The change in grid-plate impedance is particularly noticeable when it is attempted to adjust balanced or neutralized amplifier circuits with the filament cold, in which case the balance is disturbed when the cathode is heated.

As yet, no accurate expression for this grid-plate impedance has been obtained, either at the low frequencies where transit times are negligible or at the higher frequencies now particularly under investigation. The reason for this lies in the repelling force on the electron stream of the negative grid so that the assumption of current flow in straight parallel lines is not valid in so far as current from the grid to the plate is involved.

It has been shown that both the cathode-grid path and the grid-plate path contain resistive components with corresponding losses which increase with increase of frequency. This loss may be cited as a reason why triodes with negative grids cease to oscillate at the higher frequencies. If it were not for these losses, external circuits could be attached to the tube having such phase relations as to satisfy oscillation conditions, so that the negative grid triode could be utilized in the range which is now covered by the triode with positive grid.

V. Triodes with Positive Grid and Slightly Positive Plate

When the grid of a three-element tube is operated at a high positive potential with respect both to cathode and plate, electrons are attracted toward the grid, and the majority of them are captured on their first transit. Those which pass through the mesh and journey toward the plate will be captured by the plate if its potential is sufficiently positive with respect to the cathode.

In general, space-charge conditions existing between grid and plate are quite complicated. An analysis has been made by Tonks [4] which indicates several distinct classes of space-charge distribution which are possible. In the first place so few electrons may pass the grid mesh that no appreciable space charge is set up between there and the plate. In this instance a positive plate will trap them all, whereas a negative plate will return them all toward the grid. Second, with a fixed positive plate potential an increase in the number of electrons which pass the grid mesh will result in a depression of the potential distribution as illustrated at (*a*) by the curves in Fig. 9. This depression will con-

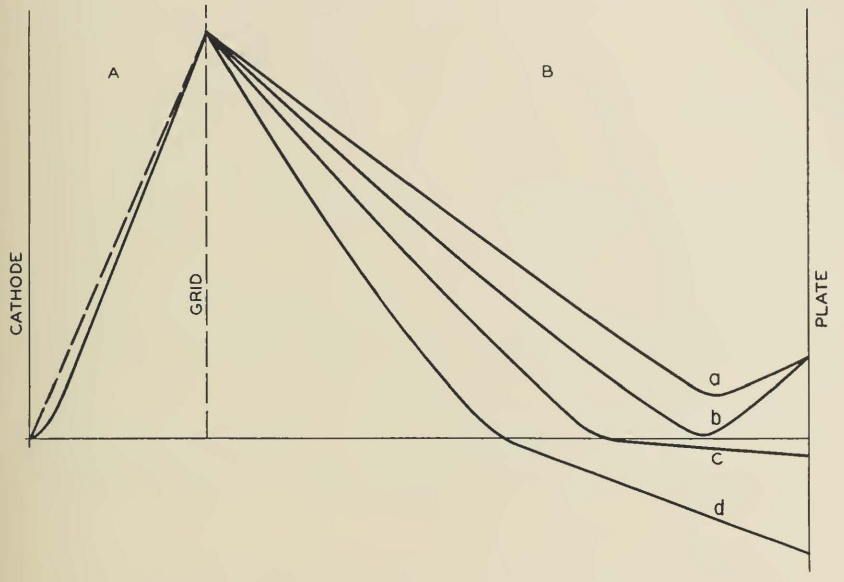


Fig. 9—Potential distributions in positive grid triodes.

tinue to increase until a potential minimum is formed. When this potential minimum becomes nearly the same as that of the cathode, either of several things may occur. If the minimum is just above the cathode potential, all electrons will pass that point and eventually reach the plate. However, an extremely small increase in the number of electrons will cause the potential minimum to become equal to the cathode potential. When this happens some of the electrons will be turned back and travel again toward the grid. These will increase the charge density existing and, therefore, cause a further depression in the potential resulting in a mathematical discontinuity so that the

[4] L. Tonks, "Space Charge as a Cause of Negative Resistance in a Triode and Its Bearing on Short-Wave Generation," *Phys. Rev.*, Vol. 30, p. 501; October (1927).

curve of the potential suddenly changes its shape with a resulting change in plate current. Again, the plate may be operated at a negative potential. In this case, none of the electrons will reach it and the potential distribution curves have the character illustrated at (*c*) and (*d*) in Fig. 9.

In attempting to apply the fundamental relations to this grid-plate region, we must choose our origin at a point where the potential distribution curve touches the zero axis and is tangent to it. Whenever such a point exists, the relations may be applied as described below. Even when this condition does not exist inside the vacuum tube, there may be a virtual cathode existing outside of the plate.

Whenever all of the electrons passing the grid reach the plate the general equations may be applied in a straightforward manner with the origin taken at the virtual cathode. Whenever some of the electrons are turned back toward the grid, slightly different equations are required, although they may be applied in the same manner. These modified equations will be derived and discussed after the application of the equations already derived has been made to the case where all of the electrons reach the plate.

Choosing the origin for this latter case at the point of zero potential or virtual cathode, we can compute the impedance between the grid-plane and the virtual cathode when we know the alternating-current velocities with which the electrons pass through the grid-plane. This has been found for the condition of complete space charge between cathode and grid and was given by (25). Likewise, it can be found on the supposition that no space charge exists in the cathode-grid region and the result will be calculated later. Thus, two limiting cases are available for numerical application.

In order to prevent confusion for the grid—virtual-cathode region where the electron flow is toward the origin rather than away from it, as was assumed in the derivation of the fundamental relations, it will be convenient to change the symbol for transit angle from ξ to $-\zeta$. This will automatically take care of all algebraic signs, currents and velocities now being considered positive when directed towards the origin.

Since we are computing the impedance between an origin at the virtual cathode and the grid plane we may apply (24) to find the potential difference, getting

$$V_1 - V_1' = -\left(\frac{2m}{e}\frac{\alpha^3}{9p^2}\right)(M + iN)[(1 - \cos\zeta) - i(\zeta - \sin\zeta)] - i\left(\frac{2m}{e}\frac{\alpha^3\beta}{9p^4}\right)\left(\frac{1}{6}\zeta^3 - \frac{4}{\zeta}(1 - \cos\zeta) + 2\sin\zeta\right), \quad (34)$$

where V_1' is the potential at the virtual cathode.

This relation is of the form

$$V_g - V_p = -(M+iN)\left(\frac{2m}{e}\frac{\alpha^3}{9p^2}\right)[(1-\cos\zeta)-i(\zeta-\sin\zeta)]+J_pZ_p, \quad (35)$$

where J_p is the plate current, and Z_p is the effective impedance:

$$Z_p = -i\left(\frac{2m}{e}\frac{\alpha^3\beta}{9p^4A}\right)\left(\frac{1}{6}\zeta^3 - \frac{4}{\zeta}(1-\cos\zeta) + 2\sin\zeta\right). \quad (36)$$

In terms of the cold capacity C_1 between plate and grid plane this becomes

$$Z_p = -\frac{i}{pC_1}\frac{6}{\zeta^3}\left[\frac{1}{6}\zeta^3 - \frac{4}{\zeta}(1-\cos\zeta) + 2\sin\zeta\right],$$

which is plotted in Fig. 10.

The form of (35) shows that the equivalent network between the plane of the grid and the plate may be represented by an equivalent generator acting in series with the impedance, Z_p. This is evidenced by the fact that the velocity $M + iN$ with which the electrons pass the grid, may be expressed in terms of the grid potential V_g by means of conditions between the grid and cathode. When complete space charge exists near the cathode, these conditions are expressed by (25) and (26). On the other hand, tubes with positive grid are sometimes operated with inappreciable space charge between grid and cathode. In this event, a similar analysis leads to values for the alternating-current velocity and potential at the grid as follows:

$$U_1 = M + iN = -\frac{\beta}{p^2}\left[\left(\frac{\eta-\sin\eta}{\eta}\right) + i\left(\frac{1-\cos\eta}{\eta}\right)\right], \quad (37)$$

$$V_g = i\frac{4\pi x}{p}A = \frac{iA}{pC}, \quad (38)$$

where η is the transit angle in the absence of space charge, and C is the electrostatic capacity between unit area of cathode and of grid plane. The right-hand side of (38) does not contain a minus sign because of the assumed current direction which is away from the cathode, as is also the convention employed in (25) and (26) where the electron charge e is a positive number.

The relations given by (35) allow the potential difference between grid and plate to be determined in terms of the total current flowing to the plate, and the total current flowing from the cathode, which ap-

pears in the velocity factor $M + iN$. In the usual case some of the alternating current flows to the grid wires and is returned through an external circuit connected to the grid. If the impedance between grid and plate is desired it is necessary to find the relation which this grid current bears to the total cathode and plate currents and to the alternating-current potentials. The calculations involved are extremely complicated because the assumption of current flow in straight lines between parallel planes is far from representing the actual conditions

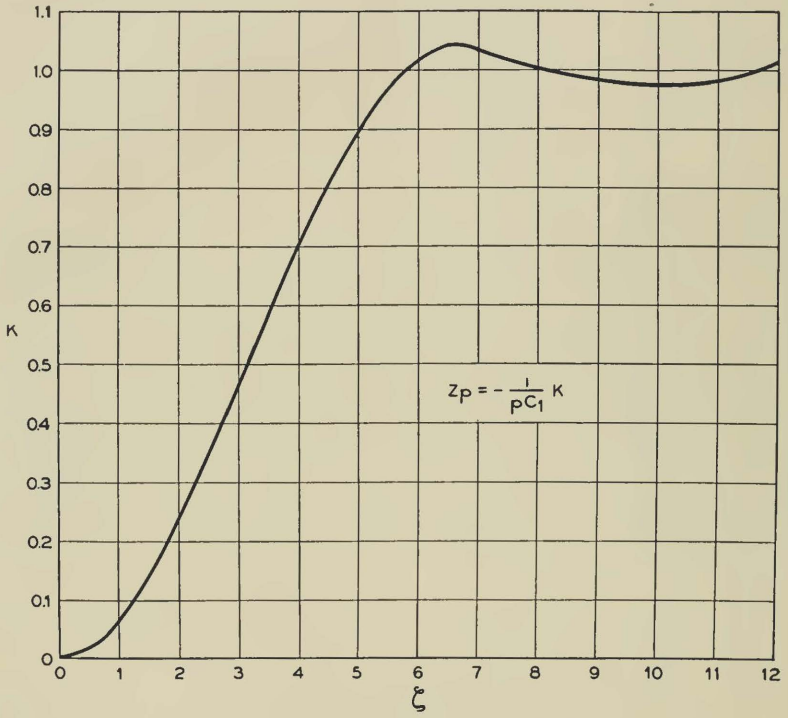


Fig. 10—Plate impedance of positive grid triodes with slightly positive plate.

in the immediate neighborhood of the grid wires. Rather than attempting an analysis of these conditions at the present time, we shall content ourselves with results already obtained, since they are applicable to the special case, which can be realized approximately in experiment, where the grid is connected to a radio-frequency choke coil of sufficiently good characteristics to prevent it from carrying away any alternating current. For this special case the current J_1 is the same both in the cathode region and in the plate region, and all encumbering assumptions involving different paths for the conduction and displace-

ment components of the current in the neighborhood of the grid wires have been done away with.

The application of the equations to this special case is dealt with in the section of this paper devoted to positive grid oscillators. Before these oscillators can be treated comprehensively, a further extension of fundamental theory is necessary. This extension comes about because positive grid oscillators are often operated with a slightly negative potential applied to the plate.

VI. Triodes with Positive Grid and Negative Plate[5]

When consideration is directed to tubes operating with positive grid but negative plate, the fundamental underlying theory must again be investigated. The reason for this lies in the fact that all electrons which penetrate through the meshes of the grid are turned back before they reach the plate, so that in the grid-plate space there are two streams of electrons moving in opposite directions. The effect of this double value for the velocity may readily be calculated in so far as direct-current components, only, are concerned. We have merely to note that the charge density is double the value which it would have in the presence of those electrons which are moving in one direction, only, so that the correct relations are obtained from the equations already derived by taking twice the value of direct current in one direction.

When alternating-current components are considered, however, matters are more complicated, but not difficult. To see what the actual relations are, let there be two possible values at any point for the instantaneous velocity, and call these two values U_a and U_b, respectively. Then the relation between force and acceleration becomes

$$\frac{eE}{m} = \frac{dU_a}{dt} = \frac{dU_b}{dt}. \tag{39}$$

Hence, at a given value of x we have by integration

$$U_a = U_b + \text{constant}.$$

But, when both values of velocity are separated into their components according to (5) we have from (39)

$$U_{a0} + U_{a1} + \cdots = U_{b0} + U_{b1} + \cdots + \text{constant}.$$

[5] Since the publication of this paper in the *Proceedings* of the *Institute of Radio Engineers*, several questions have been raised regarding the treatment presented in sections VI and VII. These are being investigated and will form the basis of another paper.

By equating corresponding terms, we find

$$\left.\begin{aligned} U_{a0} &= U_{b0} + \text{constant}, \\ U_{a1} &= U_{b1}, \\ U_{a2} &= U_{b2}, \text{ etc.} \end{aligned}\right\} \tag{40}$$

The first of these equations is trivial when the boundary conditions are inserted, for then it appears that $U_{a0} = -U_{b0}$ and the equation merely states that at a given value of x the direct-current velocity component is not a function of time.

The second equation is much more enlightening and tells us that although two values of the direct-current velocity may be present, nevertheless there is only a single value for the alternating-current component. The same conclusion holds for the higher order velocity components. This conclusion supplies the key for the solution of the general equations when applied to the stream of electrons moving in both directions between the grid and plate of the tube.

In general, the total current may be written

$$J = P_a U_a + P_b U_b + \frac{1}{4\pi}\frac{\partial E}{\partial t}. \tag{41}$$

If Σ is the total area of each of the electrode planes and

$$\Sigma = a + b,$$

where a and b are constants to be defined later, (41) may be written as follows:

$$J\Sigma = \left(P_a U_a \frac{\Sigma}{a} + \frac{1}{4\pi}\frac{\partial E}{\partial t}\right) a + \left(P_b U_b \frac{\Sigma}{b} + \frac{1}{4\pi}\frac{\partial E}{\partial t}\right) b. \tag{42}$$

In this expression, the two streams of current are clearly separated if a and b are taken so that[6]

$$P_a \frac{\Sigma}{a} = P \qquad \text{and} \qquad P_b \frac{\Sigma}{b} = P, \tag{43}$$

where P is the total charge density, equal to the sum of P_a and P_b.

The total current may now be expressed in terms of velocities, only, giving similarly to the transition from (1) to (3),

$$4\pi \frac{e}{m} J\Sigma = a\left(U_a \frac{\partial}{\partial x} + \frac{\partial}{\partial t}\right)^2 U_a + b\left(U_b \frac{\partial}{\partial x} + \frac{\partial}{\partial t}\right)^2 U_b. \tag{44}$$

[6] A more rigorous analysis, involving mean values of the motions of individual electrons, leads to the same result.

When U_a and U_b are each separated into their components according to (5), so that (44) may be resolved into a system of equations, we have for the first two equations, analogous to (6) and (7),

$$4\pi \frac{e}{m} J_0 \Sigma = (a - b) \left[U_0 \frac{\partial}{\partial x} \left(U_0 \frac{\partial U_0}{\partial x} \right) \right] \tag{45}$$

and

$$\beta\Sigma = (a+b) \left[U_0 \frac{\partial}{\partial x} \left(U_0 \frac{\partial U_1}{\partial x} + U_1 \frac{\partial U_0}{\partial x} \right) + \frac{\partial^2 U_1}{\partial t^2} + U_1 \frac{\partial}{\partial x} \left(U_0 \frac{\partial U_0}{\partial x} \right) \right]$$
$$+ (a-b) \left[U_0 \frac{\partial}{\partial x} \left(\frac{\partial U_1}{\partial t} \right) + \frac{\partial}{\partial t} \left(U_0 \frac{\partial U_1}{\partial x} + U_1 \frac{\partial U_0}{\partial x} \right) \right], \tag{46}$$

where the components of U_b have been expressed in terms of those of U_a by means of (40) and the relation that $U_{b0} = - U_{a0}$.

The solution of (45) is, as before,

$$U_0 = \alpha x^{2/3}. \tag{47}$$

where

$$\alpha = \left(18\pi \frac{e}{m} J_{0a} \frac{\Sigma}{a} \right)^{1/3}.$$

Before attempting to solve (46) we make a change of variable as in (10), writing

$$\xi = \frac{3p}{\alpha} x^{1/3} \qquad \text{and} \qquad U_1 = \frac{\omega}{\xi}.$$

This gives from (46)

$$\beta\Sigma = \left[\frac{p^2}{\xi} \frac{\partial^2 \omega}{\partial \xi^2} + \frac{1}{\xi} \frac{\partial^2 \omega}{\partial t^2} \right] \Sigma + (a - b) \left(\frac{2p}{\xi} \frac{\partial^2 \omega}{\partial \xi \partial t} \right). \tag{48}$$

In finding a solution for this, we shall restrict ourselves to the case where all of the electrons turn back at the virtual cathode, so that $a = b$ and therefore the last term of (48) vanishes. The solution of the remaining equation is then,

$$U_1 = - \frac{\beta}{p^2} \left[\sin pt + \frac{1}{\xi} F_1(i\xi + pt) + \frac{1}{\xi} F_2(i\xi - pt) \right], \tag{49}$$

which is analogous to (12).

Again, assuming the two arbitrary functions to have the form,

$$\left.\begin{aligned} F_1(i\xi + pt) &= a \sin (i\xi + pt) + b \cos (i\xi + pt), \\ F_2(i\xi - pt) &= c \sin (i\xi - pt) + d \cos (i\xi - pt), \end{aligned}\right\} \tag{50}$$

and inserting the boundary conditions, (14) and (15), we have, in complex form,

$$U_1 = (M + iN)\frac{\xi_1}{\xi}\frac{\sinh \xi}{\sinh \xi_1} - \frac{\beta}{p^2}\left(1 - \frac{\xi_1}{\xi}\frac{\sinh \xi}{\sinh \xi_1}\right), \tag{51}$$

which is a simpler equation than its analogue (18). The potential is obtained as in (24) giving,

$$\begin{aligned} V_1 = &-\left(\frac{\alpha^3 m}{9p^2e}\right)(M + iN)\frac{\xi_1}{\sinh \xi_1}[\xi \sinh \xi + i(\xi \cosh \xi - \sinh \xi)] \\ &+ \frac{\alpha^3 m\beta}{9p^4e}\left[\left(\xi^2 - \frac{\xi_1 \xi \sinh \xi}{\sinh \xi_1}\right)\right. \\ &\left. + i\left(\frac{\xi^3}{3} - \frac{\xi_1}{\sinh \xi_1}(\xi \cosh \xi - \sinh \xi)\right)\right] + \text{constant}. \end{aligned} \tag{52}$$

The alternating-current potential difference between the grid and the virtual cathode where all of the electrons are turned back may be obtained immediately from (52). As before, the variable ζ will be substituted for ξ to show that the grid-plate region is considered, and currents and velocities will be considered positive when directed towards the origin at the virtual cathode. Thus, from (52)

$$\begin{aligned} V_g - V_p = &-\frac{\alpha^3 m}{9p^2e}(M + iN)[\zeta^2 + i(\zeta^2 \coth \zeta - \zeta)] \\ &+ \frac{\alpha^3 m\beta}{9p^4e} i\left[\frac{\zeta^3}{3} - \zeta^2 \coth \zeta + \zeta\right]. \end{aligned} \tag{53}$$

The velocity, $(M + iN)$ may be expressed in terms of the alternating-current grid potential, V_g, so that the path between grid plane and virtual cathode may be represented by an effective generator in series with an impedance, as was done in (34), (35), and (36).

VII. Oscillation Properties of Positive Grid Triodes [5]

The oscillation properties of the positive grid triode are next to be investigated. In the usual experimental procedure, an external high-frequency circuit is connected between the grid and the plate of the tube. It is unfortunate that this particular arrangement greatly complicates the theoretical relations. Accordingly, a slightly modified experimental set-up will be considered. This modification consists in connecting the external circuit between the cathode and plate of the tube, rather than between grid and plate. Experimental tests have shown that the modified circuit exhibits the same general phenomena

[5] Loc. cit.

as the more usual one, the difference being mainly one of mechanical convenience in securing low-loss leads between the tube and the external circuit.

The modified circuit, then, will be employed for analysis, and the assumption will be made that the necessary direct-current connections are made through chokes which are sufficiently good so that it may be considered that no external high-frequency impedance is connected between either the grid and the plate, or between the cathode and the grid.

It is easy to see that under these conditions there can be no high-frequency current carried away by the grid. It follows that for plane-parallel structures, the alternating-current density, J_1, will be the same both in the cathode-grid region and in the grid-plate region. The arrangement thus reduces the problem to the consideration of the single current, J_1, and the resulting potential difference between cathode and plate.

There are several possible combinations of direct-current biasing potentials. For the first of these, the plate will be supposed to be biased at a potential sufficiently positive to collect all electrons which are not captured by the grid on their first transit. Complete space charge will be assumed both in the cathode region and in the plate region.

Under these conditions, we have the grid-cathode potential difference given by (26) and the grid-plate potential difference given by (35), where the velocity, $M + iN$, is given by (25). We can write,

$$\begin{aligned} V_p - V_c &= (V_p - V_g) + (V_g - V_c) \\ &= -[\text{Eq. } 35] + [\text{Eq. } 26]. \end{aligned} \tag{55}$$

It will be remembered that the current was assumed to be positive in (26) when directed away from the origin, and positive in (35) when directed toward the origin. Therefore, since the same current exists in both regions, and they are joined together at the grid, the sign of the current J_1 remains the same in both (35) and (26), its direction being from cathode to plate. The impedance looking into the cathode-plate terminals may be obtained from (55) by dividing by the amplitude A of J_1 and reversing the sign of the result to correspond to a current from plate to cathode. Letting

$$Z_0 = R_0 + iX_0 \tag{56}$$

represent the impedance looking into the cathode-plate terminals, we can write the result as follows

$$R_0 = -\frac{12r_0}{\zeta^4}\left[\left(1+\cos\eta-\frac{2}{\eta}\sin\eta\right)(1-\cos\zeta)\right.$$
$$-\left(\frac{2}{\eta}-\sin\eta-\frac{2}{\eta}\cos\eta\right)(\zeta-\sin\zeta)$$
$$\left.+(2\cos\eta\sin\eta-2)\right], \tag{57}$$

$$X_0 = -\frac{12r_0}{\zeta^4}\left\{\left(1+\cos\eta-\frac{2}{\eta}\sin\eta\right)(\zeta-\sin\zeta)\right.$$
$$+\left(\frac{2}{\eta}-\sin\eta-\frac{2}{\eta}\cos\eta\right)(1-\cos\zeta)$$
$$+\left[\frac{1}{6}\zeta^3-\frac{4}{\zeta}(1-\cos\zeta)+2\sin\zeta\right]$$
$$\left.+\left[\eta+\frac{1}{6}\eta^3-2\sin\eta+\eta\cos\eta\right]\right\}, \tag{58}$$

where η is the transit angle from cathode to grid, ζ is the transit angle from grid to virtual cathode at the plate, and r_0 is the zero-frequency resistance which would be present in a diode having the grid-plate dimensions, and the same operating direct-current voltages and current densities which occur in the grid-plate region of the triode under consideration.

Fig. 11 shows graphically the relation between R_0 and X_0 for a wide frequency range, in terms of the reference resistance, r_0. Curve A is drawn for the hypothetical condition that $\eta = \zeta$, so that the tube is exactly symmetrical about the grid. Actually such a condition could not be attained, since the grid captures some of the electrons, leaving fewer for producing space charge near the plate. The grid-plate dimension would accordingly have to be increased in order to secure the space charge, but this would cause the transit angle ζ to become larger than η. However, despite the fact that it does not correspond to a physically realizable condition, curve A is nevertheless of use in indicating the limit which is approached as the grid capture fraction is made smaller and smaller.

Curves B and C correspond to values of grid-plate transit angle equal respectively to two and three times the cathode-grid transit angle. Both these curves represent conditions which may readily be obtained experimentally, and indeed, curves lying much closer to A than does the curve B may be secured. For example, the general relation for the ratio of the transit angles in terms of the direct currents J_a and J_b in the cathode and in the plate region, respectively, when

complete space charge exists in both regions, is,

$$\frac{\zeta}{\eta} = \sqrt{\frac{J_a}{J_b}}.$$

Suppose that the grid captured half of the electrons. Then the ratio of transit angles would be 1.41. This would result in a curve lying between A and B in Fig. 11.

The numbers, $\pi/2$, π, and so forth, which are attached to the curves in Fig. 11 show the values of the grid-plate transit angle, ζ, which correspond to the points indicated.

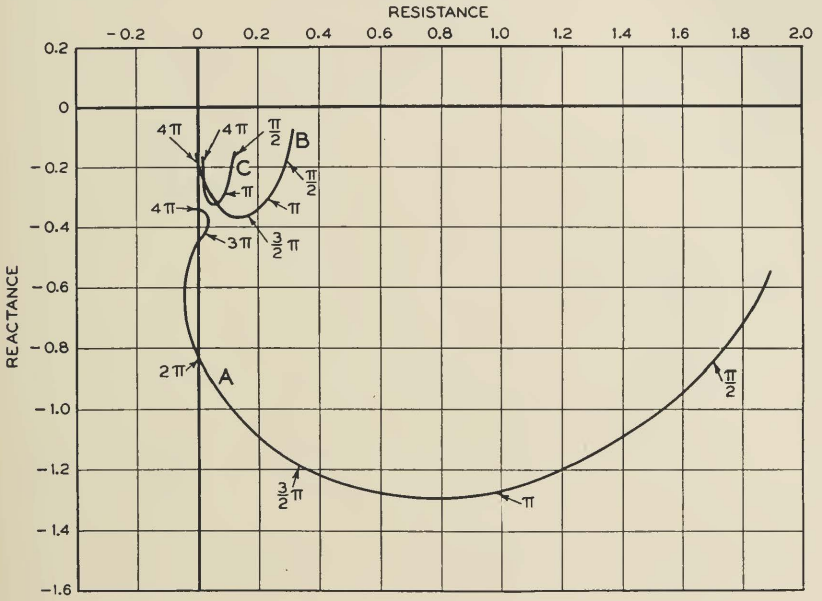


Fig. 11—$R_0 - X_0$ diagram for positive grid, slightly positive plate triode with cathode space charge.

Curve A, $\eta = \zeta$
Curve B, $\eta = 1/2\zeta$
Curve C, $\eta = 1/3\zeta$

We now come to the problem of obtaining information about the oscillation properties of a tube from a set of curves such as those shown in Fig. 11. In a very loose way, and without proof we may state the results of an extension of Nyquist's [7] rule as follows:

If an $R - X$ diagram, which in general may include negative as well as positive frequencies, encircles the origin in a clockwise direc-

[7] H. Nyquist, "Regeneration Theory," *Bell Sys. Tech. Jour.*, Vol. 11, p. 126; January (1932).

tion, then the system represented by the diagram will oscillate when the terminals between which the impedance was measured are connected together.

Verification of this rule, together with further extension to more general cases are expected to be discussed in a subsequent paper. For the present, its validity will have to be accepted on faith, but with the assurance that the applications employed in this discussion are readily capable of demonstration.

Returning to consideration of the positive grid triode with complete space charge on both sides of the grid, and a slightly positive plate, whose $R - X$ diagram is given in Fig. 11, we see at once that the diagram does not encircle the origin as it stands. Of course only positive values of frequency are included in the curves as they are shown. The inclusion of negative frequencies (never mind their physical meaning) would produce a curve which would be the image of the curve shown, a reflecting mirror being regarded as a plane perpendicular to the paper, and containing the R-axis. The curve A, for instance, would have its part corresponding to negative frequencies lying above the R-axis and forming an image of the part lying below. This is shown by the dotted curve in Fig. 12.

It is obvious that the curve of Fig. 12 will encircle the origin or not depending on what happens at infinite frequencies. However, the slightest amount of resistance in the leads to the tube will be sufficient to move the curve to the right and thus exclude the origin. This means that no oscillations would be obtained if an alternating-current short were placed between plate and cathode. The result, although in accord with experiment, is not particularly useful. The important thing is to find whether the curve can be modified by the addition of a simple electrical circuit in such a way that the origin of the resulting $R - X$ diagram for the combination of tube and circuit is encircled in a clockwise direction.

Suppose that a simple inductance is connected in series with the plate lead, and the impedance diagram of the series combination of tube and inductance is plotted. For this arrangement, the $R - X$ diagram of Fig. 12 would be modified as shown in Fig. 13. Here the part of the curve corresponding to negative values of resistance has been pushed upward until the origin is enclosed within a loop which encircles it in a clockwise direction. It is therefore to be expected that oscillations will result. As to their frequency, we can say that the grid-plate transit angle must be at least as great as 2π for this particular example. This follows by supposing a certain amount of resistance to be added in series with the circuit. The effect of this resistance will

be to move the curves on Fig. 13 bodily to the right. The lowest frequency which will just allow the origin to be included within the loop when the series resistance is reduced to zero and the inductance is adjusted, corresponds to a grid-plate transit angle of 2π.

It must be remembered that the foregoing details apply only to curve *A* of Fig. 11, and it has already been pointed out that curve *A*

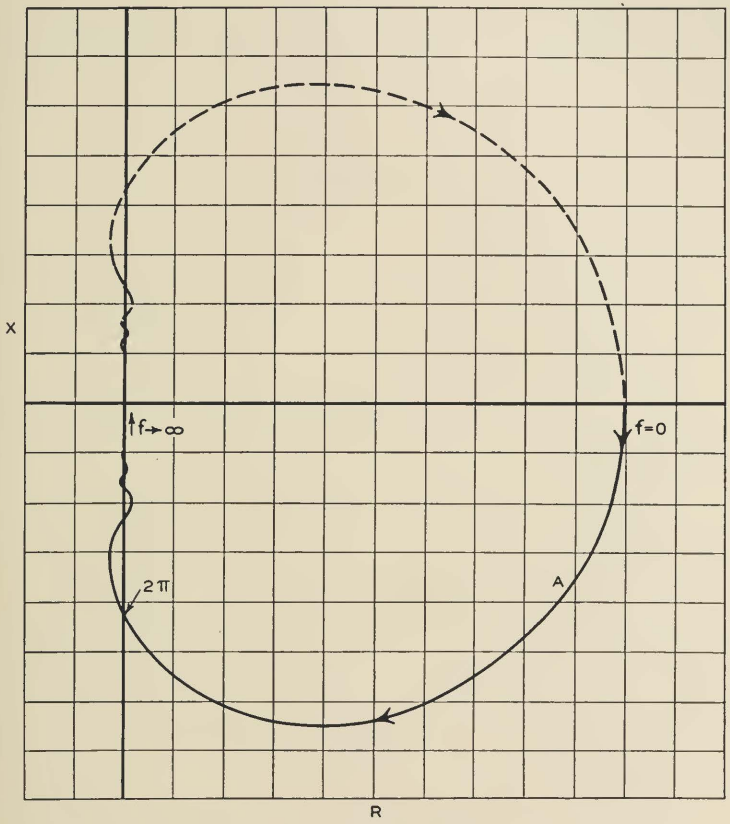


Fig. 12—Curve *A* of Fig. 11 together with the image corresponding to negative frequencies.

represents a limit which can be approached in practice, only as the grid capture fraction is made smaller and smaller. Curve *B* can well be duplicated in experiment. For this case, the lowest frequency at which oscillations may be expected is much higher than before, since the transit angle must be equal to 4π before the resistance becomes negative. Actually, conditions intermediate between the two curves may be realized, so that from a practical standpoint the transit angle

must be in the neighborhood of 3π before we may expect to secure oscillations.

This would correspond to a frequency somewhat higher than is often associated with this type of oscillation. It must be remembered however, that the particular case considered was that of a tube with its plate at a slightly positive potential, whereas the majority of the experimental frequency observations were made with the plate either slightly negative, or, if positive, adjusted so that a virtual cathode was formed inside the tube, and many of the electrons were turned back

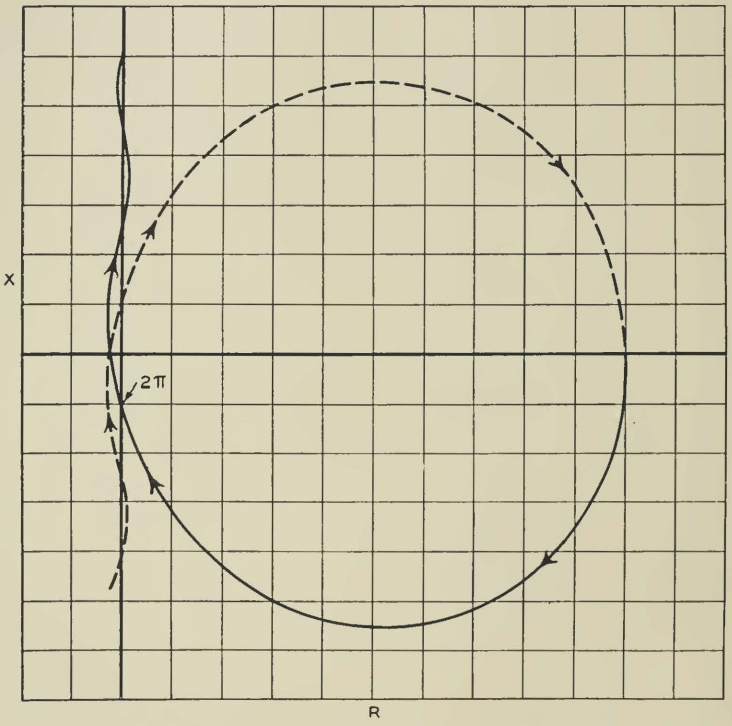


Fig. 13—Modification of Fig. 12 produced by added inductance.

before they reached the plate. The curves of Fig. 11 do not apply to these cases.

Therefore, let us see what happens when the plate is operated at a negative potential so that all of the electrons are turned back before they reach it. At the outset, it should be remarked that this condition does not prohibit the presence of direct-current plate current *after the oscillations have built up to a finite amplitude.* The analysis applies to the requirements for the starting of the oscillations, only, so that

if the plate fluctuates in potential by a very small amount, as it does for incipient oscillations, and hence does not become positive during the alternating-current alternation, then no direct-current plate current can occur when the plate is biased negatively. After oscillations have built up to an appreciable amplitude, the presence of plate current is not only possible, but is in fact to be expected.

We have at hand the mathematical tools with which to compute our $R - X$ diagram for the negative plate triode with complete space charge near the cathode. Thus, instead of substituting (35) in (55) we must substitute (53). Since complete space charge is still postulated near the cathode, (26) and (25) are still applicable. The result is:

$$R_0 = -\frac{12r_0}{\zeta^4}\left[\left(1 + \cos\eta - \frac{2}{\eta}\sin\eta\right)\zeta^2 - \left(\frac{2}{\eta} - \sin\eta - \frac{2}{\eta}\cos\eta\right)(\zeta^2\coth\zeta - \zeta) + (2\cos\eta + \eta\sin\eta - 2)\right], \tag{59}$$

$$X_0 = -\frac{12r_0}{\zeta^4}\left[\left(\frac{2}{\eta} - \sin\eta - \frac{2}{\eta}\cos\eta\right)\zeta^2 + \left(1 + \cos\eta - \frac{2}{\eta}\sin\eta\right)(\zeta^2\coth\zeta - \zeta) + (\tfrac{1}{3}\zeta^3 - \zeta^2\coth\zeta + \zeta) + (\eta + \tfrac{1}{6}\eta^3 - 2\sin\eta + \eta\cos\eta)\right], \tag{60}$$

and the corresponding diagram is shown in Fig. 14. Here the curve A shows oscillation possibilities for transit angles as small as $3/2\pi$, while a much greater amount of resistance would have to be added to the circuit in order to eliminate the negative resistance and so stop the oscillations. In all, then, this method appears to be a better way of operating the system than with the positive plate, and this conclusion is substantiated by experimental observations.

As before, an increase in the grid capture fraction moves the oscillation region up to higher frequencies.

In both of the examples cited above, and represented by Figs. 11 and 14, respectively, complete space charge was assumed near the cathode. The effect of decreasing the cathode heating current so that this charge becomes negligible may be computed by employing (37) in place of (25), and (38) in place of (26).

The resulting equations for a slightly positive plate are,

$$R_0 = -\frac{12r_0}{\zeta^4}\left[\left(\frac{\eta-\sin\eta}{\eta}\right)(1-\cos\zeta)-\left(\frac{1-\cos\eta}{\eta}\right)(\zeta-\sin\zeta)\right], \quad (61)$$

$$X_0 = -\frac{12r_0}{\zeta^4}\left\{\left(\frac{\eta-\sin\eta}{\eta}\right)(\zeta-\sin\zeta)+\left(\frac{1-\cos\eta}{\eta}\right)(1-\cos\zeta)\right.$$
$$\left.+\left[\tfrac{1}{6}\zeta^3-\frac{4}{\zeta}(1-\cos\zeta)+2\sin\zeta\right]+\tfrac{1}{4}\eta\zeta^2\right\}. \quad (62)$$

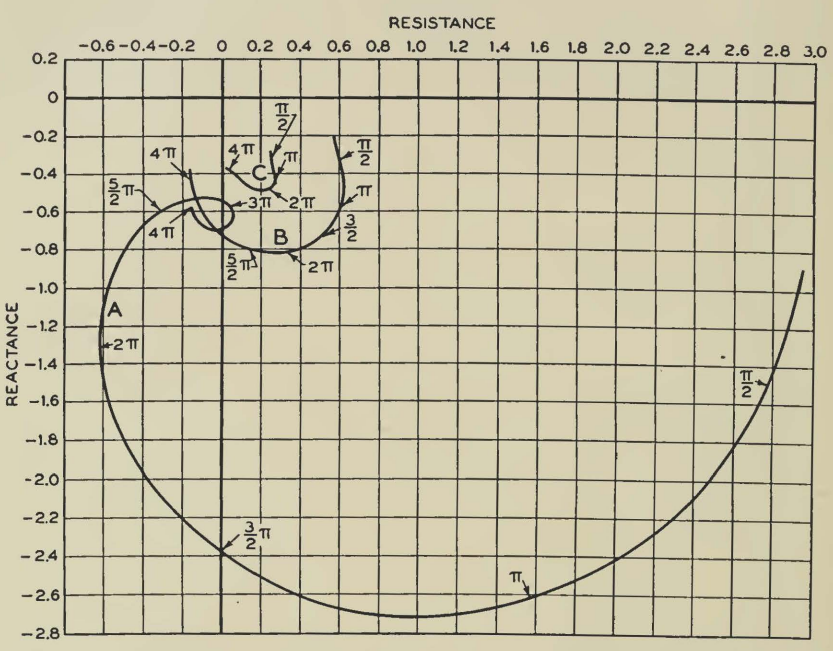


Fig. 14—$R_0 - X_0$ diagram for positive grid, slightly negative plate triode with cathode space charge.

Curve A, $\eta = \zeta$
Curve B, $\eta = 1/2\zeta$
Curve C, $\eta = 1/3\zeta$

The corresponding $R - X$ diagram is given in Fig. 15.

Again, the equations for a negative plate and no cathode space charge are,

$$R_0 = -\frac{12r_0}{\zeta^4}\left[\left(\frac{\eta-\sin\eta}{\eta}\right)\zeta^2-\left(\frac{1-\cos\eta}{\eta}\right)(\zeta^2\coth\zeta-\zeta)\right], \quad (63)$$

$$X_0 = -\frac{12r_0}{\zeta^4}\left[\left(\frac{1-\cos\eta}{\eta}\right)\zeta^2+\left(\frac{\eta-\sin\eta}{\eta}\right)(\zeta^2\coth\zeta-\zeta)\right.$$

$$+ \left(\tfrac{1}{3}\zeta^3 - \zeta^2 \coth \zeta + \zeta\right) + \tfrac{1}{4}\eta\zeta^2 \Big], \qquad (64)$$

and the $R - X$ diagram is shown in Fig. 16.

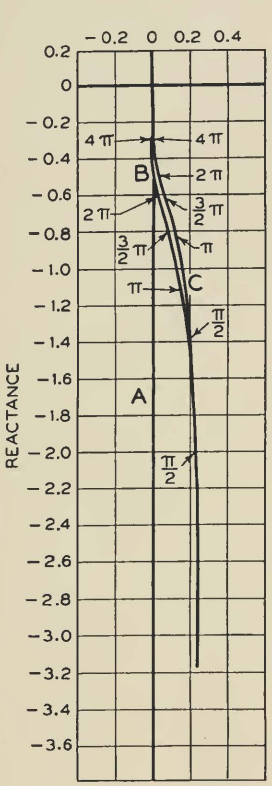


Fig. 15—$R_0 - X_0$ diagram for positive grid, slightly positive plate triode, without cathode space charge.

Curve A, $\eta = \zeta$
Curve B, $\eta = 1/2\zeta$
Curve C, $\eta = 1/3\zeta$

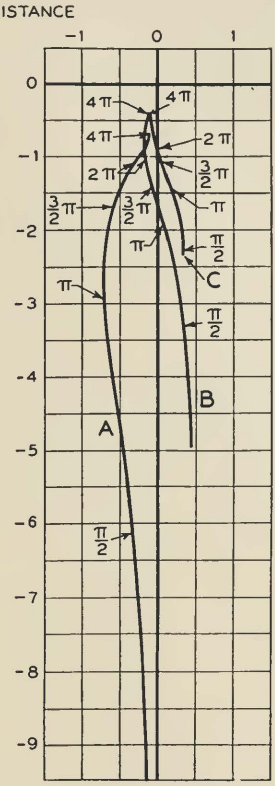


Fig. 16—$R_0 - X_0$ diagram for positive grid, slightly negative plate triode, without cathode space charge.

Curve A, $\eta = \zeta$
Curve B, $\eta = 1/2\zeta$
Curve C, $\eta = 1/3\zeta$

Inspection of Figs. 15 and 16 shows that the negative plate condition is greatly to be preferred when there is no cathode space charge. In fact, when account is taken of the difference in the scales for which Fig. 16 and the other three figures, 11, 14, and 15, are plotted, it is evident that the negative plate without space charge offers the greatest latitude in the adjustment of circuit condition. As in all of the cases, except Fig. 15, a small grid capture fraction is to be desired. If

curve A in Fig. 16 could be attained practically, it would be possible to secure oscillations even at low frequencies by connecting an inductance between the plate and cathode terminals. Curve B shows a low-frequency limit of a little less than $\frac{3}{2}\pi$ for the grid-plate transit angle.

One of the more important observations to be drawn from the curves of Figs. 11, 14, 15, and 16 is that if the inductance between plate and cathode is obtained by means of a tuned antiresonant circuit, then the circuit must be tuned to a frequency somewhat higher than the oscillation frequency. This is in order that it may effectively present an inductive impedance to the oscillating tube, so that the extended curves in the figures may encircle the origin in a clockwise direction.

Another conclusion is that there are so many different permutations and combinations of the operating conditions that it is small wonder that there have been a great many different "theories" and empirical frequency formulas advocated. For instance, operation under conditions giving an $R - X$ diagram which shows negative resistance over a small frequency range, only, such as A in Fig. 11, or B in Fig. 15, would give oscillations whose frequency would be much more nearly independent of the tuning of the external circuit than would conditions which resulted in a negative resistance over a wide frequency range, as at A in Fig. 16. In this latter case the external circuit exerts a large influence upon the frequency.

The data from which Figs. 11 to 16 were plotted are given in the appended tables. The final step in the calculation of these data was a multiplication by 12 which was performed on a slide rule. For all previous steps seven-place tables were employed because of the frequent occurrence of differences of numbers of comparable magnitude.

The effect on the frequency of a change in the operating voltages can be deduced inferentially from the curves. Thus, in general, the formulas for the transit angle have the form,

$$\xi = \frac{Kx}{\lambda\sqrt{V_0}}, \tag{65}$$

where x is the grid-origin distance,
λ is the wave-length,
V_0 is the grid potential,
K is a constant which depends on the mode of operation.

When the plate potential is changed by a relatively large amount the operation undergoes a transition from a limiting mode illustrated by one of the figures to another limiting mode shown on some other one of the figures.

On the other hand, a change in grid potential will act to change the

transit angles on the two sides of the grid in the same proportion. A modification of this generality occurs because the value of x in (65) will shift as the effective position of the virtual cathode moves about. Also, the complete space-charge condition near the plate becomes modified, and so the general relations become extremely variable. The partial space charge that exists with very negative values of plate potential, or with very high values of grid potential does not lend itself readily to mathematical treatment, so that intermediate conditions between complete and negligible space charge can be treated only by inference as to what happens between the two limiting conditions.

With inappreciable space charge on both the plate and the cathode sides of the grid, there can be no oscillations at all, since all impedances then approach pure capacities, with no negative resistance components.

A word concerning the so-called "dwarf" waves is in order before this general theoretical discussion is completed. In the curves, Fig. 14 distinctly shows this possibility in curve A, since the resistance reaches a large negative value at 2π and again at 4π. Likewise Fig. 11 shows the same possibility. On account of the resulting confusion in the figures, the higher frequency portions have not been drawn in the figures, but from (57), (59), (61), and (63) we can see what happens. Thus, for very high frequencies, η is large compared with unity, so that the formulas may be written,

$$R_0 = -\frac{12r_0}{\zeta^4}(\eta + \zeta)\sin\eta, \tag{57-a}$$

$$R_0 = -\frac{12r_0}{\zeta^2}(1 + \cos\eta + \sin\eta),$$

$$= -\frac{12r_0}{\zeta^2}\left[1 + \sqrt{2}\sin\left(\eta + \frac{\pi}{4}\right)\right], \tag{59-a}$$

$$R_0 = -\frac{12r_0}{\xi^4}\left[1 - \frac{\xi}{\eta}\cos\xi + \frac{\xi}{\eta}\cos\eta\right], \tag{61-a}$$

$$R_0 = -\frac{12r_0}{\xi^2}. \tag{63-a}$$

It is noteworthy that all of these exhibit the possibility of "dwarf" waves separated by discrete frequency intervals except (63-a). On the other hand, (63-a) gives possible conditions for operation at all high frequencies provided that the proper external circuit may be secured.

VIII. Postscript

The extension of the electronics of vacuum tubes which was described in the preceding pages must be regarded in the light of a tenta-

tive starting point rather than as a completed structure. Of the fundamental correctness of the method of attack there can be little doubt. The various simplifying assumptions, however, require careful scrutiny and doubtless some of them will be revised as time goes on and additional experience is acquired. Experimental guidance will be invaluable, and indeed certain data already have been obtained which are helpful in analysis of the assumptions. Although these data are in general qualitative agreement with the theory as outlined, the experimental technique must be refined before quantitative comparison can be made. It is hoped that the results can be made available at an early date.

Among the various assumptions which were made in the development of the theory, there are three which lead particularly to far-reaching consequences. These three may be enumerated as follows:

1. Plane-parallel tube structures
2. Current flow in straight lines
3. Small alternating-current amplitudes.

There are grounds for the belief that the assumption of plane-parallel tube structures does not exclude the application of the alternating-current results to cylindrical structures as completely as might be supposed. In the first place, the approximation of cylindrical arrangements to the plane-parallel structure becomes better as the cathode diameter is made large. Many tubes contain special cathode structures where this is the case. Furthermore, Benham [1] has obtained an approximate solution for the alternating-current velocity in cylindrical diodes where the cathode diameter is vanishingly small, and the transit angle is less than 5 radians. The resulting curves of alternating-current velocity versus transit angle have the same shape as the curves for the planar structures, and when the cylindrical transit angle is arbitrarily increased by about 20 per cent, the quantitative agreement is fair for transit angles less than 4 radians. It follows that until accurate solutions for cylindrical triodes can be obtained, the planar solutions may be expected to give correct qualitative results, and fair quantitative results when appropriate modifications of the transit angle are made. In fact, good agreement is obtained if calculations of the cylindrical transit angle are made as though the structure were planar.

The assumption of current flow in straight lines is open to some question when a grid mesh is interposed in the current path. For the positive grid triode, the objection to the assumption has been overcome by postulating a special case where an ideal choke coil prevents

the grid from carrying away any of the alternating current. Benham [1] has suggested an alternative which seems to work fairly well when the grid-cathode path of a negative grid tube is considered, but which offers grave difficulties when the grid-plate path is included. A still different alternative was employed in the present paper in connection with negative grid triodes, and successfully indicates phase angles for the mutual conductance of the tube which are qualitatively logical. The grid-plate path is still without adequate treatment, however.

As to the third general assumption: that of relatively small alternating-current amplitudes, there can be no objection from a strictly mathematical point of view, and for a very large proportion of the physical applications the assumption is thoroughly justified. Indeed, it is the only one which is successful in giving *starting* conditions for oscillators. However, when questions as to the power efficiency of oscillators or amplifiers arise, then the "small signal" theory is inadequate, and should be supplanted by an approximate theory. The form which this approximate theory should take is indicated by the standard methods of dealing with the efficiencies of low-frequency power amplifiers and oscillators where the wave shape of the plate current is assumed to be given. The application of the same kind of approximation to ultra-high-frequency circuits may eventually prove to be a simpler matter than the "small signal" theory set forth in these pages.

Besides the three main assumptions discussed above, there was a fourth assumption which, although of lesser importance, deserves some comment. This fourth assumption involves the neglect of initial velocities at a hot cathode. If all electrons were emitted with the same velocity, the theory is adequate, and may be applied as indicated by Langmuir and Compton.[3] When the distribution of velocities according to Maxwellian, or Fermi-Dirac, laws is considered, some modifications may be necessary. In general, a kind of blurring of the clear-cut results of the univelocity theory may be expected, which will be expected to result in an increase in the resistive components of the various impedances at the expense of the reactive components. Again, lack of symmetry in the geometry of the tube structure may be expected to do the same thing, since the transit angles are then different in the different directions.

Finally, however, and with all its encumbering assumptions, it is hoped that the excursion back to fundamentals which was made in this paper, has resulted in a method of visualizing the motions of the condensations and rarefactions of the electron densities inside of vacuum tubes operating at high frequencies and has shown their relation to the conduction and displacement components of the total current.

DATA FOR FIG. 11

ζ	$\eta = \zeta$		$\eta = \frac{1}{2}\zeta$		$\eta = \frac{1}{3}\zeta$		$\eta = \frac{1}{4}\zeta$	
	R_0	X_0	R_0	X_0	R_0	X_0	R_0	X_0
1.0	1.87	−0.577	0.302	−0.119			0.0633	
1.4	1.75	−0.777	0.292	−0.164			0.0604	−0.160
1.57	1.69	−0.855	0.288	−0.182	0.112	−0.172		
1.8	1.60	−0.949	0.280	−0.204	0.108	−0.193	0.0568	−0.200
2.356	1.36	−1.13	0.260	−0.241	0.0987	−0.240		
2.8	1.15	−1.23	0.241	−0.289			0.0458	−0.278
3.14	0.986	−1.27	0.226	−0.311	0.0838	−0.289	0.0418	−0.297
3.6	0.769	−1.29	0.204	−0.335	0.0748	−0.308	0.0364	−0.316
4.0	0.593	−1.27	0.186	−0.350	0.0673	−0.320	0.0320	−0.327
4.71	0.326	−1.18	0.153	−0.366	0.0457	−0.329		
5.2	0.186	−1.08	0.132	−0.368			0.0216	−0.330
5.6	0.0972	−0.987	0.116	−0.368			0.0193	−0.324
6.28	0	−0.830	0.0923	−0.358	0.0365	−0.309	0.0165	−0.306
6.8	−0.0348	−0.719	0.0767	−0.346			0.0153	−0.291
7.2	−0.0440	−0.644	0.0662	−0.337	0.0308	−0.284	0.0148	−0.278
7.85	−0.0369	−0.546	0.0515	−0.318	0.0284	−0.264		
8.4	−0.0199	−0.487	0.0414	−0.302	0.0268	−0.248	0.0144	−0.238
9.0	+0.000414	−0.444	0.0320	−0.284	0.0254	−0.233	0.0144	−0.222
9.42	0.0125	−0.424	0.0262	−0.272	0.0244	−0.222	0.0143	−0.235
10.0	0.0218	−0.407	0.0194	−0.256			0.0138	−0.198
10.99	0.0213	−0.385	0.0101	−0.232	0.0192	−0.193		
12.57	0	−0.342	0	−0.197	0.0128	−0.172	0.00962	−0.162

DATA FOR FIG. 14

ζ	$\eta = \zeta$		$\eta = \frac{1}{2}\zeta$		$\eta = \frac{1}{3}\zeta$		$\eta = \frac{1}{4}\zeta$	
	R_0	X_0	R_0	X_0	R_0	X_0	R_0	X_0
1.0	2.94	−0.932	0.581	−0.225			0.133	−0.221
1.4	2.84	−1.33	0.595	−0.304			0.136	−0.286
1.57	2.78	−1.49	0.601	−0.336	0.252	−0.302		
1.8	2.67	−1.71	0.606	−0.377	0.255	−0.330	0.139	−0.336
2.356	2.30	−2.19	0.618	−0.465	0.264	−0.381		
2.8	1.90	−2.48	0.619	−0.528			0.147	−0.401
3.14	1.56	−2.63	0.613	−0.571	0.272	−0.424	0.150	−0.410
3.6	1.06	−2.71	0.598	−0.625	0.276	−0.438	0.153	−0.416
4.0	0.645	−2.67	0.577	−0.665	0.277	−0.448	0.155	−0.417
4.71	0.00015	−2.38	0.525	−0.728	0.275	−0.460	0.157	−0.412
5.2	−0.319	−2.08	0.479	−0.762			0.158	−0.407
5.6	−0.494	−1.80	0.436	−0.784			0.158	−0.404
6.28	−0.608	−1.31	0.356	−0.807	0.253	−0.476	0.156	−0.396
6.8	−0.565	−0.990	0.292	−0.814			0.154	−0.390
7.2	−0.480	−0.796	0.241	−0.803	0.232	−0.483	0.152	−0.386
7.85	−0.290	−0.594	0.160	−0.797	0.213	−0.485		
8.4	−0.111	−0.528	0.0957	−0.773	0.196	−0.486	0.144	−0.376
9.0	+0.00226	−0.536	0.0308	−0.735	0.175	−0.485	0.138	−0.372
9.42	0.0574	−0.574	−0.0102	−0.705	0.160	−0.484	0.133	−0.369
10.0	0.077	−0.634	−0.0581	−0.666			0.127	−0.365
10.99	0	−0.690	−0.118	−0.564	0.101	−0.436		
12.57	−0.152	−0.560	−0.152	−0.414	0.0445	−0.383	0.0938	−0.348

DATA FOR FIG. 15

ζ	$\eta = \zeta$		$\eta = \frac{1}{2}\zeta$		$\eta = \frac{1}{3}\zeta$		$\eta = \frac{1}{4}\zeta$	
	R_0	X_0	R_0	X_0	R_0	X_0	R_0	X_0
1.0			0.239	−3.06			0.179	−1.58
1.4			0.228	−2.22			0.172	−1.19
1.57			0.223	−2.00	0.199	−1.39		
1.8			0.215	−1.76	0.192	−1.24	0.162	−0.979
2.356			0.193	−1.39	0.173	−1.01		
2.8			0.173	−1.21			0.131	−0.737
3.14			0.157	−1.10	0.130	−0.825	0.120	−0.693
3.6			0.135	−0.983	0.123	−0.756	0.104	−0.645
4.0			0.116	−0.903			0.0902	−0.610
4.71			0.0837	−0.788	0.0796	−0.633		
5.2			0.0643	−0.724			0.0540	−0.524
5.6			0.0503	−0.677				
6.28			0.0308	−0.605	0.0346	−0.506	0.0308	−0.455
6.8			0.0197	−0.558			0.0232	−0.425
7.2			0.0131	−0.525	0.0194	−0.444	0.0187	−0.402
7.85			0.00568	−0.477	0.0126	−0.406		
8.4			0.00203	−0.442	0.00939	−0.378	0.109	−0.343
9.0			−0.0000284	−0.409	0.00710	−0.351	0.00908	−0.318
9.42			−0.000646	−0.388	0.00608	−0.333	0.00826	−0.290
10.0			−0.000817	−0.363			0.00744	−0.284
10.99			−0.000402	−0.327	0.00408	−0.282		
12.57			0	−0.285	0.00216	−0.246	0.00385	−0.227

DATA FOR FIG. 16

ζ	$\eta = \zeta$		$\eta = \frac{1}{2}\zeta$		$\eta = \frac{1}{3}\zeta$		$\eta = \frac{1}{4}\zeta$	
	R_0	X_0	R_0	X_0	R_0	X_0	R_0	X_0
1.0	−0.176	−9.35	0.426	−4.84			0.342	−2.52
1.4	−0.306	−6.84	0.366	−3.64			0.316	−1.96
1.57	−0.363	−6.15	0.338	−3.33	0.345	−2.32		
1.8	−0.436	−5.41	0.299	−3.00	0.321	−2.11	0.287	−1.67
2.356	−0.584	−4.15	0.206	−2.46	0.263	−1.77		
2.8	−0.658	−3.44	0.137	−2.17			0.212	−1.29
3.14	−0.685	−3.00	0.0888	−1.99	0.188	−1.48	0.190	−1.21
3.6	−0.685	−2.52	0.0318	−1.80	0.149	−1.36	0.162	−1.12
4.0	−0.661	−2.17	−0.0104	−1.65	0.120	−1.27	0.140	−1.06
4.71	−0.565	−1.69	−0.0698	−1.43	0.0747	−1.13	0.108	−0.957
5.2	−0.482	−1.45	−0.0998	−1.30			0.0871	−0.898
5.6	−0.413	−1.30	−0.119	−1.21			0.0730	−0.853
6.28	−0.304	−1.11	−0.141	−1.07	0.0478	−0.907	0.0523	−0.787
6.8	−0.236	−1.02	−0.151	−0.975			0.0389	−0.742
7.2	−0.195	−0.963	−0.155	−0.910	−0.0220	−0.805	0.0296	−0.709
7.85	−0.148	−0.894	−0.156	−0.815	−0.0364	−0.743		
8.4	−0.126	−0.848	−0.152	−0.747	−0.0458	−0.695	+0.0072	−0.625
9.0	−0.113	−0.803	−0.145	−0.680	−0.0538	−0.647	−0.00162	−0.589
9.42	−0.109	−0.771	−0.138	−0.638	−0.0582	−0.615	−0.00706	−0.565
10.0	−0.107	−0.727	−0.128	−0.588			−0.0135	−0.535
10.99	−0.101	−0.653	−0.107	−0.518	−0.0668	−0.517		
12.57	−0.0760	−0.556	−0.0760	−0.439	−0.0667	−0.439	−0.0314	−0.426

Contemporary Advances in Physics, XXVII
The Nucleus, Second Part *

By KARL K. DARROW

In this Second Part the major subject is Transmutation: that is to say, the alteration or disintegration of a nucleus, the unique and distinctive part of any atom, by impacts of fast-moving corpuscles. For the last year and a half the pace of progress in this field has been increasingly rapid, and in all likelihood is destined to become yet swifter. This is partly because of the discovery—a discovery due largely to theoretical foresight—that transmutation of some elements is practicable with protons of a relatively modest energy which can be produced in laboratories without any serious difficulty. Partly it is due to the discovery of neutrons and of deutons, particles which apparently possess remarkable ability in effecting certain kinds of transmutation. Partly also it is due to advances and refinements in the methods of working with alpha-particles, the first variety of corpuscle with which disintegration of nuclei was ever achieved. People are already beginning to speak of "nuclear chemistry" as a special branch of science, and this is already almost justified by the number of cases known in which two nuclei interact and produce two others which are recognizable.

Bringing the First Part up to Date

STRANGE as it seems to speak of "bringing up to date" something that was published only six months ago, one is sometimes obliged to do so by the rapid march of science; and three of the "elementary particles" of which I spoke in the First Part were and still are so young—or to speak more carefully, our acquaintance with them is still so young—that their rôle and situation in the body of physical knowledge is changing from month to month.

The Positive Electron

Of the positive electron the most striking new thing to be said is, that there is now a new way of generating it: by impacts of alpha-particles against metals. This so far has been applied only by its discoverers, M. and Mme. Joliot; only with alpha-particles from polonium, therefore of energy 5.3 millions of electron-volts; only to five metals, of which beryllium and boron and aluminium yielded positive electrons, while silver and lithium did not. It is as yet the most efficacious way of producing positive electrons, Joliot having evoked last summer as many as 30,000 of these corpuscles per second from aluminium. This of course looks small when compared with the torrents of negative electrons which incandescent metals will pour out,

* "The Nucleus, First Part" was published in the July 1933 issue of the *Bell Sys. Tech. Jour.*, Vol. XII.

but these are not a proper standard of comparison. Rather should one say that in the autumn of 1932 positive electrons were being observed at the rate of three or four a year, and already by the summer of 1933 this rate had been enhanced to thirty thousand in the second!

The other voluntary way of generating positive electrons—by applying hard gamma-rays to heavy elements—has already been studied enough to yield the data of the following table. Here, in the first column, stand the names of various sources of gamma-rays (the one denoted as "Po + Be" is beryllium exposed to impacts of alpha-particles from polonium); in the second, the energy-values in MEV (I use this symbol hereafter for "millions of electron-volts") of the individual photons of these rays; in the third, the symbols of various metals; in the fourth, the number of positive electrons per hundred negatives, ejected from these metals by these gamma-rays; in the fifth, the authorities:

Po + Be	5	U	40	Joliots
		Pb	30	Joliots
		Pb	35	Chadwick
		Cu	18	Joliots
		Al	5	Joliots
ThC″	2.6	Pb	8	Joliots
		Pb	4	Chadwick
Ra(B + C)	1.0–2.2	Pb	3	Grinberg
Po	0.85	Pb	0	Meitner-Philipp

The percentages in the fourth column give at the moment our best available notion as to the relative plentifulness of positive electrons, produced by the several kinds of rays falling upon the several metals. One would prefer to have the total number of positives per unit intensity of the infalling rays, but that is not available at present—I presume because of the difficulty of measuring these intensities. One must remember that the data usually consist in observations of a few hundred or a few dozen cloud-tracks, so that the accuracy of these percentages cannot be great.[1]

We note that with lead the proportion of positive electrons mounts rapidly with increasing photon-energy, and that with 5 MEV-photons

[1] This perhaps is sufficient to account for a discrepancy between the general trend of the table and a value of 1/3 given by Meitner and Philipp for the ratio of positives to negatives when brass is exposed to (Po + Be). Should the table be extended and supported by a successful theory, it should then be possible to determine the frequency of gamma-rays by the percentage of positives which they produce when falling on a metal. In this connection it is interesting that Anderson's latest data indicate that positive and negative electrons are about equally abundant among the ionizing particles of the cosmic rays, a fact which suggests that if they are due to photons, these must be of a distinctly higher energy than any of those cited in the foregoing table.

the proportion goes up rapidly with the nuclear mass of the bombarded atoms.[2] Both of these rules are in harmony with the remarkable theory to which I alluded in the First Part—the theory that each positive electron (together with a negative companion) springs into being from a transmutation of light into electricity! It is supposed that a photon transmutes itself into a pair of electrons, one of each sign.



Fig. 1—Tracks of an electron-pair (positive and negative) arising in argon exposed to gamma-rays, and probably created near an argon nucleus by transmutation of a photon. (M. et Mme. Joliot)

Conservation of the net charge of the universe is assured in this hypothetical process. Conservation of mass and energy is attainable, for the speeds of the electrons may be such that their energies together are equal to the energy of the vanished photon. I take this occasion to repeat Einstein's principle, which figures so importantly in these articles. The energy E of a material particle moving with speed v (relatively to the observer) is given by the formula:

$$E = m_0(1 - \beta^2)^{-1/2}c^2 = mc^2,$$

[2] In this connection it should be noted that the source "Po + Be" emits neutrons as well as photons, and while the first-named are certainly not chiefly responsible for the positive electrons, they may produce some of these. Chadwick observed that the rays from a "Po + B" source (boron in place of beryllium), which consist of neutrons plus some photons of about the same energy as the photons of ThC", evoked from lead a distinctly larger percentage of positive electrons than do the rays of ThC".

in which c stands for the speed of light in vacuo; β for v/c; and m_0 for a constant. The ratio of E to c^2 is the function of v and m_0 which this equation defines, and is denoted by m and called the mass of the particle: it is in this sense that mass and energy are equivalent. We may (mentally) divide the mass of the moving particle into two terms m_0 and $(m - m_0)$, and the energy into two terms m_0c^2 and $(m - m_0)c^2$. We may further call m_0 the rest-mass and m_0c^2 the energy associated with the rest-mass; and we may call $(m - m_0)c^2$ the kinetic energy and $(m - m_0)$ the mass associated with the kinetic energy or the extra mass due to the motion of the particle. Such will be the terminology used in these articles, although this definition of kinetic energy is only approximately the same as the classical and familiar one.[3]

Returning to the argument about the transmutation of light into electrons, or more precisely, of a photon into an electron-pair: conservation of mass and energy is attainable, for the two electrons may have such speeds—call them β_1c and β_2c—that the sum of $m_0(1 - \beta_1^2)^{-1/2}c^2$ and $m_0(1 - \beta_2^2)^{-1/2}c^2$ is equal to the energy $h\nu$ of the photon. But the demand for conservation of momentum makes apparently serious trouble. If we assume that a photon voyaging through the depths of space suddenly converts itself spontaneously into a pair of electrons, and if then we attempt to impose both conservation of momentum and conservation of energy, the equations lead us straightway into an inescapable muddle, in which the original assumptions contradict each other. We are driven therefore to infer that the imagined process is impossible. But this seeming catastrophe of the theory turns out to be a blessing. What is observed is not after all the transmutation of a photon in the depths of empty space, but a process which occurs in the depths of plates of lead and other heavy elements. If we suppose that such a transmutation occurs near to a massive nucleus, then this may receive some of the energy and some of the momentum of the photon; and the equations show that the momentum which it takes may be quite sufficient to permit the process to occur, while the energy which it takes is so small that for practical purposes we may still pretend that the whole of the energy of the photon is divided between the electrons (though we certainly should not forget about the small fraction which goes to the nucleus). All the principles are thus fulfillable: conservation of charge requires that there should be

[3] The classical definition of kinetic energy is $(1/2)mv^2$; the present or relativistic definition, viz. $(m - m_0)c^2$, is an infinite series of which the first term is identical with the classical definition. The difference between the two definitions increases as the speed of the particle increases, but so far as I know there has not yet been an actual case in which it is of practical importance.

two electrons of opposite signs; conservation of energy, that they should have appropriate speeds; conservation of momentum, that the process should occur only near a massive nucleus.

This is the most alluring of all theories, for it is the doctrine that the substance of matter and the substance of light are ultimately the same, being interconvertible. It therefore demands, and is surely destined to receive, the sharpest and fullest of testing; the more so because there is a rival in the field, the theory that the positive electron exists beforehand and from all time in the nucleus of the atom, and is ejected from it by the photon. The newest way of producing positive electrons by alpha-particle impact seems to speak in favor of the latter. One could indeed suppose that the kinetic energy of the alpha-particle is transformed into an electron-pair, directly or through the intermediacy of a transiently-existing photon; but this would be an artificial idea unless it were to be supported by a basic theory or by observing that the positive electrons are often paired with negatives (which the Joliots do not say). In favor of the former theory speak the facts that in several scores of cases paired electrons have been observed—*i.e.*, two electrons of opposite signs were seen to spring from the same point (so far as the eye could tell)—when metal plates were bombarded with gamma-rays; and the further fact that the energies of these electron-pairs and of individual positives did not surpass those of the infalling photons, though they approached it often.[4] There are always apparently unpaired positives and many more unpaired negatives; but one may always say that with some of the pairs it happened that one member remained in the metal and the other got away, while many of the negatives are surely electrons which have been expelled from their places by photons acting in the well-known ways. Further, there are more or less forcible indications that some part of the absorption of gamma-rays in heavy metals may be ascribed to the formation of electron-pairs, and some part of the radiation scattered from the metals when gamma-rays fall on them may be attributed to the reunion of two electrons of opposite sign which re-transmute themselves into light; but some of the data are not checked, and the time seems not ripe for reviewing them. In the hands of Oppenheimer and Plesset the transmutation theory has supplied other quantitative

[4] See First Part of this article, pp. 304–305, *B.S.T.J.*, July 1933. The *kinetic* energies of electron-pairs and *a fortiori* of positive electrons should not come within one million electron-volts of the energy-value of the photons, for the rest-mass of two electrons amounts approximately to a million of these units. This rule has lately been strengthened by evidence from Anderson and his colleagues, who in a couple of hundred of additional cases find no violation of it; the distribution-in-energy curves for pairs and for (apparently) isolated positives extend up to the predicted upper limit, and there they fall to the horizontal axis. More evidence of this kind has been accumulated by Blackett (loc. cit. footnote 5).

predictions meet for testing, and it is likely that in six months more a great deal will be learned.[5]

The Deuton

The newly-discovered isotope of hydrogen of mass-number 2—H^2, "heavy hydrogen," or, to adopt Urey's name for it, "deuterium"—has suddenly become the most popular and the most eagerly sought-after of all chemical substances. This is because of the notable chemical and physical differences between it and its compounds on the one hand, H^1 and the corresponding compounds of H^1 on the other. So great are these differences that by the usage of twenty years ago H^2 would probably have been called a new element, and indeed it deserves all the prestige that would accrue to it from being so denoted; but to violate the present and most wisely-based of usages, whereby an element is characterized by atomic number rather than by the ensemble of its properties, would be mistaken.[6]

Deuterium is so rare by comparison with H^1 (Urey's "protium") that it would still be very unfamiliar, but for the unexpected and remarkable efficacy of the electrolytic method of separating water molecules comprising H^2 atoms from water molecules comprising none but H^1 atoms. It turns out that if an aqueous solution is electrolyzed until only a very tiny fraction of the original liquid remains, the proportion of the former kind of molecule in that tiny residue is anomalously large. Washburn seems to have been the first to suspect that this might happen; he procured samples of the residues from electrolytic cells which had been operated continuously in commercial plants for two and three years, and sent them to Urey, who performed a spectrum-analysis and observed "a very definite increase in the abundance of H^2 relative to H^1." Shortly afterwards the method was put into operation on a grand scale by G. N. Lewis and his collaborators, with spectacular results. In one experiment, for instance, they started out with twenty liters of water, electrolyzed it until there remained but half a cc. of liquid, and found that in this residue deuterium atoms made up two-thirds of all the hydrogen atoms which were left. For months thereafter, nearly every paper on deuterium and on the deuton which was published began with an acknowledgment to Lewis for a small amount of water rich in heavy hydrogen which the fortunate author had received from him.

[5] For a fuller account of the situation as it now stands, see an article of mine in the *Scientific Monthly*, January 1934; also one by P. M. S. Blackett, *Nature* **132**, pp. 917–919 (Dec. 16, 1933), which incidentally contains some further data.

[6] I should think that the case of deuterium by itself would make it necessary henceforth to define the concept "element" altogether from the concept "atomic number," forsaking all the earlier definitions.

Interesting as are the chemical and physical properties of deuterium and its compounds, we are here concerned only with the nucleus of the H^2 atom, the deuton (all the other suggested names seem to be fading out). The accepted value for its mass is that given by Bainbridge, 2.0131 on the standard scale in which the mass of the O^{16} atom is 16 exactly. Of its spin I shall speak in a later article. Its powers of transmutation are remarkable, and quite unlike those of H^1; if first a beam of H^1 nuclei (protons) and then an equal beam of deutons be directed against targets of various elements, the number of fragments observed per unit time is greater for some elements and less for others, and their ranges in general are different. In some cases it seems possible that the deutons themselves are being split into protons and neutrons, a result of great importance if it can be established beyond question. We shall consider the data at length.

The neutron

Most of what has newly been learned about the neutron will find appropriate places elsewhere in this article. There should be a separate section about the deflections suffered by neutrons when they impinge on or pass close to nuclei without transmuting them—the topic known as "scattering," "interception," or (badly) "absorption" of neutrons. This topic however is scarcely ripe for description in such an article as this, the experiments being difficult and the inferences from the data being highly controversial. I therefore postpone it to some future occasion, remarking only that it seems established that a neutron may pass within a very short distance indeed from a nucleus—only a very few times 10^{-13} cm from the centre thereof—without interacting with that nucleus in any perceptible way.

Masses of the Lighter Atoms

There are now thirteen of the lighter atoms of which the masses—in terms of the mass of the O^{16} atom taken as 16 exactly—have been determined to four and even to five significant figures. Most of these values were mentioned in the First Part, but it will be convenient to have them all tabulated here. They are the masses of complete atoms, nuclei accompanied by their full quotas of orbital electrons. The uncertainties quoted are the "probable errors"; where Aston originally gave the maximum possible uncertainty, this has been divided by 3 (see First Part, footnote 10). Values marked with an asterisk are from Bainbridge, the others from Aston; the value for H^1 has been obtained by both.

H^1	1.007775 ± .000035	C^{12}	12.0036 ± .0004
* H^2	2.01363 ± .00008	N^{14}	14.008 ± .001
He^4	4.00216 ± .00013	O^{16}	16.0000 (standard)
* Li^6	6.0145 ± .0003	F^{19}	19.000 ± .002
* Li^7	7.0146 ± .0006	* Ne^{20}	19.9967 ± .0009
* Be^9	9.0155 ± .0006	* Ne^{22}	21.99473 ± .00088
B^{10}	10.0135 ± .0005	* Cl^{35}	34.9796 ± .0012
B^{11}	11.0110 ± .0005	* Cl^{37}	36.9777 ± .0019

The table of the chemical atomic weights reproduced in the First Part has suffered two alterations: a very slight change in the given value for K, from 39.10 to 39.096; and an important change in the chemical atomic weight of carbon, which rises from 12.00 to 12.011, and now permits of an abundance of C^{13} easier to reconcile with the observed intensities of the spectrum-lines of this substance than was the abundance, or rather the scarcity, implied by the former value.[7]

The list of isotopes detected by Aston's mass-spectrograph has been enlarged by the following examples,[8] which the reader may enter upon Fig. 6 of Part I: neodymium, $Z = 60$, $A-Z = 83$; samarium, $Z = 62$, $A-Z = 85, 86, 87, 90, 92$; europium, $Z = 63$, $A-Z = 88, 90$; gadolinium, $Z = 64$, $A-Z = 91, 92, 93, 94, 96$; terbium, $Z = 65$, $A-Z = 94$.

New Developments in Transmutation: The Apparatus

In the two years and a quarter which are all that have elapsed since I published in this *Journal* an article on transmutation,[9] the situation in this field has vastly changed, and the prospects for the future have been amplified immensely. So lately as the early spring of 1932, disintegration of a nucleus had not yet been demonstrably achieved except by alpha-particles possessing energy not smaller than three millions of electron-volts. Schemes for producing five- and ten-million-volt ions were already under way, being ardently pushed onward because it was supposed that transmutation would never be effected by any agency much feebler. But in the course of 1930, Cockcroft and Walton of the Cavendish Laboratory had been emboldened by a theory (I will describe it later) to imagine that protons of only a few hundred thousand electron-volts might be able to transmute, and to risk their time and labors in the task of developing powerful streams of such particles. After two years of work they

[7] See an item in *Nature*, **132**, 790–791 (Nov. 18, 1933). In the table of masses on p. 303 of the First Part, change 1.0078 to 1.0072 and 4.002 to 4.001 (the former values refer to complete atoms, not bare nuclei).

[8] F. W. Aston, *Nature* **132**, 930–931 (Dec. 16, 1933).

[9] "Contemporary Advances in Physics XXII," *Bell System Technical Journal*, **10**, 628–665 (October 1931). I refer to this article hereinafter as *Transmutation*.

were justified in the event; for they detected fragments proceeding from targets of lithium bombarded by their protons, with energy-values anywhere from half-a-million down to only seventy thousand electron-volts.

It would be hard to overstate the joyful surprise of this announcement. Transmutation, of some elements at least, was easier by far than had been thought! It would not after all be necessary to fare forth into the unknown, and face at once the problems of applying voltages without precedent; successes which had seemed doubtful at best and assuredly distant were after all to be had by a relatively slight extension of a known technique. All over Europe and America people began making plans for applying these voltages, so much less formidable than those which had previously been thought indispensable. Nevertheless the first who confirmed and extended the work of Cockcroft and Walton were those who had aimed from the start at the higher and harder goal: Lawrence and his colleagues at Berkeley. Their work had not been wasted, for they instantly found themselves able to measure the disintegration of lithium by protons all the way up to 710,000 electron-volts; and within four months they had carried the upper limit onward to 1,125,000, and as I write these lines they have just announced that the limit has soared to three millions! From Pasadena also comes word of transmutation achieved by protons, and deutons, and helium nuclei, endowed with energy by voltages ranging downward from nine hundred thousands.

These are not the only novel results of the last two years and a quarter. The neutron has disclosed itself not only as a product, but as an agent of transmutation, able to alter nuclei which have thus far resisted both the alpha-particles and the protons which have been showered upon them in laboratories. The disintegrations effected by alpha-particles have been studied with ever-increasing minuteness and detail, and are beginning to show that nuclei are structures capable of existing in various normal states and excited states, characterized by distinctive energy-values. The emission of alpha-particles from radioactive nuclei has been studied with a new precision, and leads to the same conclusion. The astonishing feats achieved with bombarding particles of lesser energy have not lessened the hope of achieving startling things with particles of greater.

Cockcroft and Walton, inspired by theory, had built an apparatus for producing half-million-volt protons, and had proved them able to transmute. The proton-streams had not, however, been greater than five microamperes (one microampere or $\mu a = 6.28 \cdot 10^{12}$ protons per second). Next Oliphant and Rutherford, inspired by that result,

proceeded to build an apparatus in which the maximum voltage should not go above a quarter of a million, but in recompense the stream of protons should be raised to a hundred microamperes. Another alteration: previously the stream had been a mixture of protons with heavier ions and neutral particles—now Oliphant and Rutherford introduced a magnetic field, adjustable and strong enough to bring either the protons or the more massive ions separately against the target. The magnetic field also assures that all the particles striking the target shall have nearly the same speed, something not completely guaranteed by the constancy of the voltage.

The scheme of this device is sketched in Fig. 2, where the course of the proton-stream is traced (rather too pictorially, I fear!) in a sweeping arc from its origin in the discharge-tube *R*, to the target *T* where the element to be transmuted awaits the impacts. In the discharge-tube all the parts are of steel, and the block *C* and cylinder *B* conjointly form the cathode, while the oil-cooled block *D* and cylinder *A* conjointly form the anode. This unusual material and structure are required partly to minimize cathode-sputtering, and partly to take care of the great amount of heat which is steadily developed in the tube, inasmuch as for the best supply of protons a voltage of 20,000 and a current of many milliamperes are demanded. Something like a twentieth of the current in the discharge is borne through the hole in the cathode by protons (and other positive ions of greater mass, if such there be); and in the space between *C* and *E* these particles receive from an electric field most of the kinetic energy with which they strike the target. In this space and in the region where the magnetic field comes into play, the density of the gas must be kept extremely low, despite the fact that there is an open passage into these spaces from the discharge-tube where the density must always be great enough to sustain the discharge and the supply of protons. This is a task for powerful pumps, which must be kept continuously at work pumping away from the lower chambers the gas which is steadily draining out of the discharge-tube through the hole and must as steadily be replenished by feeding fresh hydrogen in from above. It is no small part of the difficulty of the experiment, that the discharge-tube and the source of its power and the source of its hydrogen must all be maintained at scores or hundreds of thousands of volts above the potential of the ground, in order that the observing-apparatus may itself be at ground-potential. The transmutations are observed by detecting the fragments which issue through the very thin mica pane of the window *W*.

Until the building of this apparatus proton streams had been so scanty, that to bring about disintegrations in measurable number it had been needful to project the protons against thick layers of dense matter. In going through these layers they were slowed down and

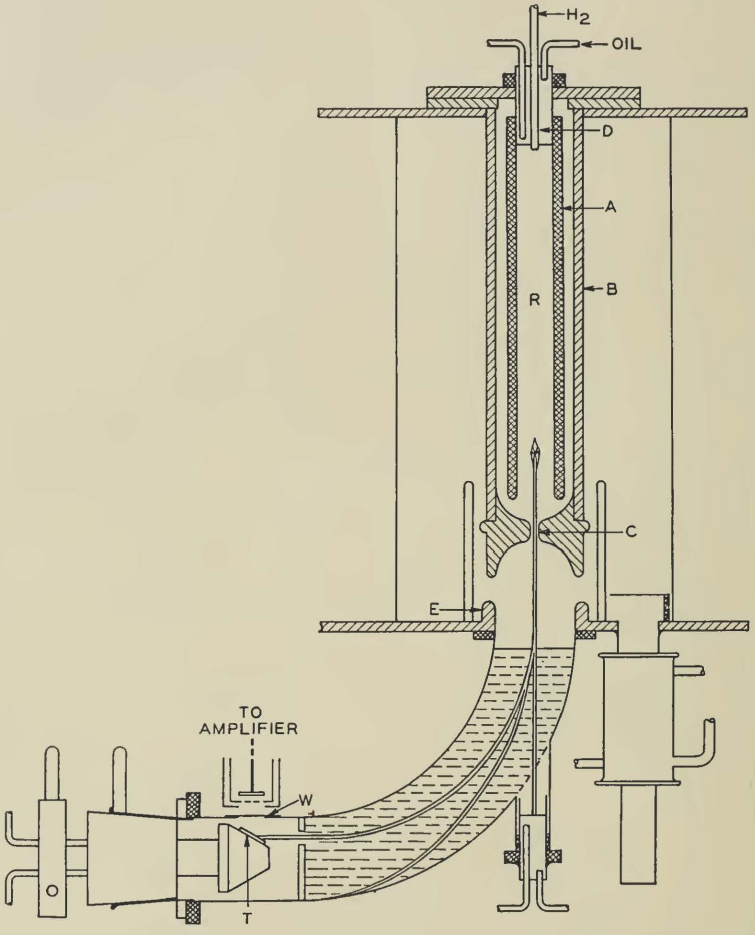


Fig. 2—Apparatus of Oliphant and Rutherford for producing transmutation by intense streams of protons. (*Proceedings of the Royal Society*)

stopped, and there was no direct way of telling whether the observed transmutations were achieved by protons of full speed, or by those which had already lost some energy, or both. Moreover if the energy of the bombarding particles was raised, the number of disintegrations infallibly went up, but a part of this increase (and sometimes the

whole of it) was certainly due to the fact that the faster particles went farther into the layer and struck more nuclei. There is no need to labor the point: it is obviously desirable to do the experiments with a film so thin that each oncoming proton either strikes a nucleus with its full and unabated initial energy, or else goes through the film and away without any impact at all. This ideal was closely approached by Oliphant and Rutherford, when they got countable numbers of fragments from films of lithium and boron (deposited on blocks of steel or iron) which were so thin as to be invisible, and of which the latter was known to consist of only seven-tenths as many atoms as would suffice to cover the iron surface with a single monatomic layer. (The curves of Fig. 16 were obtained with these films.)

This is a success which proves it possible to investigate films consisting each of only a single isotope of the element in question; for feeble as are the ways of separating isotopes in all but a few very favorable cases, they yet are powerful enough to produce pure monatomic layers. This article will amply show how valuable will be the privilege of getting data from a single isotope, of lithium or boron for example; already there are several cases of important antagonistic theories, the decisions between which will be given once and for all by such data.

The apparatus devised by E. O. Lawrence and developed in his school at Berkeley is of a singular ingenuity, inasmuch as in it ions are accelerated until their energies are such as would be derived from an unimpeded fall through a potential-difference of literally millions of volts, and yet the greatest voltage-difference at any moment between any two points of the apparatus is only a few thousands. It owes its elegant compactness to the lucky fact that when a charged particle is moving in a plane at right angles to a constant magnetic field, and consequently is describing a succession of circles, the time which it takes to describe a single circle is the same whatever its speed. One sees this readily by writing down the familiar equation,

$$mv^2/\rho = Hev/c,$$

in which e, m, v stand for the charge (in electrostatic units), mass, and speed of the ion and ρ for the radius of curvature of the circle, and on the right we have the force exerted by the magnetic field H upon the ion and on the left the so-called "centrifugal force" to which it is equal. The radius ρ varies directly as v, but the time $T = 2\pi\rho/v$ which the ion takes to describe a circle is independent of v. This is no longer true if the ion is moving so fast that the foregoing classical equation must be replaced by its relativistic analogue, but fortunately

the desired results are attained without forcing the speed to such heights.

Suppose now that while the ion is describing its consecutive circles each in a time T, its speed is suddenly increased; it continues to make circles, of a larger radius but with the same duration. Suppose that the increase occurs twice in each cycle, at intervals $T/2$; the path is a succession of semicircles each broader than the one preceding but all described in equal time. Now we arrive at Lawrence's device. The ions circulate in a round flat metal box, sliced in two along one of its diameters (Figs. 3, 4); and every time that one of them passes from

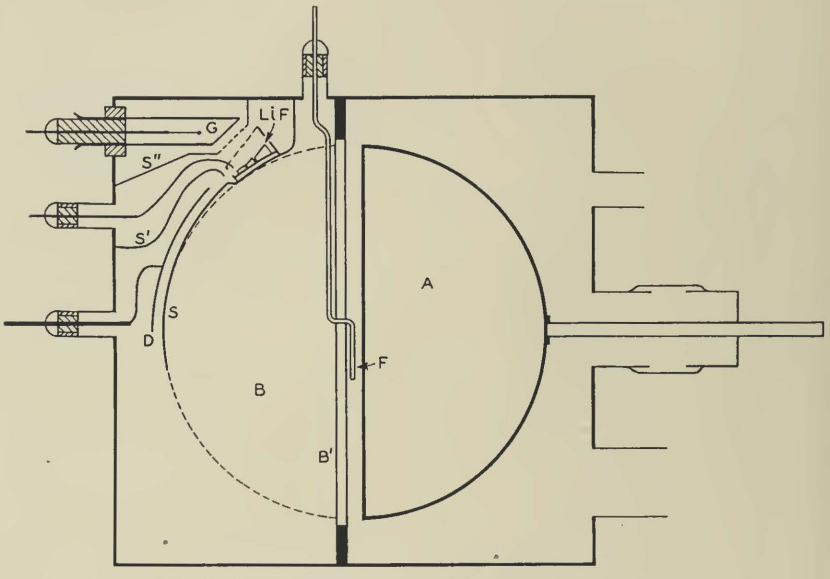


Fig. 3—Diagram of the Lawrence apparatus for cumulative accelerations of protons and other ions with auxiliary magnetic field. (After Henderson)

within half-box B to within half-box A it is accelerated by a voltage-difference existing between B and A, and every time that it passes from within A to within B it is again accelerated by a voltage-difference between B and A. Of course if this voltage-difference remained the same, the ion would lose at the latter passage just the energy which it gained at the former; but here is precisely the distinctive feature of the method: *the potential-difference between the two half-boxes is reversed in sign between each two consecutive passages.* So rapidly do the successive passages follow on one another, that if the intervals between them were unequal it would probably be impossible to devise any mechanism that would perform the potential-reversals at the proper

moments, but the felicitous law of the equality of the intervals makes all easy—all that is needed is to connect an oscillator of the proper frequency (determined by the strength of the magnetic field and the charge and mass of the ions) across the pair of half-boxes.[10]

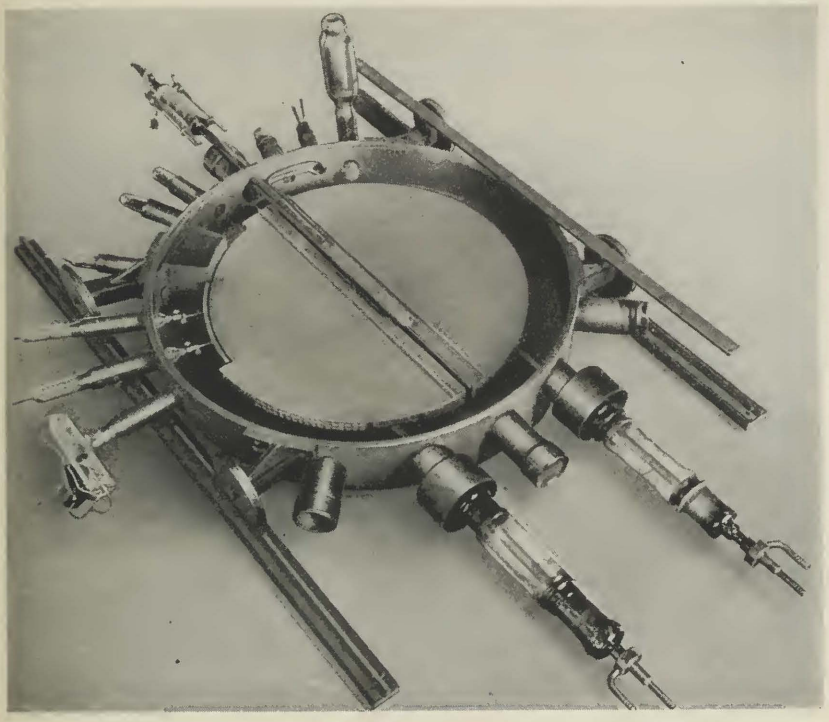


Fig. 4—Photograph of the apparatus sketched in Fig. 3. (E. O. Lawrence)

The sketch of Fig. 3 is of the apparatus wherewith Henderson observed the transmutation of lithium by protons (pp. 140–142). It is filled with hydrogen of a low density, so that electrons proceeding from the hot filament F at the center ionize the gas and produce a sufficient number of protons. These are whirled around and around in ever-widening semicircles, till after a number of circuits which may be as high as one hundred and fifty they arrive at the boundary of the

[10] One may do without the magnetic field, arranging to have the ions proceed along a straight line and to accelerate them at definite points along that line, by voltages produced in rhythm by an oscillator; the points of application of the voltages must be spaced according to a particular way, and the apparatus is inconveniently long, being longer the lighter the ion; it has been successfully employed with mercury ions by Lawrence and some of his colleagues.

half-box B opposite the charged electrode D, which deflects them enough to bring them into the cup-shaped receptacle at the far end of which the crystals of lithium fluoride are spread. The fragments of lithium nuclei which are observed are those which escape to the left in such directions as to enter the Geiger counter G. In this apparatus the radius of the outermost circle was 11.5 cm., the magnetic field 14,000 gauss; the potential-difference between the half-boxes never attained as much as 5000 volts, but it was reversed 4.2 millions of

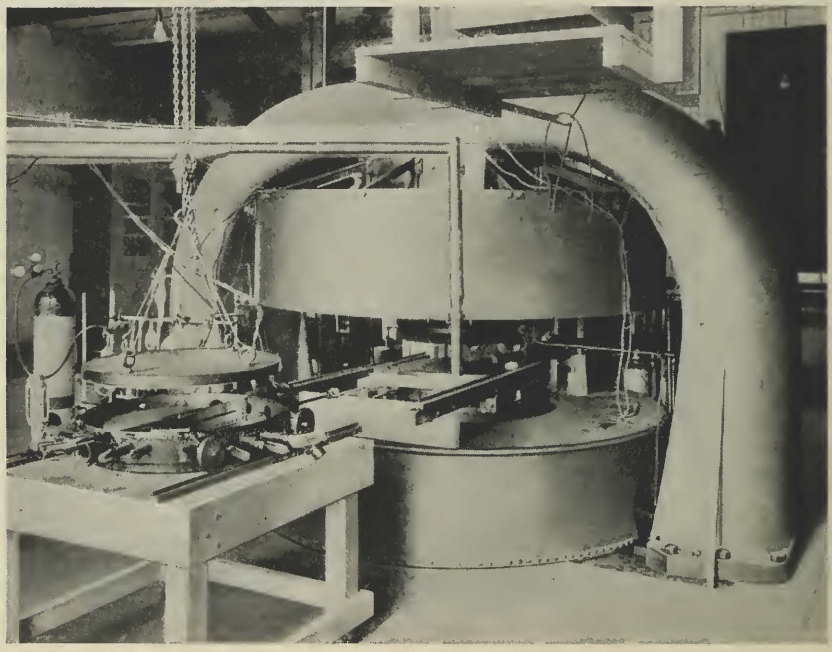


Fig. 5—The Lawrence apparatus for cumulative acceleration, beside the colossal magnet between the poles of which it is placed. (Lawrence)

times in a second, so that after three hundred reversals the protons had arrived at the limit of the box and were ready to strike the lithium nuclei with an energy of 1.23 MEV. By using a bigger pair of half-boxes in a more extensive magnetic field, this energy could be augmented; and by viewing the size of the magnet in Fig. 5 one sees what an augmentation is now imminent. The currents were inferior to those achieved by the apparatus of the Cavendish school, being mostly of the order of a few millimicroamperes (one millimicroampere or $m\mu a = 10^{-3}\ \mu a$).

Detection and Measurement of Transmutation

While thus the scope of transmutation has been so vastly extended in the past eighteen months, there is one limit which has not yet been passed. *No product of transmutation has yet been detected by any chemical means.* Many a plate of metal has been bombarded with protons or with alpha-particles, but no man has seen it change into a plate of another metal, nor alter in any of its chemical properties; many a tubeful of gas has been bombarded, but no man has observed the qualities or the spectrum-lines of another gas appearing in the content of the tube. All that is ever observed is an outpouring of material particles from the piece of bombarded matter; particles of such a nature, that they must come from the nuclei of the atoms. One expects this statement to go out of date from one morning to the next; but at the moment of this writing it is still as true as it was in 1919 when Rutherford first disintegrated nuclei, and broader in one respect only. From 1919 until 1932, one would have said "charged particles"; but since the winter of 1932, it is known that either charged particles or uncharged may be driven out of nuclei, by the appropriate impacts.

Thus there are two great experimental problems, and not one only; beside the problem of producing the streams of bombarding corpuscles, there is that of detecting and of recognizing the particles which fly forth from the bombarded nuclei—the "fragments," I will say. There is a grave objection to this term, and to the common name "disintegration" for the process. Both suggest a picture of the nucleus as a structure of pre-existing pieces which the impact breaks apart and scatters. This picture is surely incorrect, for there are cases in which the fragments contain the susbtance of the impinging corpuscles. In fact, if we define "fragment"—as we should—to include the part which in most of the experiments does not escape from the target bulk, we may say that this kind of case is frequent, and perhaps indeed that there is no other kind! Nevertheless we seem to be unable to get along without the words "disintegration" and "fragment."

For detecting protons and more massive fragments which are charged, there are three methods.

The *first method* (A) is that of observing the scintillations, which fast charged particles produce when they impinge on fluorescent screens. This is the classic and historic method, by which were made the earliest proofs of transmutation by impact of alpha-particles (which I described at length in the earlier article) and also the earliest proof of transmutation by protons. Of late years this method has been largely displaced by the others. Few people outside of the Cavendish

Laboratory and the Institut für Radiumforschung in Vienna have ever submitted themselves to the long, tedious and nerve-racking process of counting thousands of dim flashes for periods of hours in darkened rooms with dark-adapted eyes; and if two disagreed as to what was observed, there was no objective way of deciding between them. The newer methods abolish this strain; they can readily be so shaped as to leave a permanent record, which anyone may consult and analyze for himself; and they are capable of measuring the ionizing power of the fragments. Nevertheless the eldest method still retains the unique advantage that no barrier whatever, not even a gas, need intervene between the detecting screen and the source of the fragments; and also it is often employed by those accustomed to scintillations as a check upon the others.

The *second method* (B) is that of the expansion-chamber or cloud-chamber of C. T. R. Wilson, whereby the tracks of ionizing particles across a gas are made visible by droplets of water which condense upon the ions. This is the splendid invention which is the joy of all who write or lecture on atomic physics, since it enables them to decorate their exposition with pictures which make real the things of which they speak. It has virtue for the investigator also, especially since it may show in a single vivid photograph how many fragments there are formed in a single process, what are the directions in which they fly away, and how far they are able to travel through the gas. The curvature of the track in an applied magnetic field supplies the value of the momentum of the particle which made the track, if the nature of the particle be known; and this last may often be guessed from the aspect of the track, or assured by independent data. The major disadvantage of the method is, that the apparatus records only the particles which fly off during about a hundredth of a second, and then lies idle for several seconds or even minutes while it is being prepared for its next brief interval of effectiveness.

The *third method* (C)—or group of methods rather, for the variants are legion—is the detection by purely electrical methods of the ions which the fragments produce as they shoot across the gas of an ionization-chamber. A fast-flying charged particle loses on the average 30 to 35 electron-volts for every ion, or rather every ion-pair, which it produces.[11] To see the utility of this theorem, turn it around; the number of ion-pairs produced by a fast charged particle going through a gas is about a thirtieth of the number of electron volts which it loses in its transit. A fast alpha-particle, such as are spontaneously emitted by radon, or constitute the fragments springing out

[11] "Electrical Phenomena in Gases," pp. 52, 70–71.

of lithium bombarded by protons, has about eight million electron-volts; if it enters an ionization-chamber filled with gas so dense that it is brought completely to a stop, the ion-pairs appearing are about a quarter of a million. The upper limit occurring in practice is possibly twice as high, but is very rarely met with; there is no lower limit, but every incentive to push downward and ever downward the least amount of ionization which can be detected.

Twenty-five years ago, it would have been impossible to detect by electrical means so few as a quarter of a million ions. (The total number produced *e.g.* by an alpha-particle was determined by measuring the total ionization produced by a known and very great number of particles.) This problem was however destined to be solved in many ways, which I will group under four headings:

(C1) By arranging to have each particle touch off a brief but violent discharge, something like an invisible spark, in the gas of the ionization-chamber. There is a strong electric field applied between the electrodes of the chamber, whereby the "primary" ions which the particle forms as it travels across the gas are caused to produce (directly and indirectly) vast numbers of extra or "secondary" ions; and these suffice to make a sensible effect in the external circuit. The idea was first put into practice by Rutherford and Geiger in 1908, and the scheme is commonly known by Geiger's name. One of the electrodes must be either a fairly sharp point or a fairly thin wire, and there are a number of empirical rules (some partially understood, some not at all) about the size and shape of the chamber, the proportioning and the conditioning of the electrodes, the nature and the purity and the density of the gas, and the magnitude of the field. The voltage across the gas must lie within a definite range, often pretty narrow; if it is lower the particles do not produce discharges, if it is higher a single discharge may last indefinitely. The ratio of the number of secondary to the number of primary ions is usually not constant and usually not measured; most of the various forms of the device serve solely to detect or count the particles, and they are known as "Geiger counters." Often a loudspeaker is connected into the circuit of the ionization-chamber, and each discharge produces an audible clack, so that by the Geiger method one hears the passage of a corpuscle as by the Wilson method one sees it. Sometimes the discharges are recorded and the record examined at leisure.

(C2) By modifying the foregoing scheme so that the number of secondary ions shall be proportional to the number of primary ions, and a measurement of their total charge shall give at least a relative value of the ionizing-power of the traversing particle. This is a

recent achievement of Geiger and Klemperer. The process may be called *internal amplification* of the primary ionization, the amplification being in a constant proportion, or, as people carelessly call it, "linear."

(C3) By developing an electrometer or electroscope so sensitive that it is able to detect and even measure the total charge of a few thousands of ions, without amplification. This was first achieved, or at any rate applied to transmutation, by G. Hoffmann of Halle, and his associate Pose; the latter was able to observe fragments of aluminium nuclei (ejected by alpha-particles) which produced as few as three thousand ion-pairs. The major difficulty seems to be, that the electroscope takes a large fraction of a minute to perform its deflection and then recover its readiness to respond to another particle. Pose in his experiments observed only some thirty fragments to the hour.[12]

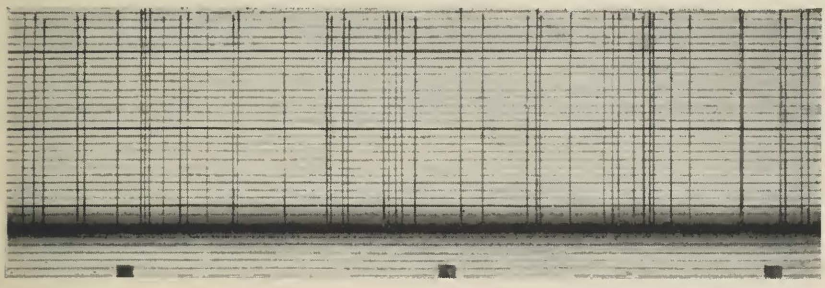
(C4) By applying *external amplification* to the feeble impulse which the primary ions due to a single particle produce in the external circuit, and which is imperceptible to an electroscope of normal and convenient quickness of response. This is done by developing the superb techniques of amplification which modern vacuum-tubes have rendered feasible, and like the three foregoing schemes is an achievement of the last few years, having been carried on especially by Wynn-Williams of the Cavendish Laboratory and Dunning of Columbia.

I show as Fig. 6 three records made with Dunning's apparatus, wherein every vertical line is due to an ionizing particle, and is proportional in length to the number of ions which the particle produced in crossing a shallow chamber.[13] The lines of great and nearly uniform length which appear in record (*a*) are due to alpha-particles from polonium; these all had nearly the same speed and were moving in nearly parallel lines when they entered the chamber, and it is evident that in crossing the gas they all made nearly the same amount of ionization; they left with a good deal of their initial kinetic energy unspent. The lines in record (*b*) are caused by protons; their diversity in length is chiefly due to the wide variety of speeds which the protons had when they entered the chamber, for these were fragments of the disintegration of aluminium by alpha-particles, and therefore had a broad distribution-in-speed (page 147). As these words imply, and as I will stress presently, the ionization produced by a charged particle

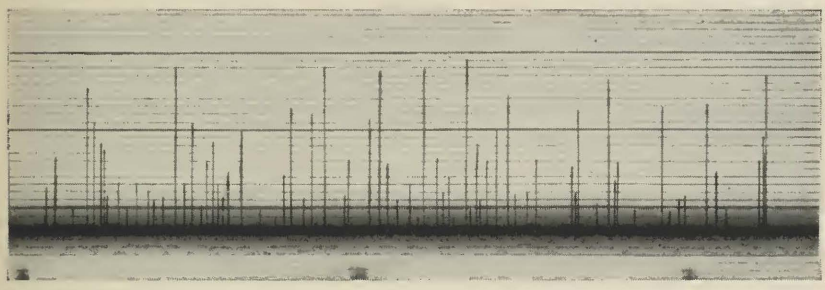
[12] It deserves to be recorded that in their blank experiments, Hoffmann and Pose during one research observed deflections at the average rate of 1.22 per hour, but observed altogether 197 of them!

[13] I am much indebted to Dr. Dunning for these pictures, made especially for this article. He writes of (*b*): "The minimum amount of ionization detectable here is well under 1000 ions; probably it could be pushed down to 250 ions." Consecutive dots at the bottom of each record mark off the minutes.

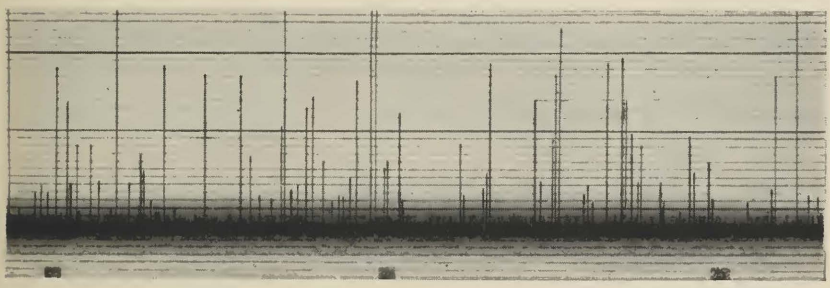
of given kind depends on its speed; the greatest amount which (in this particular chamber) a proton could ever produce, with its most favorable speed, is indicated by the longest lines in (*b*), and one sees that even these are definitely shorter than the lines in (*a*) due to



(*a*)



(*b*)



(*c*)

Fig. 6—Three records of the ionization produced by individual particles in a shallow ionization-chamber: (*a*) alpha-particles of nearly the same speed, (*b*) protons of various speeds, (*c*) particles of several kinds which had been set into motion by impacts of neutrons. The fogging along the base-lines is much fainter in the original records than in these reproductions. (J. R. Dunning)

alpha-particles.[14] The lines in record (*c*) were obtained when neutrons were traversing the chamber and a piece of paraffin outside of it; not they, but the charged nuclei which they strike and impel, are producing the record. Those lines which are longer than any in (*b*) are certainly not due to protons; they must be caused by recoiling nuclei of the atoms of the gas which fills the chamber (air), and which have various speeds because the neutrons strike them more or less glancingly (and probably do not themselves all have the same speed). The shorter lines are due in part to such nuclei, chiefly to protons ejected from the paraffin in such directions that they cross the chamber. Some are very short indeed, half-lost in the dusky haze due to the perpetual wiggling of the oscillograph mirror caused by gamma-rays; they are made by the fastest of the protons. Observations by expansion-chambers and with applied magnetic fields have proved this classification of the particles.

All of these methods are available for detecting charged particles which are protons or alpha-particles or corpuscles of a yet greater mass than these. For electrons the problem is harder.

An electron of given energy—say x thousands of electron-volts—is able to make roughly as many ion-pairs in a gas as could a proton or an alpha-particle of equal energy: that is to say, about $30x$. Nevertheless it produces much less ionization in an ordinary chamber than either of these last. This seeming paradox is due to the facts that the ion-pairs produced by the electron are relatively far apart and the loss of energy per centimeter of path is correspondingly low, so that in an ionization-chamber of reasonable dimensions and customary density of gas the traversing electron produces only a few hundreds or perhaps one or two thousands of ion-pairs before it reaches the opposite side of the chamber and plunges into the wall.

This is made evident by the Wilson method, the tracks of electrons appearing much thinner—less richly peopled with droplets, that is to say—than those of alpha-particles or protons. The expansion-chamber therefore is available for observing fast electrons, and so to a certain extent is the Geiger counter, which skilful observers can adjust so that it will react to these bodies. None of the other methods has yet been used with success. The ions produced by a single electron in an ionization-chamber are apparently too few to observe without amplification or even to amplify successfully, and the scintillations too faint. If one has neither expansion-chamber nor Geiger counter available, the only thing to be done is to measure the total ionization

[14] The contrast is much more striking than the records suggest, for the amplification was fourfold greater when (*b*) and (*c*) were made than when (*a*) was made.

produced by great numbers of electrons, and attempt to estimate these numbers. This is done in the study of the beta-rays or fast electrons emitted from radioactive nuclei, and in the study of cosmic rays; but the method has not yet been applied to the rays emitted from atoms undergoing transmutation by impacts, and apart from Joliot's observations on positive electrons (page 102 *supra*) nothing yet is known of any electrons which may be emitted by these.

Since individual electrons are so difficult or impossible to observe by the customary methods, one might suppose that at any rate they never annoy the observer. This unluckily is not so; for if electrons are numerous, they may keep the electrometer needle (in the method C4, for example) in a perpetual tremor, producing a so-called "background" over which even the strong sharp impulses due to alpha-particles or protons may fail to stand out. It is even possible for a chance coincidence or near-coincidence of several electrons to make a record which cannot be disinguished from that of a single particle of greater ionizing power. The scintillation-method suffers from a like defect, for if the fluorescent screen is heavily bombarded with electrons —or with gamma-rays, which liberate electrons from the fluorescent stuff and the surrounding matter—it shines all over with a feeble glow, against which the flashes made by more massive ions are difficult to discern. The most casual student of transmutation cannot fail to notice that polonium is generally used, of recent years, as the source of alpha-particles for bombardment. Probably he infers that either it is especially abundant or else supplies especially fast particles. But in both respects polonium is inferior to another customary source, radon mixed with its descendants radium A and radium B. It is used because it emits no gamma-rays but feeble ones of low penetration, whereas the other source pours out abundant and powerful photons which flood any nearby ionization-chamber with electrons and confuse the electrometer. Dunning's amplifying circuit, whereby he detected charged nuclei set into motion by neutrons, was so devised as to discriminate against the feeble but many impulses produced by these electrons and in favor of the occasional stronger ones produced by the massive particles; and this device enabled him to use a source of the latter type providing fifty times as many alpha-particles to engender the neutrons, as the largest amount of polonium ever employed.

Neutrons, I recall, are detected by observing the protons and more massive nuclei which they convert by impact into fast-flying ionizing particles, and photons by observing the electrons on which they have the like effect; the problems of getting the data are thus not new, it is the problem of interpreting them which is changed.

The next important question is, how the fragments are identified as protons, or as alpha-particles, or otherwise, from the data. Few as yet are the cases in which the identification is full and undeniable. In the earlier paper I described Stetter's measurements of charge-to-mass ratio for the fragments produced by impacts of alpha-particles against boron, carbon, fluorine and aluminium, which gave values identical with that for protons within the observational uncertainty of five per cent. As for the fragments produced by impacts of protons, the best direct evidence is that which appears in Fig. 7. Cockcroft

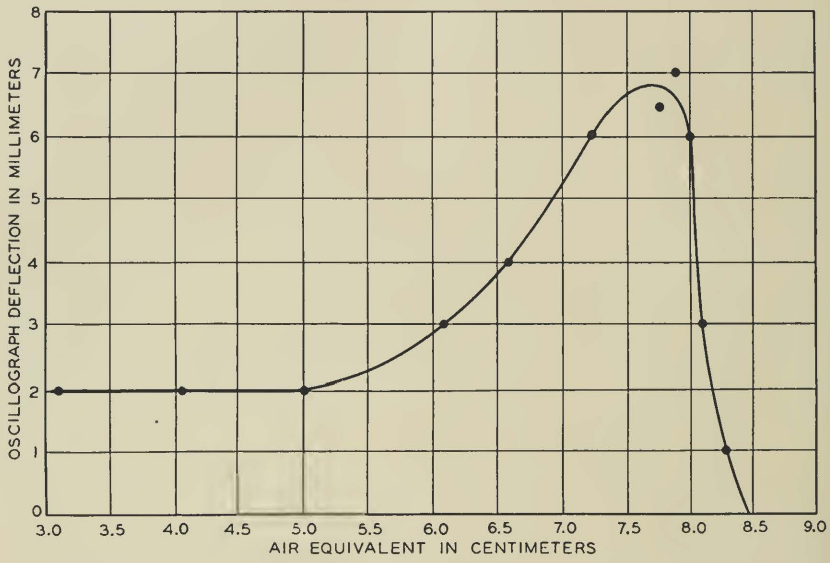


Fig. 7—Ionization produced in a shallow chamber by fragments (of the transmutation of lithium by protons) which have passed through screens of various thicknesses. (Cockcroft and Walton)

and Walton had an ionization-chamber only 3 mm. across, and the fragments from bombarded lithium traversed it completely, producing a few thousand ion-pairs apiece which were detected and measured with the aid of an amplifying circuit of Wynn-Williams according to the method C4. When mica sheets were interposed in the path of the fragments from the lithium, they were slowed down but still kept energy enough (so long as the sheets were not too thick altogether) to travel across the chamber; and the curve of Fig. 7 represents the number of ion-pairs produced per fragment, as function of a quantity x proportional to the thickness of mica which the fragments have

traversed ("air-equivalent" of the mica, p. 127 *infra*).[15] The point is, that exactly the same curve was obtained when a beam of alpha-particles was projected through the same thicknesses of mica into the same chamber. Mere similarity in the shape of the curves would prove nothing, for this is the shape obtained with all kinds of charged particles, electrons and protons and more massive charged nuclei; in particular, every such curve rises from zero to a maximum and thereafter descends continually as the energy of the particles is raised indefinitely upward from the least value sufficient for ionization.[16] However, the ordinates of the curve of Fig. 7 are *equal* to those of the alpha-particle curve, and about four times as great as would have been observed with protons; and this it is which proves the fragments to be alpha-particles. Almost as good a proof could be made by two measurements: by measuring the range of the fragments and the total ionization produced by any fragment in a chamber deep enough to swallow it up, and comparing the latter datum with the ionization produced in the same chamber by an alpha-particle of equal range. This proof, or some other substantially like it, has been adduced in certain cases. When alpha-particles are the agents of the transmutation, the same test has proved in several cases that the fragments are protons. In some cases the test has not yet been applied.

I have already had to speak of interposing mica in the path of the fragments, in order to learn something about them. This is a procedure with which it is necessary to be familiar. It would be very pleasant indeed to be able to apply electric and magnetic deflecting fields to a narrow stream of fragments all flying in the same direction, for one could then spread it out into a velocity-spectrum, and not only identify the corpuscles perfectly but also determine their distribution-in-range, which as we shall presently see is of the first importance. This has not yet been done, partly (I presume) because of the high fieldstrengths that would be needed, chiefly because the available streams of particles are too scanty. It will be a happy day when at last we get streams of fragments so intense that they can be dispersed into a velocity-spectrum which will appear imprinted on a photographic film, as has been feasible for years with beta-rays. For the time being we must be content with curves such as many figures in this article display, Figs. 8 and 9 and 11 for example.

[15] The quantity plotted as ordinate is obtained from such records as those of Fig. 6, in which every fragment produces a vertical line. Cockcroft and Walton observed many such lines for each thickness of mica, and ascertained in each case the most frequently-occurring value of line-length.

[16] "Electrical Phenomena in Gases," pp. 40–44, 70–71. Such a curve as that of Fig. 7 is sometimes called a "Bragg curve."

These are curves in which the abscissa stands for the thickness of a special kind of matter (air of a standard density) interposed in the path of the fragments, and the ordinate for the number of fragments detected on the far side of that matter; I will call them "integral distribution-in-range" curves representing the number $f(x)$ of particles able to traverse thickness x. Were one to differentiate them, one would get the "differential distribution-in-range" curves, representing a function $f'(x)$ such that $f'(x)dx$ stands for the number of particles able to traverse thickness x but not additional thickness dx—the particles which are said to have "ranges" between x and $x + dx$.

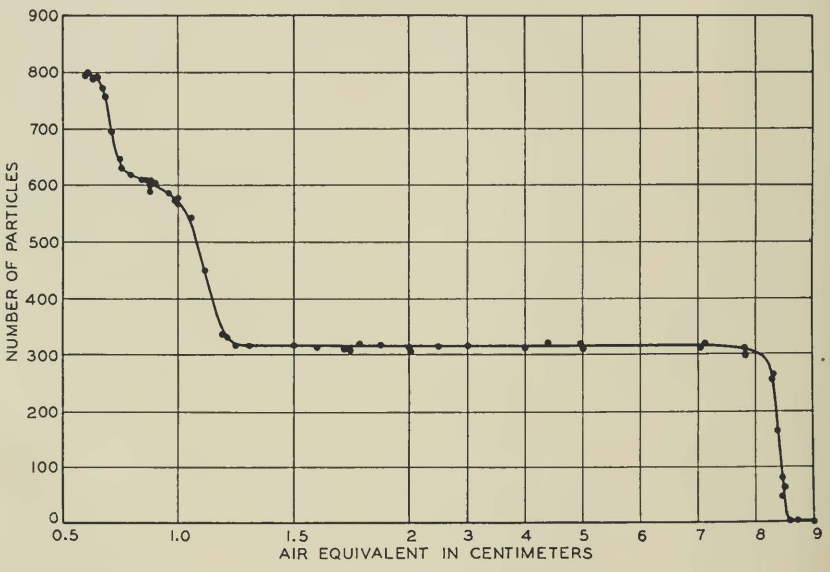


Fig. 8—Integral distribution-in-range curve of the fragments resulting from bombardment of lithium by protons. (Oliphant Kinsey & Rutherford)

These, however, are usually not plotted,[17] and one must accustom himself to draw the proper inferences from the integral curves.

The clearest of these to read are those which are shaped like a staircase, with steep rises connecting horizontal parts called paliers or plateaux. A steep rise extending over a narrow interval of x signifies a "group" of fragments all having ranges close together. A plateau extending over a broad interval of x signifies that no particle has a range comprised anywhere in this interval. An integral curve in the form of a staircase therefore implies the analogue of a line-spectrum,

[17] One of the rare examples is reproduced in "Transmutation," *B. S. T. J.*, Vol. X, p. 650 (Oct. 1931), from the work of Bothe and Fränz.

the particles being classifiable into groups each with its characteristic speed. But if in such a curve there is a long sloping arc (as in Fig. 9), it implies the analogue of a continuous spectrum, there being particles of all ranges over a notable interval.

The "stopping" or "absorbing" screens which are used in determining these curves are usually sheets of mica or of aluminium. The curves are not however plotted against the actual thickness of the interposed strata of mica or whatever else the substance may be, but against the "air-equivalent" or thickness of the stratum of air of standard density [18] which is known by separate experiments to have the same effect in slowing down and stopping charged particles, the

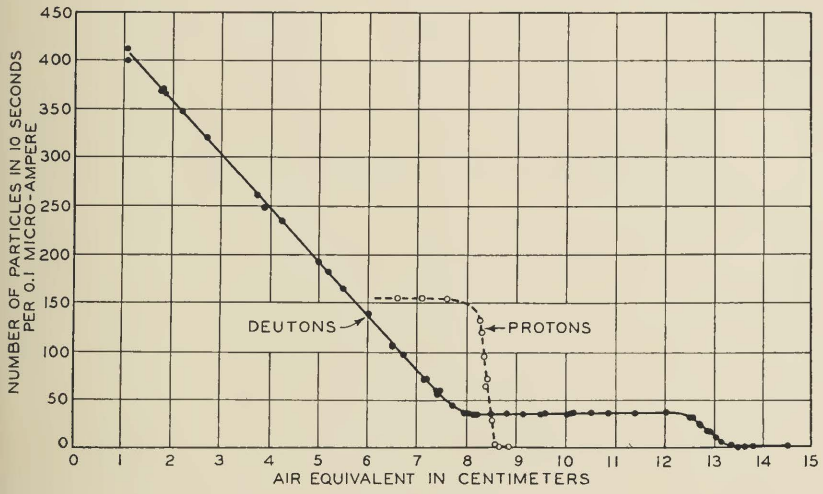


Fig. 9—Integral distribution-in-range curve of the fragments resulting from bombardment of lithium by deutons. (Oliphant Kinsey & Rutherford)

same "stopping-power." It is the air-equivalent which is the quantity x of the preceding paragraphs and the abscissa (often termed "absorption") of Fig. 8 and nearly all other such figures. The ratio between the actual thickness of a layer of matter and the equivalent thickness of air is roughly (but only roughly) the reciprocal of the ratio of their densities. The sheets of metal or of mica used in the experiments are therefore very thin (it has been possible to make screens of mica so tenuous that their air-equivalent is only 0.15 mm.) and the thinnest must be bolstered up by stiff metal grids, of which the wires block a considerable fraction of the beam. It is also possible

[18] There are unluckily two standards of density, one being that of air at 0° C. and 760 mm. Hg, the other that of air at 15° C. and 760 mm. Hg; see "Transmutation," footnote on p. 643, *B. S. T. J.*, Oct. 1931. The latter is used in this article.

to use air (or some other gas) of adjustable density; when the scintillation-method is employed, the gas may fill the entire space between the source of the fragments and the fluorescent screen; with other methods of detecting the fragments, it must be contained in a cell which the stream enters and leaves through windows of mica or similar substance.

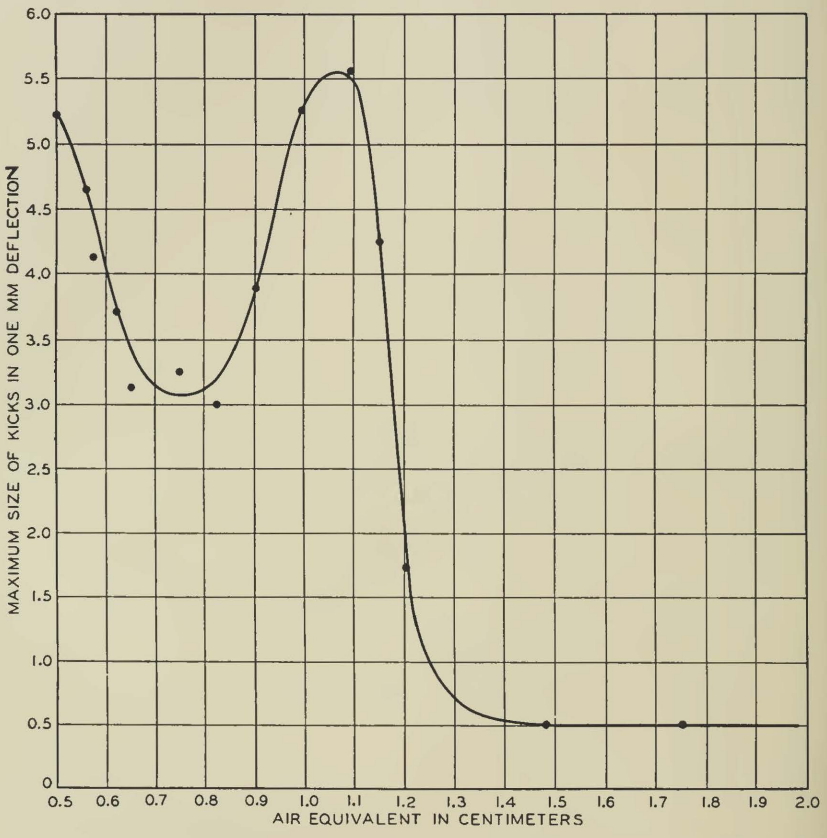


Fig. 10—Ionization produced in a shallow chamber by the least penetrating fragments from the transmutation of lithium by protons. (Oliphant Kinsey & Rutherford)

There is an interesting and important way of confirming the steps in an integral curve such as those of Figs. 8 and 9. Near the rise of such a step, the thickness of the intercepting matter is such that many particles are approaching the ends of their ranges when they emerge from the last of the screens. Suppose that this last screen is adjoined by a very thin ionization-chamber, like that with which the curve

of Fig. 9 was obtained. Let the air-equivalent x of the total thickness of the screens be varied, and let the average number of ions produced per particle in the chamber be measured and plotted as function of x. Recalling Fig. 7 and what was said in respect to it, the reader will see that the resulting curve should have a peak wherever the integral distribution-in-range curve has a step. This has been verified several times, and there are cases in which these peaks have been taken as

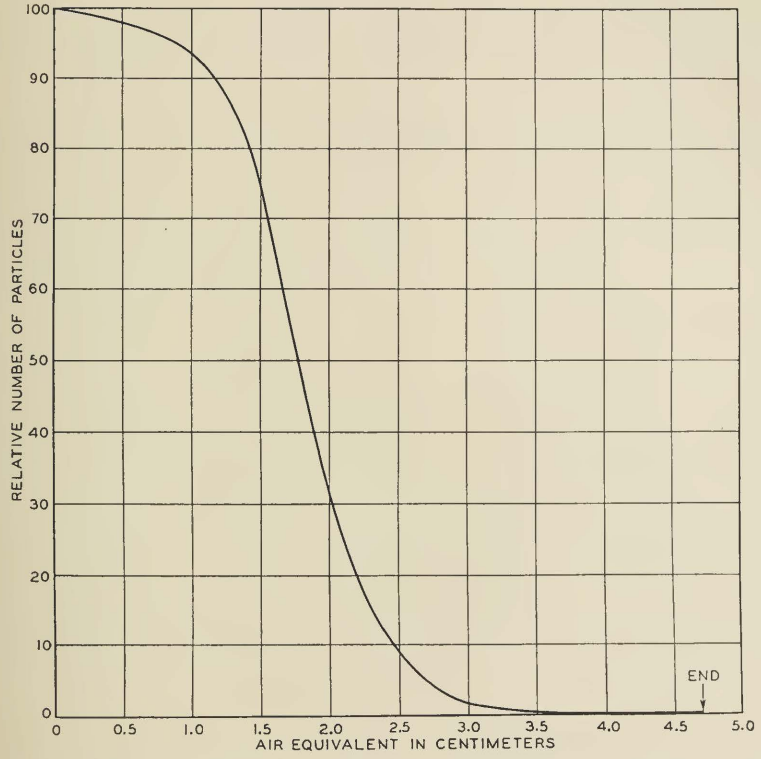


Fig. 11—Integral distribution-in-range curve of the fragments resulting from the bombardment of boron by protons. (Oliphant & Rutherford)

clearer evidence for the existence of groups than the shape of the integral curve itself (Fig. 10). Peaks may also appear in a curve of which the ordinate is the *total* ionization produced in the very thin chamber by all the fragments which enter it.

Anyone at all acquainted with physical experiments will readily suspect that the steps of actual integral curves "ought to be" steeper than they are. I mean: that he will form the hypothesis that perpen-

dicular rises would be observed instead of rounded-off and sloping ones, if only the pencil of fragments passing through the absorbers were ideally narrow and cylindrical, and were produced by bombardment of atoms with particles all of the same speed; and he will attribute the rounding-off of the steps to the facts that the fragments actually form a divergent and conical beam, and the atoms from which they come have been struck by impinging particles of diverse speeds. This idea is strongly supported by the facts that the steps are notably steepened when the divergence or "aperture" of the beam of fragments is reduced, and when the diversity of speeds among the bombarding particles is narrowed.

The former of these variables is controlled by the slits and diaphragms which bound the beam, and the latter by the thickness of the bombarded target whence the fragments proceed; for the bombarding particles are slowed down as they dive deeper into the target, and nuclei at different depths receive impacts of different energy, and thus there is a wider diversity of speeds among the particles when they finally make their impacts than there is among them when they start from their source. But as one cuts down either the thickness of the target or the aperture of the beam of fragments, one reduces the number of fragments which come to the detecting apparatus, and reaches a limit when this number becomes too small to be observed in any convenient time. Progress in approaching ideal conditions therefore depends on progress in multiplying the number of fragments by multiplying the strength of the bombarding beam. We may count on a yet greater steepening of such steps as those of Fig. 8, when the enormous streams of bombarding protons produced by Oliphant and Rutherford are applied to very thin films and the distribution-in-range of the resulting fragments is measured. A corresponding improvement of the curves obtained when alpha-particles are the bombarders is still in the not-immediate future. Whether under ideal conditions the steps would be absolutely perpendicular, and all the fragments of a group have exactly the same speeds, is not as yet to be safely inferred from the data.

There remains the great problem of converting distribution-in-range curves into distribution-in-speed or distribution-in-energy curves, and thus determining the energy or the speed of fragments belonging to a group of which the range is known. The recent developments of research in transmutation and in cosmic rays have elevated this to the rank of the major problems of physics. For alpha-particles of ranges of 8.6 cm. and less, it is practically solved by empirical means; for such alpha-particles are supplied in such abundance by radioactive

bodies that it has already been feasible to measure by deflection-methods the speeds corresponding to a large number of different ranges, and plot an empirical speed-*vs*-range curve which is fixed by so many points of observation that there is no important uncertainty in making interpolation between these. For alpha-particles of range superior to 8.6 cm., such as often occur among fragments of transmutation, it has heretofore been necessary to extrapolate; but very lately the empirical curve has been extended onward to 11.6 cm., thanks to a powerful new magnet at the Cavendish which is able to deflect the paths of alpha-particles of even such rapidity.[19] With protons our knowledge of the range-*vs*-energy relation is less extensive and less accurate, and an improvement thereof should be one of the first and most important by-products of the new methods for imparting high energies to ions. For charged nuclei of other elements than hydrogen and helium, relatively little is assured (what is known has been found out chiefly by Blackett and his school)[20] ; but this lack has not as yet been much of an impediment to the study of transmutation, except in certain cases involving impacts by neutrons.

Transmutation by Impacts of Protons and Deutons

The earliest element to be transmuted by protons in the laboratory —indeed the first to be transmuted by man with any agent other than the alpha-particle—was lithium. It was fortunate that Cockcroft and Walton began with this element, for its behavior turned out to be uniquely lucid. In most disintegrations, a single fragment is detected, and there must be a massive residue which remains unseen, staying hid within the substance of the bombarded target. But in some at least of the transformations which occur when lithium nuclei are struck by protons or deutons, there seems to be no hidden residue; every fragment is observed and recognized. These are processes of "nuclear chemistry" of which we fully discern both the beginning and the end; and they are described by the quasi-chemical equations:

[19] Rutherford et al., *Proc. Roy. Soc.* **139**, 617–637 (1933). The empirical curve departs slightly from a third-power law (range proportional to cube of speed) and the results are expressed by an empirical formula for the departure. See also G. H. Briggs, *Proc. Roy. Soc.* **139**, 638–659 (1933).

[20] See N. Feather, *Proc. Roy. Soc.* **141**, 204 (1933) and literature there cited. The observations are made upon tracks which appear in Wilson chambers when the contained gas is bombarded by alpha-particles, and which are the tracks of objects of atomic mass that have suffered violent impacts. It is presumed (though not always proved) that these objects are solitary or "bare" nuclei, not accompanied by any of the orbital electrons which attended them before the impacts. Some (but not all) of the data conform to the empirical rule that the ratio of the ranges of two nuclei of masses m_1 and m_2 and of charges Z_1e and Z_2e, when the two have the same speed, is $(m_1/m_2)(Z_1/Z_2)^{1/2}$.

$$_1H^1 + {}_3Li^7 + T_0 = 2{}_2He^4 + T_1, \tag{1}$$

$$_1H^2 + {}_3Li^6 + T_0 = 2{}_2He^4 + T_1, \tag{2}$$

of which the first has already appeared in Part I. of this article.

These are to be regarded as equations for mass and energy, owing to the equivalence of these two entities. Attached to the symbol of each atom are its mass-number as superscript and its atomic number as subscript (and, incidentally, every such equation must balance when considered as an ordinary equation in either the mass-numbers or the atomic numbers). The symbols T_0 and T_1 stand for the total kinetic energy of the particles *before* and the particles *after* the transmutation, expressed in mass-units. (I recall from Part I. that a mass-unit is one-sixteenth the mass of an $_8O^{16}$ atom, and that one million electron-volts is equal to 0.00107 of one mass-unit.) The other symbols then stand for the rest-masses of the nuclei of the atoms in question. It would be proper, and in accordance with the spirit of relativity, to leave out the symbols T_0 and T_1 and consider each of the other symbols as standing for the *total* mass of the nucleus, viz. the sum of its rest-mass and the extra mass resulting from its speed. When hereinafter the symbols T_0 and T_1 are absent from such an equation, the others are thus to be interpreted.

The suggestion thus is, that when a proton meets with a $_3Li^7$ nucleus or a deuton with a $_3Li^6$ nucleus, either process ends in the formation of two helium nuclei—alpha-particles—out of the substance of the original bodies. It is further suggested that these nuclei share kinetic energy amounting to T_1; and if they are emitted in directions making equal angles with that of the impinging particles—the "symmetrical case" which (as we shall see) is most commonly observed—they must share T_1 equally in order to assure conservation of momentum. Now the rest-masses of all the nuclei figuring in equations (1) and (2) are accurately known through the work of Aston and of Bainbridge. Taking them from Table I and substituting them into the equations, and using the electron-volt for our unit, we get:

$$T_1 = T_0 + 16.8 \cdot 10^6, \tag{3}$$

$$T_1 = T_0 + 22.2 \cdot 10^6, \tag{4}$$

in the two cases,[21] and therefore expect alpha-particles paired with one another, their kinetic energies amounting altogether to these values.

[21] For these numerical values and their uncertainties, see K. T. Bainbridge, *Phys. Rev.* (2), **44**, 123 (July 15, 1933).

It is the verification of these predictions which gives us such great confidence that we have recognized the processes which really happen.

I have already said how Cockcroft and Walton proved that the fragments, when lithium is bombarded by protons, are alpha-particles. The integral distribution-in-range curve of these fragments, obtained by Oliphant Kinsey and Rutherford with the apparatus of Fig. 2 and proton-currents running up to $50\mu a$, appears in Fig. 8; and that for the fragments created when deutons are used instead of protons appears in Fig. 9. In both of these one cannot but be struck by the beautiful long horizontal plateaux, and the sharpness of the steps which end them on the right. The groups of fragments of which these steps are the signs have ranges stated by the observers as 8.4 and 13.2 cm respectively, with uncertainties of $\pm$ 0.2 cm. (These figures are evidently taken from the bottom of the step, probably because it is assumed that under ideal conditions of narrow beam and thin bombarded film—the actual beam had a divergence of about 15° and the actual target was thick—the step would rise vertically from the point whence it actually begins to rise obliquely.) The corresponding energy-values are estimated as 8.6 and 11.5 MEV (millions of electron-volts) respectively; and as T_0, the energy of the impinging protons, is at most two-tenths of a million, these values may be compared directly with the halves of the numbers in equations (3) and (4). Meanwhile at Berkeley, Lewis Livingston and Lawrence were driving deutons with an energy of 1.33 MEV—no longer negligible—against lithium, and observing fragments with a range of 14.8 cm., corresponding to an energy of 12.5 MEV; and this is to be compared with half of 23.7 millions on the right-hand side of equation (4).

The agreement in the case of protons impinging on lithium is admirable, and well within the uncertainty of the data. The agreements in the cases of deutons impinging on lithium are ostensibly not so good, but this is not so serious as it seems at first glance, because of the required extrapolation of the range-*vs*-energy curve of alpha-particles (page 131), and because it is not always the "symmetrical case" which occurs. For the present there is no compelling reason to suppose that equation (2) is contradicted by the data.

A further point susceptible of test: if the processes described by equations (1) and (2) are actual, the alpha-particles of the stated ranges must be shot off in pairs, the two members of each pair flying off in almost opposite directions—in directions which would be exactly opposite were it not for the original momentum of the proton, but which because of that momentum must make with one another an angle slightly (and calculably) less than 180°. Cockcroft and Walton

made the test with a pair of Geiger counters set on opposite sides of the bombarded lithium, and got a positive result; but it is the expansion-chamber which is suited by its nature for supplying the most magnificent of proofs. To achieve this, one must put the bombarded target of lithium in the middle of the chamber, and photograph the tracks from above; and since the bombarding stream must come through vacuum while the chamber must be filled with moistened air, the target must be separated from the air by walls of mica thick

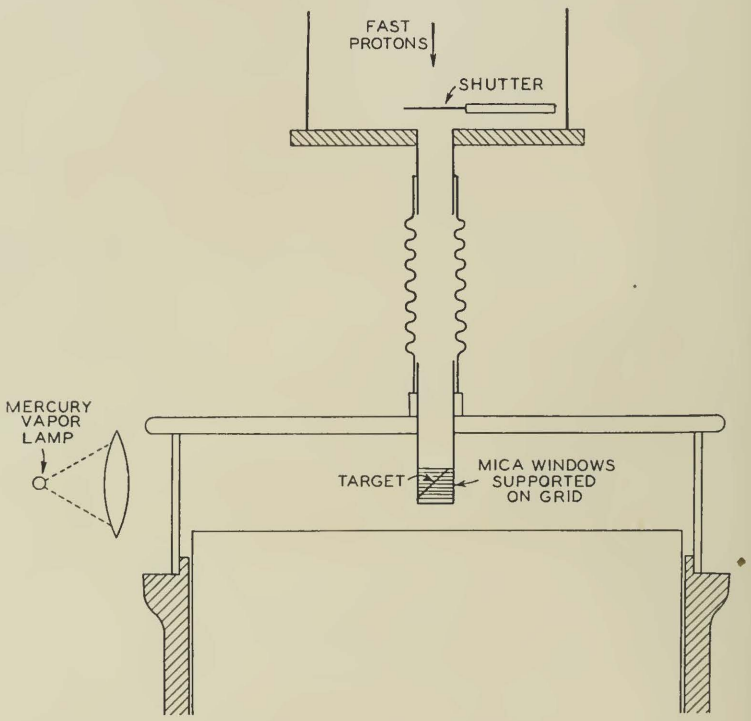


Fig. 12—Diagram of arrangement for observing tracks of fragments by the expansion-method. (After Dee and Walton)

enough to withstand the pressure and thin enough to let the fragments pass. The scheme is clearly depicted in Fig. 12. One notices that the design is such that the pairs which are observed are those of which the directions are nearly at right angles to the proton-beam—the "symmetrical case" aforesaid.

This experiment was first performed by Kirchner of Munich, who got several pictures of paired fragments from lithium bombarded by protons. Fig. 13 shows an example. (The third track is rather

annoying, but it was quite an achievement so to adjust the conditions as to get so few as three.) Many splendid examples have lately been published by Dee and Walton of the Cavendish, and Fig. 14 is outstanding among them because the bombarding stream was a mixture of protons and deutons, and the picture shows two pairs of fragments, one apparently due to each of the processes which I have been describing. Those of the pair marked b_1b_2 have the range of 8.4 cm. agreeing with equation (1), while those marked a_1a_2 go definitely farther and even escape from the chamber, which makes it impossible to measure their ranges. Dee and Walton therefore made the walls of the target-capsule thicker, so that more of the energy of the frag-



Fig. 13—Tracks of paired fragments, He nuclei resulting from impact of a proton on a Li^7 nucleus. (Kirchner; *Bäyrische Akademie*)

ments should be consumed in them; the pairs which were obtained with bombarding deutons now ended in the chamber and in the field of view, and their ranges agreed with the 13.2 cm. obtained from the curve of Fig. 9. At least two more of these pairs appear in Fig. 15. Verification of a theory could scarcely go further or be more vivid! Yet there is the additional point, that Kirchner found the angle between the paired paths in his pictures to differ from 180° by just about the amount required by the momentum of the proton.

However not every fragment observed when lithium is bombarded, either by protons or by deutons, results from these superbly simple interactions. Notice in Fig. 8 the two very much rounded steps, suggesting groups of short ranges (1.15 cm. and 0.65 cm.); these are confirmed by the maxima in the curve of Fig. 10 which has already

been explained (page 129). Only tentative theories of these have been made, and it would be of little use to expound them here.[22] Notice then in Fig. 9 the beautiful long *sloping* line adjoining the plateau, and implying a continuous distribution over a wide interval of ranges extending up to 7.8 cm. The numerous shorter tracks of Fig. 15 are due to particles belonging to this continuum. Observe last the integral distribution-in-range curve for the fragments from

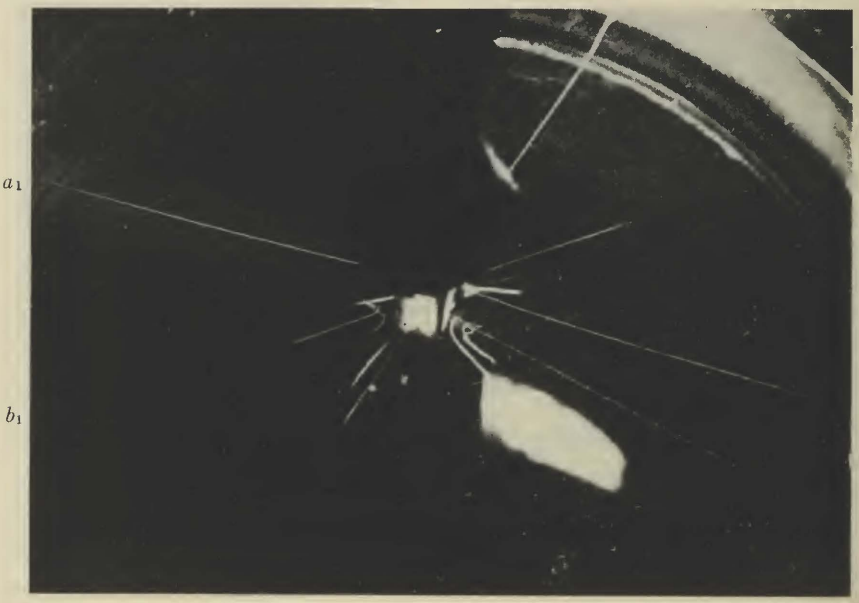


Fig. 14—Tracks of paired fragments, He nuclei believed to result from impact of a proton on a Li^7 nucleus and from impact of a deuton on a Li^6 nucleus. (Dee and Walton; *Proceedings of the Royal Society*)

boron bombarded by protons, Fig. 11; notice that it displays no definite step, but consists of a single sloping arc implying a continuum extending to an upper limit, which on a magnified curve is found to be at 4.7 cm.

It is now suggested that in both of these two last cases we have processes in which there are not two, but three final fragments:

$$ {}_1H^1 + {}_3Li^7 + T_0 = 2\,{}_2He^4 + {}_0n^1 + T_1, \tag{5}$$

$$ {}_1H^2 + {}_5B^{11} + T_0 = 3\,{}_2He^4 + T_1, \tag{6}$$

[22] Dee has just announced (*Nature*, **132**, 818–819; Nov. 25, 1933) that these short-range fragments are frequently paired. In doing the experiment he admitted the primary protons into the expansion-chamber through a thin mica window, the target being within.

the symbol ${}_0n^1$ in equation (5) standing for a neutron. When there are three fragments, conservation of momentum no longer demands that the available energy be equally divided among the three, but admits of an infinity of distributions. It is not difficult to find the highest fraction of T_1 which either of the two alpha-particles in case (5), or any of the three in case (6), may receive; this amounts to very nearly one-half in the former, to two-thirds in the latter case.

In equation (5) the rest-masses of all the charged nuclei are known; that of the neutron is still subject to some controversy, but if we



Fig. 15—Various tracks produced during bombardment of lithium by deutons. (Dee & Walton; *Proceedings*)

tentatively put Chadwick's value 1.0065 for it we get for $(T_1 - T_0)$ the value 16 millions of electron-volts. T_0 again is negligible, so that we are to compare half of this figure with the energy corresponding to the range 7.8 cm.—the right-hand end of the sloping part of the curve of Fig. 9—which is 8.3 millions. The agreement is entirely satisfactory. With boron the result is not so pleasing, for T_1 by equation (6) should be more than eleven millions, and two-thirds of this differs rather seriously from the energy-value corresponding to the end of the curve of Fig. 11, which is 6 millions. Kirchner got a photograph in which three coplanar tracks of the same appearance

diverge at mutual angles of 120° from a point in a boron target bombarded by protons, and Dee and Walton have noticed a number of trios of paths springing from such a target, but without being quite sure that they are not mere coincidences.[23]

Having now met with a case in which there may *not* be a balance between the two sides of such an equation as (6), we should now pause to inquire what can be done about such cases. Of course, such a disagreement might mean that the actual process is something entirely different from the one postulated in the equation, but it may not be necessary to make such a complete surrender of the theory. In equations (1) to (6), it is everywhere assumed that all the energy is retained by the material particles, in the form of kinetic energy or of rest-mass. Suppose that the process described by one of these equations, (6) for instance, is confirmed in every respect excepting that the final kinetic energy of the fragments is found to be less, by some amount Q, than the value of T_1 computed from the equation. One might then assume that the missing energy Q is radiated away in the form of one or more photons. Alternatively one might assume that the missing energy is retained by one of the material fragments in the form of "energy of excitation"; the rest-mass of the fragment, so long as it retained this energy and remained in the excited state, would then be correspondingly greater than its normal rest-mass, and the equation would be balanced if this abnormal value of mass were inserted into it in place of the normal one. Such explanations are frequently offered nowadays. They suffer, of course, from the disadvantage of being too easy; one can always postulate the necessary photons or excited states to explain any observed positive value of Q. But if they can ever be supported by independent proof of these excited states or photons, they will become much more convincing.

Lithium and boron are by far the best-studied of nuclei, in respect to their interactions with protons and deutons. It is true that our knowledge of the distribution-in-range curves of the fragments is still confined to comparatively low values of the energy of the bombarding particles, values less than 300,000 electron-volts. With higher energies it is to be presumed that the steps at the right-hand ends of the curves in Figs. 8 and 9 would move to the right, to the extent pre-

[23] If in the case of boron bombarded by protons it be assumed that two of the He nuclei fly off in directions making symmetrical angles $(\pi - \theta)$ and $(\pi + \theta)$ with the direction of the third, the distribution-in-θ of the disintegrations can be deduced from the curve of Fig. 14; it turns out that the most probable cases are those in which $\theta = 60°$ nearly, and all the three particles have nearly the same energy. A like deduction may be made for lithium bombarded by deutons, the neutron playing the part of third alpha-particle in the foregoing case; it is inferred that again the most probable types of disintegration are those in which all three share almost equally in the energy.

scribed by the increase of T_0 in equations (1) and (2); and so should the right-hand end of the sloping part of the curve in Fig. 9, and the extremity of the curve of Fig. 11. There is an indication of the first of these expected changes in the observation already quoted from Lewis Livingston and Lawrence, of 14.8-cm. fragments ejected from lithium by 1.33-MEV protons (page 133). We must wait for future data to test the others, and to see what happens to the heights of the steps and the general shape of the uninterpreted parts of the curves. Already however we have data bearing on the so-called "disintegration-function," or the relation of the total number of emitted fragments to the energy of the bombarding particles.

To speak of "total number of fragments" is to suggest too much. The present knowledge suffers from two limitations: the counts of fragments are made with apparatus which does not enclose the target completely and must be separated from the target by a screen, so that the fragments counted are only those which start off within a limited solid angle of deflections and have sufficient range to penetrate the screen. One generally makes a tentative correction for the former limitation, by assuming that the fragments go off equally in all directions and multiplying the number observed by the factor $4\pi/\omega$, where ω stands for the solid angle subtended by the detector as seen from the target. This factor may well be wrong, but perhaps does not vary seriously with the energy of the bombarding particles, so that at least the trend of the curve may not be distorted. For the latter limitation we have not the knowledge to make any allowance; it must always be stated that the count is of fragments having more than such-and-such a range, or such-and-such an energy. Every kind of device for observing transmutation suffers from some such lower limit, set either by the sensitivity of the device itself, or by the stopping-power of the wall which bounds it.

With their dense streams of protons and exceedingly thin films (page 113) Oliphant and Rutherford obtained the curves of Fig. 16: the disintegration-functions of lithium and boron, with respect to incident protons, up to proton-energies of some 200,000 electron-volts. The wall between the target and the gas of the ionization-chamber had an air-equivalent of 2.50 cm., and consequently the curves pertain only to fragments having ranges greater than this.[24] The rise from the axis is gradual, not abrupt; one might say that the shape of the curves suggests that the protons have, not a definite *capability* for transmuting which begins suddenly at a critical energy, but a *probability* of trans-

[24] I hear from Dr. Oliphant that the trend of the curve for the short-range fragments is just the same.

muting which increases smoothly from zero (though this suggestion might not occur to anyone not having foreknowledge of the current theory!). The least energy at which transmutation is observable should then depend entirely on the strength of the proton-stream and the sensitiveness of the apparatus; von Traubenberg, with a stream

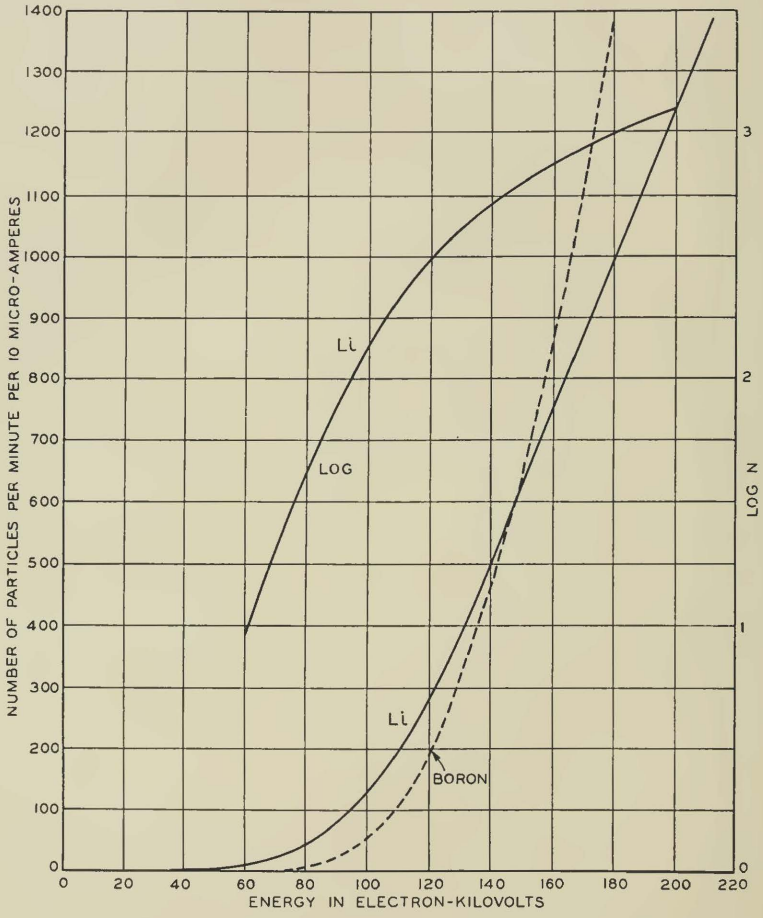


Fig. 16—Disintegration-functions of thin films of lithium and boron. (Oliphant & Rutherford)

perhaps as strong as that of Oliphant and Rutherford, observed one to three fragments per minute at 13,000 volts.

The curve of Fig. 17 extends very much further—all the way to 1.125 MEV—but was obtained with so thick a target of lithium (lithium fluoride, to be precise) that the protons came to a stop in the

mass, and the disintegrations observed at any voltage might have been produced by particles of any energy up to the maximum corresponding to the voltage. It comes from the Berkeley school, the data being procured chiefly by Henderson.[25] It refers only to fragments of ranges superior to 5.32 cm., a grave limitation, accepted in order to make sure that none of the primary protons could get into the detector (a Geiger counter). From 400,000 volts onward, the curve of Fig. 17 conforms

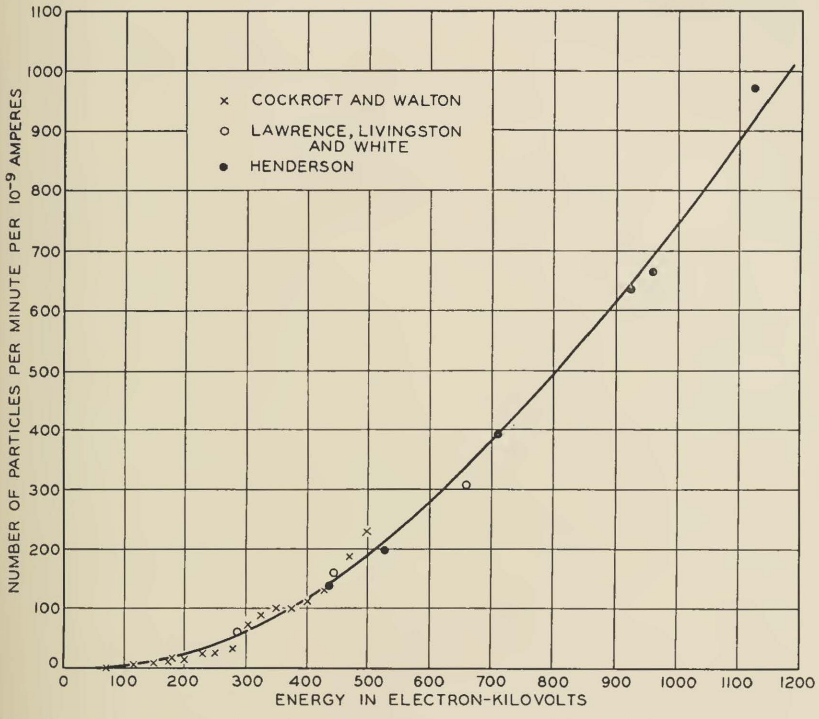


Fig. 17—Disintegration-function of lithium measured with a thick layer of lithium fluoride. (Henderson)

to a simple and somewhat surprising assumption: *viz.* the assumption that a proton of energy superior to 400,000 is neither more nor less efficient in disintegrating lithium than a proton of only 400,000 electron-volts, and that the whole of the rise in the curve from this voltage onwards is entirely due to the fact that the faster the proton, the farther it dives into the target and the more chances it has to

[25] The curve also fits the data of Cockcroft and Walton within the uncertainty of experiment, due regard being had to the difference in the values of the solid angle (letter from Dr. Henderson). In their work the screen between target and detector had an air-equivalent of 3 cm. (letter from Dr. Cockroft). The curve of Fig. 8 shows that this had the same effect as Henderson's 5.32 cm.

impinge on a nucleus before it is slowed down and its energy reduced beneath this particular value. The curve of Fig. 15 for lithium should then become horizontal at abscissa 400. At lower voltages, both curves concur in implying that the probability of disintegration depends on the energy of the proton. I will revert to this topic in a later article.

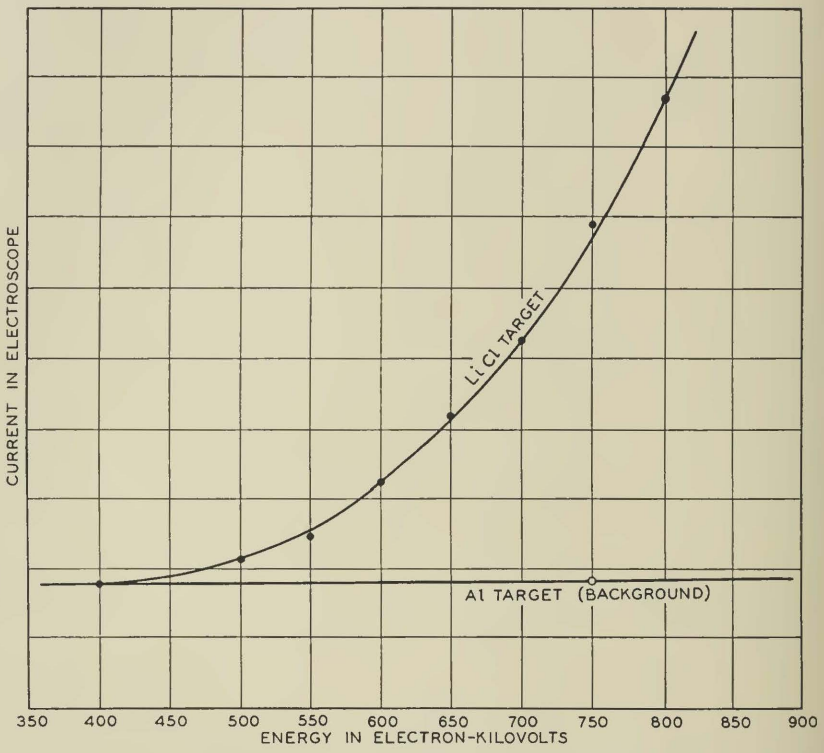


Fig. 18—Intensity of the mixture of neutrons and gamma-rays resulting when lithium is bombarded by deutons. (Crane & Lauritsen)

There are also modes of disintegration of lithium by deutons and by protons, in which neutrons and gamma-rays are emitted. These have been observed in Pasadena by Crane, Lauritsen and Soltan. Deutons are the more efficient of the two, but protons are sufficiently potent to have enabled Crane and Lauritsen to trace the curves of Fig. 18, in which the significant quantity is the difference between the ordinates of the two.[26] The ionization-chamber was walled inwardly

[26] Dr. Lauritsen writes me that the readings from which the lower curve is drawn were unchanged when the high voltage was removed; presumably therefore they represent the "background" due to the natural leaks of the electroscope. I am indebted to his letter for other as-yet-unpublished statements.

with paraffin, to accentuate the effect of the neutrons; it was however found that the readings were not considerably lessened when the paraffin coating was absent, and consequently Lauritsen infers that most of the effect is due to gamma-rays proceeding from the bombarded atoms. This inference is sustained by the fact that when the rays responsible for the effect are caused to pass through leaden screens, the ionization falls off exponentially with the thickness of the lead; and the value of the exponent suggests that the energy of the photons is about 1.5 MEV. One can easily think of a process whereby deutons might evoke neutrons from lithium nuclei:

$$ {}_1H^2 + {}_3Li^7 + T_0 = 2{}_2He^4 + {}_0n^1 + T_1, \qquad (7) $$

but with protons no plausible interaction comes readily to mind. Perhaps there is a two-stage process, the protons producing the reaction described by equation (1), the resultant He^4 nuclei striking other lithium nuclei and evoking neutrons. Or perhaps the neutrons and the gamma-rays alike result from the same processes as produce the groups of short-range alpha-particles revealed in Fig. 8. Questions of this intricate kind will probably predominate in the study of transmutation, in the years to come; and experiments on thin films will play a very important part in settling them, both because the likelihood of two-stage processes will be reduced, and because it may be possible to learn which isotopes are involved.

Little indeed is definitely known about the disintegration, by protons or deutons, of any other elements than lithium or boron. Charged fragments have been observed proceeding, in relatively small but yet appreciable number, from bombarded targets made of a great variety. But in many of these cases they may be due, so far as any of the observations tell, to a minute contamination of the target by boron derived from the glass of the enclosing tube; and the danger of this possible source of error was vividly brought out by Oliphant and Rutherford, when at first they observed such fragments, but ceased altogether to observe them when the original glass of their tube was replaced by a special boron-free variety! Beryllium and fluorine are the only elements, other than lithium and boron, of which these experimenters were sure of detecting fragments; for those of fluorine they were able to plot a disintegration-function and a distribution-in-range, which differed sufficiently in aspect from those of lithium and boron to exclude the possibility that these might be responsible; those of beryllium were too scanty for such tests. The elements with which they got no charged fragments, or only a few per minute, were the following: Fe, O, Na, Al, N, Au, Pb, Bi, Tl, U,

Th. But their observations were confined to protons of relatively low energy-values,—their upper limit was little over 200,000 electron-volts—and do not prove that faster particles are incapable of transmutation. The Berkeley school has already published a number of observations made with protons of energies ranging up to 710,000, and with deutons of energies attaining the unprecedented height of 3 MEV; and they find fragments in abundance from a wide diversity of targets.

Beryllium deserves a special paragraph, since it yields neutrons when bombarded, whether with alpha-particles from radioactive bodies; or with helium ions extracted from a discharge and endowed artificially with energies of 600,000 electron-volts and upward; or with deutons. The first of these processes is the one which led to the discovery of the neutron; the second, which incidentally marks the first employment of artificial alpha-particles (since these helium ions are alpha-particles in all but origin, except for the unimportant difference that each possesses an extra-nuclear electron while it is approaching the target) is a recent achievement of the Pasadena school (Crane, Lauritsen and Soltan); the third was achieved both at Pasadena and at Berkeley. These three processes are now in rivalry with one another, and it remains to be seen which will be producing the greatest number of neutrons, a year or five years hence. It is still very doubtful how the third takes place: perhaps the deuton merges with the beryllium nucleus, as in the other cases the alpha-particle is supposed to do (page 155), or perhaps it knocks a pre-existent neutron out of the beryllium structure and goes unaltered on its way. This too is a problem for the future, and one in the solving of which the charged fragments likewise observed will probably play a part.

The deuton itself is in all probability a complex particle; might it not be shattered in impinging against a nucleus, especially some heavy nucleus? This is the interpretation offered by Lawrence of the fact that in sending streams of deutons against targets of several different kinds, he observed charged fragments which were protons (not alpha-particles!) forming a group having a definite range and a definite energy not depending at all on the substance of the target. With 1.2-MEV deutons this characteristic energy of the protons is 3.6 MEV. A singular rule governs this quantity: if the energy of the bombarding particles is increased, that of the protons goes up by just the same amount—deutons of energy $(1.2 + x)$ MEV evoke protons of energy $(3.6 + x)$ MEV. The rule has been verified for values of x up to 1.8. Such a rule is just what one would expect, were there no other frag-

ments than the protons, excepting fragments of such great mass that they could take up the necessary momentum without taking an appreciable amount of kinetic energy. The heavy nucleus by itself is able to do this. However there are also neutrons, of which the energy is sufficient to let them be detected, and therefore by no means negligible. This is gratifying for the theory, inasmuch as if a proton is separated from a deuton, the residue should be a neutron (or else another proton and a free electron); but one is then obliged to assume that the neutron always takes the same kinetic energy, whatever that of the impinging deuton may have been. This seems rather odd, but nothing prohibits it. Streams of alpha-particles have been sent against compounds ("heavy water") containing deuterium in abundance, but as yet no neutrons have been detected coming off.

Transmutation by Impacts of Alpha-Particles [27]

Impact of an alpha-particle against a nucleus may result in the springing-off of one or more (or none) of four kinds of corpuscles: protons, photons, neutrons, positive electrons.

Transmutation with production of protons

This is the earliest-discovered type, of which I told at length in "Transmutation." The discovery was made by Rutherford in 1919 in experiments on nitrogen. At present the Cavendish school considers that this mode of transmutation has been proved for thirteen elements, none of atomic number greater than 19: the list comprises B, N, F, Ne, Na, Mg, Al, Si, P, S, Cl, A, K. The most frequently and fully studied cases are those of boron, nitrogen and aluminium.

The evidence that the fragments are protons is rather variegated. In some cases this has been proved by deflection-experiments; [28] recently it has been proved in some other cases by measuring both the range of the fragments and the ionization which they individually produce in a shallow chamber or a deep one (page 125); some observers are able to tell the scintillations due to protons from those which are due to alpha-particles.

Integral distribution-in-range curves of the fragments have been obtained for boron, nitrogen, fluorine, sodium, magnesium, aluminium and phosphorus. Most of them show more or less conspicuous plateaux, of which the most magnificent appear in the celebrated curves of Pose for aluminium, reproduced in "Transmutation"

[27] An expanded version of this section, with citations of additional data and reproductions of some curves, appears in the Physics Forum of the *Review of Scientific Instruments* for February 1934.

[28] "Transmutation," pp. 636–640, *B. S. T. J.*, Oct. 1931.

(Figs. 6, 7); from this there are all gradations of distinctness downward, ending with cases in which it is uncertain whether the ideal curve would be a smoothly-descending one, or would have a succession of short plateaux which in the actual curve are rounded off into indistinguishability.

By "ideal curve" in the foregoing sentence I mean, as heretofore (page 130), that which would be obtained with an infinitely narrow beam of fragments proceeding in a single direction and produced by alpha-particles all of a single speed and proceeding in a single direction. I must also add that many thousands of fragments should be counted, as otherwise the results are likely to be distorted by statistical fluctuations. It appears that in most of the experiments with bombarding alpha-particles, the departure from the ideal is much more considerable than in the best of the experiments with bombarding protons. The targets are usually so thick that the speeds of the alpha-particles vary considerably as they go through, and often so thick that these are swallowed up and every energy of bombarding particle, from the initial maximum down to zero, is represented among the impacts. This matters much more than it does with protons, because here the energy of the primary particles is often much greater than that of the fragments, and a small percentage variation of the former may entail a big one of the latter. The solid angles subtended by the exposed part of the target as seen from the source of the alpha-rays on the one hand, from the detector on the other, are frequently both large. This is particularly serious, because it appears that the ideal distribution-in-range curve would vary with the angle between the directions of the impinging particle and of the fragment. In some experiments the number of fragments observed has been too small to be immune to statistical fluctuations, and it is surprising that the plateaux in Pose's curves should be so clear despite this handicap.

Where two or more observers have studied a single element, there is generally enough concordance among their statements to assure the onlooker that at least the major groups of protons are recognizable. The prettiest case thus far is that of nitrogen: three researches on the integral distribution-in-range curve agree in showing a sharply-marked group of range about 17.5 cm (for protons ejected forward by full-speed alpha-particles from polonium, energy 5.3 MEV). The flattest plateau and sharpest step are to be seen in a curve by Chadwick Constable & Pollard, who approached very nearly to the ideal experiment in one respect, by using a stratum of nitrogen so thin that its air-equivalent was only 3 mm. All the protons of range superior to about 6 cm. belong to this group; there is another of inferior range, lately discovered

by Pollard. Phosphorus and sodium have been studied only by Chadwick Constable & Pollard, who find for the former a single group, for the latter a smoothly-descending integral curve which may betoken total absence of groups, or may be resolved, by some future and closer approach to the ideal curve, into a close succession of bends and corners. The four remaining elements—B, F, Mg, Al—show at least three groups apiece, and indeed Chadwick and Constable deduce four *pairs* of groups for aluminium and three for fluorine. To illustrate the degree of concurrence between different observers, I quote the values for the groups of aluminium—that is to say, values of the ranges of the protons belonging to these groups, ejected forward by 5.3-MEV alpha-particles —from the four authorities. Pose gives 28.5, 49.6, and 61.2 (cm of air-equivalent); Steudel, 33, 49, 63; M. de Broglie and Leprince-Ringuet, 30, 50, 60; Chadwick and Constable give 22, 26.5, 30.5, 34, 49, 55, 61, 66. More detailed comparisons had best be left to those who have practice in this field.

While nearly all of the data have been obtained by other methods than that of the expansion-chamber, a few beautiful pictures have been taken in which there appears the track of an alpha-particle passing through nitrogen, and this track is seen to end at a fork.[29] One of the tines of the fork is a long thin track, apparently that of a proton; there is only one other, and this is short and thick. It is inferred that these reveal the only fragments which there are, and that, in the usual though somewhat objectionable phrase, the alpha-particle has fused with the residual nucleus. The process is then expressed by the equation:

$$ {}_7\mathrm{N}^{14} + {}_2\mathrm{He}^4 + T_0 = {}_8\mathrm{O}^{17} + {}_1\mathrm{H}^1 + T_1, \tag{8} $$

the symbols being chosen according to the same principles as in equation (1). It is commonly assumed, though in no other case with such good evidence, that this happens in most if not in all cases, so that when a nucleus of atomic number Z and mass-number A is transmuted by an alpha-particle, the process often is:

$$ {}_Z\mathrm{M}^A + {}_2\mathrm{He}^4 + T_0 = {}_{Z+1}\mathrm{M}^{A+3} + {}_1\mathrm{H}^1 + T_1, \tag{9} $$

with an obvious symbolism. This is called "disintegration with capture" (though it is the case in which the objection to the name "disintegration," page 117, is gravest). The other conceivable case of "disintegration without capture" would be described thus:

$$ {}_Z\mathrm{M}^A + {}_2\mathrm{He}^4 + T_0 = {}_{Z-1}\mathrm{M}^{A-1} + {}_1\mathrm{H}^1 + {}_2\mathrm{He}^4 + T_1. \tag{10} $$

[29] "Transmutation," Figs. 10 and 11.

Disintegration-with-capture is very advantageous for the theorist, since when there are only two fragments after the interaction the principle of conservation of momentum suffices to determine the kinetic energy of either in terms of that of the other and that of the alpha-particle. In equation (9), T_0 stands for the kinetic energy of the alpha-particle, T_1 for the sum of the kinetic energies of the proton and the residual fragment, which call T_p and T_r respectively. Now excepting in the cloud-chamber experiments, it is only the proton which is detected, and therefore only T_p can be estimated from the data; but if the disintegration is by capture, then T_r and consequently T_1 can be deduced from T_0 and T_p. If however there are three or more final fragments, measurement of T_p is not sufficient to determine T_1. Also even in the case of disintegration-by-capture there will be uncertainty if the transmuted element is a mixture of two or more isotopes, since the value of T_r corresponding to an observed T_p will depend on the mass of the atom which is transmuted.

In a case of disintegration-by-capture, the simplest possible assumption is that $(T_1 - T_0)$ has a perfectly definite value, independent of T_0: there is conversion of a definite amount of kinetic energy into rest-mass (or vice versa), whatever the velocity of the alpha-particle may be. This may be tested by varying T_0; it may also be tested to some extent by observing protons ejected in various directions (relatively to the initial direction of the alpha-particles) since although the sum of T_p and T_r (which is T_1) should be the same for all of these protons those two quantities individually should vary, and T_p in particular should depend in a definite manner on the direction of the protons. Yet in nearly all such tests, the target is so thick that the alpha-particles impinging on various nuclei have very various speeds. How then shall we know which speed of proton to associate with which speed of alpha-particle, which value of T_p belongs with which of T_0? One naturally begins by assuming that the fastest of the primary particles produce the fastest of the protons. But plausible as this assumption seems at first, there are several cases known in which it is not true: cases in which a definite group of protons is evoked by alpha-particles of a definite interval of speeds, and neither faster nor slower particles are capable of producing them.

This phenomenon of "resonance," as it is called,[30] was first observed by Pose in the experiments on aluminium to which many pages were devoted in "Transmutation." It is evidently an important quality of nuclei, destined to be prominent in experiment and theory both.

[30] There is a tendency to use the term "resonance" to express the mere existence of groups, irrespective of whether they are evoked by alpha-particles of narrowly limited speeds. This is to be deprecated.

This makes it desirable to consider at some length how resonance may be detected. There are the following ways:

(*a*) When the target is thick, one may vary the energy K_0 which the particles possess when they strike the target-face K_0 (usually by varying the density of gas between the target-face and the source of the alpha-particles) and plot the integral distribution-in-range curve for many different values of K_0. Let us suppose that there is a certain proton-group evoked only by alpha-particles having energy between K_a and K_b, the notation being so chosen that $K_b < K_a < K_0$. Then it will be found that as K_0 is lowered, the step and plateau which reveal the group will remain unaltered until K_0 drops below a certain critical value (to be identified with K_a) after which they will fade out.

(*b*) In the foregoing conditions, one may use a very thin ionization-chamber and plot instead of the integral distribution-in-range curve a curve of the sort in Fig. 10, or the sort described on page 129 of which the ordinate stands for the number of fragments producing more than a certain chosen amount of ionization in the chamber. There will be various peaks in the curve corresponding to various groups, and if any of these is produced by "resonance" it will at first remain unaltered and then gradually disappear as K_0 is lowered.

(*c*) When targets thin enough to be completely traversed by the alpha-particles are available, one may leave K_0 unchanged and increase the thickness t of the target. The energies of the impinging particles in a target then vary from K_0 down to a minimum value K_1 which depends on t. If curves of any of the foregoing kinds be plotted for various values of K_1, and if any of the groups is produced by resonance, then the step or the peak corresponding to this group may be absent when K_1 is high (i.e. with the thinnest target) and will then make its appearance when K_1 is lowered past a certain critical value (again to be identified with K_a).

(*d*) If the target is so very thin that the loss of speed suffered by the alpha-particles in going through is negligible, and K_1 is sensibly equal to K_0, then when K_0 is varied the groups should appear and disappear when it becomes equal to K_a and K_b, respectively.

(*e*) Without subjecting the fragments to any analysis, one may simply measure the total number thereof (or rather, the total number having ranges superior to some fixed minimum) as function of K_0. Suppose the target to be thick; then, if all the proton-groups are evoked by resonance, the curve should display a sequence of steps and plateaux; if in addition to such there are groups which are evoked by particles of any energy over a wide interval, the steps need not vanish, but the plateaux should slope upward and may be curved.

If the target is very thin (in the sense of the previous paragraph) the curve ought to show a peak for each group. Such curves, by the usage of page 139, may be styled "disintegration functions" (the term "excitation-function" is also used).

(*f*) Finally, when the target is thick the mere existence of sharp steps in the integral distribution-in-range curves, may be taken as a suggestion of resonance, since if a group were evoked by alpha-particles of a wide range of energies it would probably have a broad distribution of speeds. But this is not a very strong argument by itself.

Despite this great variety of ways of testing for resonance, the situation is still confusing and confused.

Aluminium has been the object of most of the tests, doubtless because it figured in Pose's discovery. He used methods (*a*) and (*c*) and found resonance distinctly and even vividly displayed by the 60-cm. and the 50-cm. group, and not at all by the 25-cm. group. Chadwick and Constable used (*a*) and (*b*), and concluded that there is resonance for six at least of their eight groups, the two members of a pair appearing and disappearing together. (The remaining pair was elicited by alpha-particles of a limited interval of energy-values extending from a lower limit K_b to the highest value of K_0 which they had available.) They also used (*e*) with a very thin sheet of aluminium (air-equivalent 0.8 mm.) and got a curve with two well-defined peaks. But Steudel also had recourse to method (*e*), and the curve he got swept smoothly upward; it is true that his target was notably thicker (air-equivalent 5.2 mm.) and yet one would not expect such a thickness to blot out the peaks if they exist. Harder yet to explain away is the evidence of M. de Broglie and Leprince-Ringuet, who made test (*d*) with sheets of aluminium of air-equivalent 2.5 mm., and observed all three of Pose's groups over a wide range of values of K_0.—As for the other elements: boron and fluorine and magnesium have all been tested by method (*a*), and there are strong indications of resonance for all three, strongest for fluorine. Nitrogen has been studied by Pollard with a modification of (*e*), and he finds that resonance is displayed by the 6-cm. group but not by the stronger and better-known group of longer range.

Evidently this is a field which yearns for further cultivation, with more powerful sources of transmuting particles to make possible the use of narrower and more homogeneous beams of these, narrower pencils of fragments and thinner strata of matter. The discovery of the capacity of protons to transmute has probably diverted from it some of the attention which otherwise it would by now have received, but the lost ground will doubtless be made up in the course of years, after the

developments which that discovery has hastened shall have brought about the generation of streams of artificial alpha-particles more numerous by far than the natural ones. Meanwhile we must be content with scanty data and with fragmentary tests of the important question already mentioned: whether the energy transformed from rest-mass to vis viva or reversely—the quantity here denoted by $(T_1 - T_0)$, elsewhere commonly by Q, designated in German as the *Tönung* of the process—is a definite and characteristic quantity.

Certainly about resonance is essential to these tests; for if resonance exists, we have to correlate the energy of a group of protons with that particular energy of the alpha-particles which evokes the group; but if resonance does not occur, then probably the best we can do is to correlate the energy of the fastest of the ejected protons of a group with that of the fastest of the impinging particles—and if we make the latter guess when it ought not to be made, there will be trouble! Perhaps the most impressive evidence is that available for aluminium. Chadwick and Constable evaluated $(T_1 - T_0)$ for all of their eight groups: the six for which they demonstrated resonance, and the two which were evoked by alpha-particles of a limited interval of energies extending up to the highest which they used, which was 5.3 MEV. They find that $(T_1 - T_0)$ has a common value of $+2.3$ MEV for four of their groups—to wit, the longer-range members of their four pairs—and a common value of zero for the other four. Haxel plotted the integral distribution-in-range curves for the protons ejected by alpha-particles of several yet higher energies, running up almost to 9 MEV; he detected two groups; they did not display resonance, but he correlated the highest energy represented in each with the highest represented among the impinging particles, and he too found $+2.3$ MEV and zero for $(T_1 - T_0)$ in the two cases![31] Blackett analyzed eight examples of transmutation of nitrogen observed with the cloud-chamber (here he had the unique advantage of being able to observe the track of the residual nucleus and estimate its energy) and he reported for $(T_1 - T_0)$ a mean value of -1.27 MEV with a mean deviation of 0.42 from the mean. Future confirmation awaited this work also: Pollard, analyzing his integral distribution-in-range curves, made a computation of $(T_1 - T_0)$ for the 6-cm. group which exhibits resonance, and another for the 17.5-cm. group which does not, correlating in this latter case the energy of the fastest protons with that of the fastest alpha-particles; the results were -1.32 and -1.26 MEV.

[31] The precision of these values can hardly be estimated from what Chadwick and Constable say, but some idea of it can be gained from a graph in Haxel's article, *ZS. f. Phys.* **83**, p. 335 (1933), and *loc. cit.* footnote 27.

Such are the cases where there is the strongest proof for the twin doctrines that disintegration is by capture, and that a definite amount of energy is transformed between rest-mass and vis viva. The reader will have noticed in the latter case, that $(T_1 - T_0)$ appeared to be the same for a group which exhibits resonance and for another group which does not. This if certain may be taken to mean, that a particular group of protons—one may speak more graphically, and say: a particular proton in a particular level of the nitrogen nucleus—can be extracted by alpha-particles of a narrowly-limited range of energies between critical energy-values K_a and K_b, and can also be extracted by alpha-particles of *any* energy superior to a third critical value K_c which is greater than K_a and K_b. There is a good interpretation of this notion in the contemporary theory, which I reserve for the next article. It will also have been noticed that two different values of $(T_1 - T_0)$ were given for a single case, that of aluminium (there are also two for fluorine). This is to be taken as meaning that the residual nucleus may be left in either of two conditions, one of which may be the normal state, while the other must be an excited state (page 138). One then infers that the nucleus when left in the excited state will presently go over to the normal state, emitting a photon having an amount of energy equal to the difference between the two values of $(T_1 - T_0)$. It is very tempting to suppose that the gamma-rays known to be emitted from some elements during alpha-particle bombardments have this origin, but the measurements are not yet precise enough to prove this.[32]

In a case of disintegration-by-capture, the residual nucleus denoted by ${}_{z+1}\mathrm{M}^{A+3}$ in equation (9) might or might not be exactly the same as the nucleus of the known chemical atom (if such there be) of atomic number $(Z + 1)$ and mass-number $(A + 3)$. Can this be tested by comparing the rest-mass of the former with the mass of the latter as measured by Aston or Bainbridge? Unfortunately nothing of value can be concluded unless the atoms ${}_{z+1}\mathrm{M}^{A+3}$ and ${}_{z}\mathrm{M}^{A}$ have both had their masses determined with an accuracy permitting them to appear in the Table on page 109; and on inspecting this table one finds (with some surprise) that this is true for only one of the known processes, *viz.* the transmutation of fluorine. Assuming disintegration to be with capture, the process would be the following:

$$ {}_{9}\mathrm{F}^{19} + {}_{2}\mathrm{He}^{4} = {}_{10}\mathrm{Ne}^{22} + {}_{1}\mathrm{H}^{1} + (T_1 - T_0) \qquad (11) $$

Putting for $(T_1 - T_0)$ the value $+1.67$ MEV given by Chadwick and Constable, and for the rest-masses of the nuclei the values given in the

[32] Heidenreich has analyzed the data for boron, and concludes that they permit of this interpretation. (*ZS. f. Phys.* **87**, 675–693; 1933.)

Table, we get 23.002 for the left-hand member and 23.0043 for the right-hand member. The agreement is within the uncertainty of the data; so also would it have been, had $(T_1 - T_0)$ been ignored. Its importance is perhaps enhanced by the fact that it is *ex post facto:* the mass of Ne^{22} was inaccurately known at the time of the experiments of Chadwick and Constable, and there was ostensibly a disagreement.

I repeat that it is not proved that transmutation occurs in every case by capture; and an isolated value of $(T_1 - T_0)$, such as one often sees computed from a single observation on a particular group evoked by a particular beam of alpha-particles, is not necessarily valid.

Transmutation with production of neutrons

This mode of transmutation has been proved, according to the Cavendish school and the Joliots, for the elements Li, Be, B, F, Ne, Na, Mg, and Al. The outstanding cases are those of beryllium and boron, with lithium and fluorine following after. Negative results have been reported by the Joliots for H, C, O, N, P and Ca, and there is no record of a positive result for He. Positive results have been reported for quite a number of elements both light and heavy by the Vienna school.

There is nothing which can properly be called a distribution-in-range curve for neutrons; but there is something which is potentially as useful—the integral distribution-in-range curve of the protons emanating from a thin layer of matter rich in hydrogen, placed between the source of the neutrons and the detector. If one can measure the speed of a proton recoiling in a known direction from the impact of a neutron, one can deduce the speed of the neutron; in particular, if one can measure the speeds of the protons projected straight forward by central impacts of the oncoming neutrons, one may consider their speeds as practically the same as those of the neutrons themselves.[33] It is thus a proper procedure to obtain the integral distribution-in-range curve of the protons projected forward, and convert it into a distribution-in-energy curve which is that of the protons and the neutrons alike. It has however not been an easy procedure, on account of the sparseness of the available sources of neutrons and hence of the streams of recoiling protons. Chadwick has published a solitary curve of this sort, relating to the neutrons from beryllium ejected by the alpha-particles of polonium; and Dunning has obtained a curve displaying good plateaux and steps, relating to the neutrons from beryllium ejected by yet faster alpha-particles.[34] Steps and plateaux, as heretofore, signify groups of protons and consequently groups of neutrons. Feather has

[33] Cf. Part I, page 300.

[34] To be published in the article mentioned in Footnote 27, and by Dr. Dunning himself.

achieved the feat of taking and examining no fewer than 6900 cloud-chamber photographs in order to deduce the distribution-in-speed of neutron-streams from the tracks of the recoiling nuclei of various kinds of atoms. Most observers publish no curves, but give only verbal accounts in which they state the thickness (in air-equivalent) of the intercepting screens athwart the proton-beam, for which they observed a notable falling-off of the strength of that beam; or else they state what groups they believe in, inferring them presumably from observations of that type. This makes tiresome and unsatisfactory reading.

Much of recent research is meant to detect the very fastest neutrons emitted from a given element, for a reason which will presently be obvious if it is not already. Chadwick gives 3.35 MEV for the energy of the fastest neutrons ejected from boron by polonium alpha-particles, and 12 MEV for those similarly ejected by beryllium, while Dunning gives 14.3 MEV for those which beryllium emits when bombarded by the somewhat faster alpha-particles from radon.

Curves called "disintegration-functions," or more commonly "excitation-functions," have been plotted several times for the neutrons from beryllium and once at least for those from boron. One must realize an important distinction between them and the curves obtained when the fragments are alpha-particles or protons, as in Figs. 16 and 17. When the fragments are charged particles, it is practically certain that *all* of them which reach the detector at all are duly detected. When the fragments are neutrons it is certain that the only ones detected are those which strike protons (or other nuclei) hard enough and squarely enough to give them a considerable amount of energy and enable them to produce a good many ions in the ionization-chamber; and it is equally certain that those constitute but a small fraction of the total number of neutrons, most of which go through the expansion-chamber unperceived. Would that this were at least a *constant* fraction! we could then rely on the shape of the so-called excitation-curve, while realizing that all its ordinates must be multiplied by some unknown but constant factor. But we must not suppose even this; it is practically certain that the factor varies with the speed of the neutrons, and hence in all probability with the speed of the primary alpha-particles; and hence the so-called excitation-curve must be distorted from the true curve of number-of-atoms transmuted *versus* energy-of-alpha-particles. (Also the distribution-in-range curves must be distorted.)

With these severe limitations in mind, one may consider the published excitation-functions. The most striking are those obtained with very thin films of beryllium, one by Chadwick and one by Bernardini,

which agree in showing a rather sudden rise of the curve from the horizontal axis, then a peak, then a valley and then a sweeping rise. It is hardly likely that the peak and the valley are entirely due to distortion of a truly smoothly-rising curve by the aforesaid agency; and the argument of paragraph (*e*) of page 149 leads us to infer a group of neutrons displaying resonance, in addition to other neutrons for which perhaps there is no resonance. Curves obtained with thick targets of beryllium or of boron have conspicuous steps, carrying the same implication. Those for boron (Chadwick and the Joliots) and some of those for beryllium (Rasetti, Bernardini) suggest but a single group, but there are other curves for beryllium suggesting two (in recent work of Chadwick's) and even four (Kirsch and Slonek). Thus, although the first four tests of resonance which I listed above (page 149) have as yet remained untried for emission of neutrons, the fifth has given some pretty convincing evidence in its favor.

It is always assumed that transmutation with emission of a neutron is a case of disintegration-by-capture, though no one has proof of this yet. The imagined process may be symbolized thus:

$$ {}_{Z}\mathrm{M}^{A} + {}_{2}\mathrm{He}^{4} + T_0 = {}_{Z+2}\mathrm{N}^{A+3} + {}_{0}n^{1} + T_1 \qquad (12) $$

Such equations as this are used for evaluating the rest-mass of the neutron, it being assumed that the rest-mass of the residual nucleus ${}_{Z+2}\mathrm{N}^{A+3}$ is identical with that of the nucleus of the atom of mass-number $(A + 3)$ and atomic number $(Z + 2)$. One encounters at once the difficulty that there are neutrons of a wide range of speeds, and consequently a wide range of values of T_1. It is necessary to assume that the slower neutrons leave behind them a nucleus in an excited state (page 138) and that only the very fastest leave behind them the normal nucleus which is to be identified with that of the isotope $(A + 3)$ of the element $(Z + 2)$. Doing this, Chadwick got consistent values for the mass of the neutron from the observations on boron and on lithium, assuming the nucleus M of equation (12) to be that of B^{11} and that of Li^{7} respectively.[35] To obtain a consistent value from the neutrons of beryllium, one would have to observe some at least having an energy as great as 12 MEV (when $T_0 = 5.3$ MEV). Those observed in the earlier work on beryllium were all much too slow. One of the driving motives of recent research has been the desire of finding at least a few of adequate energy; and it appears that this desire has at last been fulfilled.

[35] Were we to assume B^{10} and Li^{6}, the nucleus N would correspond to an isotope as yet unknown; this is a powerful but not an absolutely imperative argument against these choices. There is also the question of whether, if resonance occurs, the right correlation is being made between values of T_1 and values of T_0 (page 151).—The equation for the transmutation of boron has been worked out in Part I., pp. 323–324.

To guess at the total number of neutrons emitted (say) from beryllium it is necessary to know the excitation-curve and to make an estimate of the factor aforesaid. I confine myself to quoting from Chadwick: "The greatest effect is given by beryllium, where the yield is probably about 30 neutrons for every million alpha-particles of polonium which fall on a thick layer."

Transmutation with production of positive electrons

This mode of transmutation, as I mentioned earlier, has been observed by the Joliots with Be, B and Al, the primary corpuscles being polonium alpha-particles. Nothing has yet been published about distribution-in-range or disintegration-function. Positive electrons of energy as high as 3.1 MEV have been observed proceeding from aluminium.

Aluminium thus affords a case of an atom which under alpha-particle bombardment may emit from its nucleus a particle of any of three kinds: a proton, a neutron, a positive electron. It has been suggested by Joliot that there is actually only one process, in which a proton emerges either intact, or else split into a neutron and a positive electron which are its hypothetical components. If this can be verified it will have important bearings on various fundamental questions, including that of the mass of the neutron.[36] Boron also emits particles of all three kinds, but here the situation is complicated by the possibility that not all of the three proceed from the same isotope.

Transmutation by Neutrons

Transmutation by neutrons has been observed only with the Wilson chamber, and therefore rarely: there are a few scores of recorded cases, the fruit of twenty or thirty thousand separate photographs taken some by Feather at the Cavendish, some by Harkins and his colleagues at Chicago. What is observed is a pair of tracks diverging from a point in the midst of the gas contained in the chamber; it is inferred that the (invisible) path of a neutron extends from the neutron-source to the point of the divergence, and that the observed tracks are those of two fragments of a nucleus which that particle has struck. "Fragment" must be taken in the generalized sense of page 117: the substance of the neutron may be comprised in either or both of the two. Each case must be separately analyzed, taking into account the directions and the ranges of the fragments (it is here that the question of the range-*vs*-energy relations of massive nuclei, footnote 20, becomes crucial). It is possible to infer that in many cases the neutron is ab-

[36] See the reference in Footnote 27.

sorbed into the fragments—"disintegration with capture"—and even to estimate $(T_1 - T_0)$, which turns out to be usually if not always negative. There are some difficulties here, since in certain cases the process which is observed seems to be the converse of one of the well-known processes of generating neutrons, and yet $(T_1 - T_0)$ does not appear to have values equal in magnitude and opposite in sign for the two. The most startling feature of transmutation by neutrons is, that it occurs with nuclei which seem to be immune to other transmuting agents, notably carbon and oxygen. Other elements with which it occurs are nitrogen, fluorine, neon, chlorine and argon.

Acknowledgments

I am greatly indebted to Monsieur F. Joliot, Professor E. O. Lawrence, Dr. J. R. Dunning and Dr. P. I. Dee for providing me with prints of several of the photographs which appear in this article (Figs. 1, 4, 5, 6, 14, 15); and to Dr. Dunning for criticism and advice in respect to several sections of the text.

References

Transmutation by Protons and Deutons

Cavendish school:

J. Cockcroft & E. T. S. Walton: *Proc. Roy. Soc.* **A129**, 477–489 (1930); **136**, 619–630 (1932); **137**, 229–242 (1932).
P. I. Dee: *Nature* **132**, 818–819 (25 Nov. 1933).
P. I. Dee & E. T. S. Walton: *Proc. Roy. Soc.* **A141**, 733–742 (1933).
M. L. E. Oliphant & E. Rutherford: *Proc. Roy. Soc.* **A141**, 259–281 (1933). The same with R. B. Kinsey: *ibid.* 722–733.

Berkeley school:

M. C. Henderson: *Phys. Rev.* (2) **43**, 98–102 (1933).
E. O. Lawrence & M. S. Livingston: *Phys. Rev.* (2) **40**, 19–35 (1932).
Letters and abstracts by E. O. Lawrence, M. S. Livingston, M. G. White, G. N. Lewis, M. C. Henderson: *Phys. Rev.* (2) **42**, 150–151, 441–442 (1932); **43**, 212, 304–305, 369 (1933); **44**, 55–56, 56, 316–317, 317, 781–782, 782–783 (1933).

Other schools:

H. R. Crane, C. C. Lauritsen & A. Soltan, *Phys. Rev.* (2) **44**, 514 (1933) (effect of He+ ions); *ibid.* 692–693; Crane & Lauritsen, *ibid.* 783–784; **45**, 63–64 (1934).
C. Gerthsen: *Naturwiss.* **20**, 743–744 (1932).
F. Kirchner: *Phys. ZS.* **33**, 777 (1932); **34**, 777–786 (1933); with H. Neuert, **34**, 897–898 (1933). *Sitzungsber. d. kgl. Bäyrischen Akad.* 129–134 (1933). *Naturwiss.* **21**, 473–478, 676 (1933).
H. Rausch v. Traubenberg, R. Gebauer, A. Eckart: *Naturwiss.* **21**, 26 (1933); *ibid.* 694.

Transmutation by Alpha-Particles

Transmutation with emission of protons:

P. M. S. Blackett: *Proc. Roy. Soc.* **A107**, 349–360 (1925).
W. Bothe: *ZS. f. Phys.* **63**, 381–395 (1930); *Atti del convegno di fisica nucleare*, Roma, 1932.
W. Bothe & H. Fränz: *ZS. f. Phys.* **43**, 456–465 (1927); **49**, 1–26 (1928).

W. Bothe & H. Klarmann: *Naturwiss.* **35**, 639–640 (1933).
M. de Broglie & L. Leprince-Ringuet: *C. R.* **193**, 132–133 (1931).
J. Chadwick, J. E. R. Constable & E. C. Pollard: *Proc. Roy. Soc.* **A130**, 463–489 (1931).
J. Chadwick & J. E. R. Constable: *Proc. Roy. Soc.* **A135**, 48–68 (1932).
K. Diebner & H. Pose: *ZS. f. Phys.* **75**, 753–762 (1932).
W. D. Harkins: with R. W. Ryan, *J. Am. Chem. Soc.* **45**, 2095–2107 (1923); with H. A. Shadduck, *Proc. Nat. Acad. Sci.* **2**, 707–714 (1926); with A. E. Schuh, *Phys. Rev.* (2) **35**, 809–813 (1930).
O. Haxel: *ZS. f. Phys.* **83**, 323–337 (1933).
F. Heidenreich: *ZS. f. Phys.* **86**, 675–693 (1933).
G. Hoffmann: *ZS. f. Phys.* **73**, 578–579 (1932).
C. Pawlowski: *C. R.* **191**, 658–660 (1930).
E. C. Pollard: *Proc. Roy. Soc.* **A141**, 375–385 (1933).
H. Pose: *Phys. ZS.* **30**, 780–782 (1929); **31**, 943–945 (1930). *ZS. f. Phys.* **60**, 156–167 (1930); **64**, 1–21 (1930); **67**, 194–206 (1931); **72**, 528–541 (1931). With F. Heidenreich: *Naturwiss.* **21**, 516–517 (1933).
E. Steudel: *ZS. f. Phys.* **77**, 139–156 (1932).
Additional early references given at the end of *Transmutation.*

Transmutation with emission of neutrons:

G. Bernardini: *ZS. f. Phys.* **85**, 555–558 (1933).
J. Chadwick: *Proc. Roy. Soc.* **A142**, 1–25 (1933).
N. Feather: *Proc. Roy. Soc.* **A142**, 689–714 (1933).
F. Joliot & I. Curie: *J. de Phys.* (7) **4**, 278–286 (1933).
G. Kirsch & W. Slonek: *Naturwiss.* **21**, 62 (1933).
F. Rasetti: *ZS. f. Phys.* **78**, 165–168 (1932).

Transmutation with emission of positive electrons:

I. Curie & F. Joliot: *J. de Phys.* (7) **4**, 494-500 (1933).

Transmutation by Neutrons

N. Feather: *Proc. Roy. Soc.* **A136**, 703–727 (1932); **142**, 689–709 (1933).
W. D. Harkins, D. M. Gans & H. W. Newson: *Phys. Rev.* (2) **44**, 529–537 (1933).
Letters and abstracts by W. D. Harkins, D. M. Gans, H. W. Newson: *Phys. Rev.* (2) **43**, 208, 362, 584, 1055 (1933); **44**, 236, 310, 945 (1933).
F. N. D. Kurie: *Phys. Rev.* (2) **43**, 771 (1933).

Abstracts of Technical Articles from Bell System Sources

Attenuation of Overland Radio Transmission in the Frequency Range 1.5 to 3.5 Megacycles per Second.[1] C. N. Anderson. Data on the effect of land upon radio transmission have been obtained during the past few years in connection with various site surveys. These data are for the general frequency range 1.5 to 3.5 megacycles per second and for various combinations of overwater and overland transmission as well as entirely overland. The generalizations in this paper are chiefly in the form of curves which enable one to make approximations of field strengths to be expected under the conditions noted above. The relation of these data to transmission in the broadcast frequency range is shown, and from the over-all picture, curves are developed which enable field strength estimates to be made for overland transmission in the extended frequency range.

The Radio Patrol System of the City of New York.[2] F. W. Cunningham and T. W. Rochester. The application of radiotelephony to municipal police work in New York City is described from the organization viewpoint. Brief references are made to historical backgrounds and description of apparatus, and the steps taken to select a receiver suitable for local conditions are outlined. The method of controlling the patrol force by radio is described at some length with examples, and a summary of results during the first year is given to show the value of this means of communication to police work.

Electrical Disturbances Apparently of Extraterrestrial Origin.[3] Karl G. Jansky. Electromagnetic waves of an unknown origin were detected during a series of experiments on atmospherics at high frequencies. Directional records have been taken of these waves for a period of over a year. The data obtained from these records show that the horizontal component of the direction of arrival changes approximately 360 degrees in about 24 hours in a manner that is accounted for by the daily rotation of the earth. Furthermore the time at which these waves are a maximum and the direction from which they come at that time changes gradually throughout the year in a way that is accounted for by the rotation of the earth about the

[1] *Proc. I. R. E.*, October, 1933.
[2] *Proc. I. R. E.*, September, 1933.
[3] *Proc. I. R. E.*, October, 1933.

sun. These facts lead to the conclusion that the direction of arrival of these waves is fixed in space; i.e., that the waves come from some source outside the solar system. Although the right ascension of this source can be determined from the data with considerable accuracy, the error not being greater than ± 7.5 degrees, the limitations of the apparatus and the errors that might be caused by the ionized layers of the earth's atmosphere and by attenuation of the waves in passing over the surface of the earth are such that the declination of the source can be determined only approximately. Thus the value obtained might be in error by as much as ± 30 degrees.

The data give for the coordinates of the region from which the waves seem to come a right ascension of 18 hours and a declination of − 10 degrees.

A Precision, High Power Metallographic Apparatus.[4] Francis F. Lucas. In 1927 the design of an advanced type of metallographic apparatus became of interest. Preliminary designs were prepared and discussed at a conference in Jena, Germany, with the scientific staff of Carl Zeiss. The Zeiss works was commissioned to construct the apparatus. The work was directed by Professor A. Kohler, an outstanding authority on the optics of the microscope, head of the mikro-department of the Zeiss works, and Professor Walter Bauersfeld, a director of the Zeiss Foundation and inventor of the Planetarium.

In this paper the author discusses the considerations which led to the design and describes the construction of the apparatus. It is the largest and the most powerful metallurgical microscope ever constructed. Capable of yielding crisp, brilliant images at magnifications of 4000 to 6000 diameters, the design required great mechanical stability, freedom from creep, absolute freedom from outside disturbances, the means to illuminate the specimen with light of any selected wave-length or group of wave-lengths within the visible spectrum and the highest order of achievement in optical equipment.

[4] Published in abridged form in *Metal Progress*, October, 1933.

Contributors to this Issue

H. S. Black, B.S. in Electrical Engineering, Worcester Polytechnic Institute, 1921. Western Electric Company, Engineering Department, 1921–25; Bell Telephone Laboratories, 1925–. Mr. Black's work has had to do with the development of carrier telephone systems.

Arthur G. Chapman, E.E., University of Minnesota, 1911. General Electric Company, 1911–13. American Telephone and Telegraph Company, Engineering Department, 1913–19, and Department of Development and Research, 1919–. Mr. Chapman is in charge of a group engaged in developing methods for reducing crosstalk between communication circuits, both open wire and cable, and evaluating effects of crosstalk on telephone and other services.

Karl K. Darrow, B.S., University of Chicago, 1911; University of Paris, 1911–12; University of Berlin, 1912; Ph.D., University of Chicago, 1917. Western Electric Company, 1917–25; Bell Telephone Laboratories, 1925–. Dr. Darrow has been engaged largely in writing on various fields of physics and the allied sciences.

Frederick B. Llewellyn, M.E., Stevens Institute of Technology, 1922; Ph.D., Columbia University, 1928. Western Electric Company, 1923–25; Bell Telephone Laboratories, 1925–. Dr. Llewellyn has been engaged in the investigation of special problems connected with radio and vacuum tubes.

VOLUME XIII APRIL, 1934 NUMBER 2

THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

The Carbon Microphone: An Account of Some Researches Bearing on Its Action—*F. S. Goucher* 163

Open-Wire Crosstalk—*A. G. Chapman* 195

Symposium on Wire Transmission of Symphonic Music and Its Reproduction in Auditory Perspective:

Basic Requirements—*Harvey Fletcher* 239

Physical Factors—*J. C. Steinberg and W. B. Snow* 245

Loud Speakers and Microphones—*E. C. Wente and A. L. Thuras* 259

Amplifiers—*E. O. Scriven* 278

Transmission Lines—*H. A. Affel, R. W. Chesnut and R. H. Mills* 285

System Adaptation—*E. H. Bedell and Iden Kerney* 301

Abstracts of Technical Papers 309

Contributors to this Issue 313

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

50c per Copy *$1.50 per Year*

THE BELL SYSTEM TECHNICAL JOURNAL

Published quarterly by the
American Telephone and Telegraph Company
195 Broadway, New York, N. Y.

EDITORIAL BOARD

Bancroft Gherardi — H. P. Charlesworth — F. B. Jewett
L. F. Morehouse — E. H. Colpitts — O. B. Blackwell
D. Levinger — O. E. Buckley — H. S. Osborne
Philander Norton, *Editor* — J. O. Perrine, *Associate Editor*

SUBSCRIPTIONS

Subscriptions are accepted at $1.50 per year. Single copies are fifty cents each. The foreign postage is 35 cents per year or 9 cents per copy.

Copyright, 1934

PRINTED IN U. S. A.

The Bell System Technical Journal

April, 1934

The Carbon Microphone: An Account of Some Researches Bearing on Its Action *

By F. S. GOUCHER

A great variety of speculations in regard to the physics of microphonic action has arisen because of the complexity of behavior when current passes through a so-called "loose contact" which forms the essential element in a carbon microphone. Technical difficulties arising from the minuteness of the contact forces and movements between contacts when in a sensitive microphonic state have retarded the establishment of a quantitative theory.

Recent studies of carbon contacts have led to a satisfactory picture of the nature of such contacts and their mode of operation when strained, both from the elastic and the electrical point of view. The surfaces of the carbon particles are microscopically rough and when two such surfaces are brought together under the action of compressional forces, both the number of hills in intimate contact and the contact area between hills vary through deformations which are primarily elastic. Changes in electrical resistance under strain are consistent with the assumption that current passes through the regions in intimate contact.

Introduction

FEW electrical devices are as widely used as the "carbon microphone" and few have given rise to as much speculation in regard to their mode of action. That the problem has proved elusive is shown by the fact that in Bell Telephone Laboratories it has been regarded as perennial. However, recent researches have thrown a considerable amount of light upon it and it therefore seems fitting to bring before you this evening a brief survey of the subject and an account of some of the latest experimental work.

The widespread use of the "carbon microphone"—it is employed almost exclusively throughout the world in commercial telephone service—is due primarily to its unique property of being its own amplifier. In converting acoustical into electrical waves, it magnifies the energy about one thousandfold. Other microphones, such as the condenser or electromagnetic type, are unable to do this and so require separate amplifiers when used in practice. For this reason, it seems unlikely that the carbon microphone will be supplanted in the near future for at least the great bulk of telephone work.

* Presented before the Franklin Institute, March 2, 1933. Published in the *Journal of the Franklin Institute*, April, 1934.

The essential element of this device is what has come to be called the "loose contact"—or, as its name implies, a contact between two conductive solids, metals as well as carbons, held together with small forces. The ability of "loose contacts" to transmit speech was discovered independently by Emile Berliner in this country and Professor D. E. Hughes in England. Following Hughes' discovery, Mr. Spottiswood, the president of the British Association in 1878, described it thus: "The microphone affords another instance of the unexpected value of minute variations—in this case, electric currents; and it is remarkable that the gist of the instrument seems to be in obtaining and perfecting that which electricians have hitherto most scrupulously avoided, viz., 'loose contacts.'" Hughes applied the word "microphone" to his instrument because of its remarkable "ability to magnify weak sounds." The word itself is a revival of a term first introduced by Wheatstone in 1827 for a purely acoustical device developed to amplify weak sounds. Although originally confined to the "loose contact" type of instrument, the term microphone has more recently been used—particularly in broadcast, public address, and sound picture work—for any device which converts sound into corresponding electric currents.

Evolution of the Carbon Microphone

The story of the development of the "loose contact" type of microphone is a fascinating one and, although it is beyond the scope of this

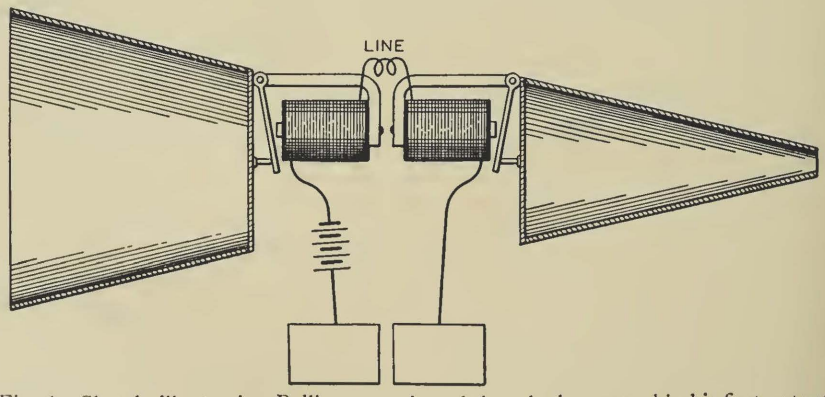


Fig. 1—Sketch, illustrating Bell's conception of the telephone, used in his first patent application of 1876.

paper,[1] I should like to refer briefly to a few of the stages in the evolution of the present day instrument. You will recall that Bell's original telephone (Fig. 1) was electromagnetic in principle and acted

[1] For a more complete account see paper by H. A. Frederick, "The Development of the Microphone," *Bell Telephone Quarterly*, July, 1931.

both as a transmitter and as a receiver. It was, however, very inefficient and Bell himself suggested that some other principle such as that of variation of electrical resistance might overcome the difficulty. He therefore devised the liquid transmitter in which a small platinum wire (Fig. 2), attached to a drumhead of gold-beaters

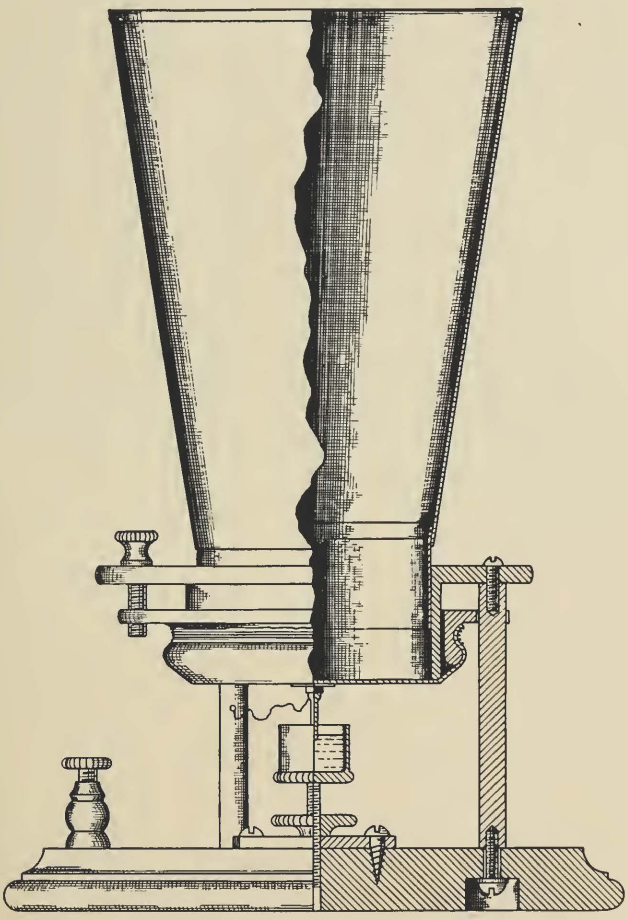


Fig. 2—Bell's liquid transmitter.

skin, is dipped into a small quantity of acidulated water in a conducting cup. The extent of the area of contact between the liquid and the wire is altered by the motion of the latter, thus altering the resistance in a continuous manner. It was with this instrument that the first complete sentence, "Mr. Watson come here—I want you," was successfully transmitted on March 10, 1876. This achievement

stimulated others to work on the problem of a variable resistance element and many new devices appeared in the next few years, the most sensitive of which utilized a single loose contact, carbon in one form or another being used as the contact material.

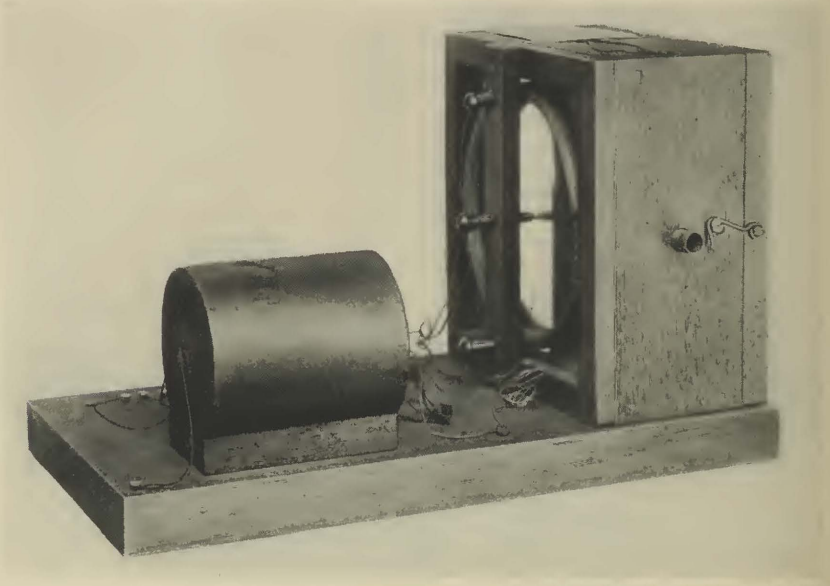


Fig. 3—Berliner's first single contact microphone, invented in 1877, employing a metal-to-metal contact.

Figure 3 shows Berliner's first successful model consisting essentially of a metal contact pressed against a metal diaphragm. This was developed later into a carbon-to-carbon contact along the same lines (Fig. 4).

Hughes, too, used metal in his first successful attempt at transmitting sounds. Only three ordinary nails were required to demonstrate the great sensitivity of loose contacts to acoustical vibrations (Fig. 5). Hughes later developed the pencil type of microphone (Fig. 6) in which carbon was used. It was the forerunner of many practical devices developed along this line.

More rugged, reliable and permanent than either of these types was the Blake transmitter shown in Fig. 7. It utilized a metal-to-carbon contact and it owed its success to the mechanical control of the contact pressure. This instrument was used for many years by the Bell System.

Then came the Hunnings or the first of the granular carbon micro-

phones (Fig. 8), the immediate ancestor of the granular carbon type used today. Hunnings used powdered "engine coke." It carried

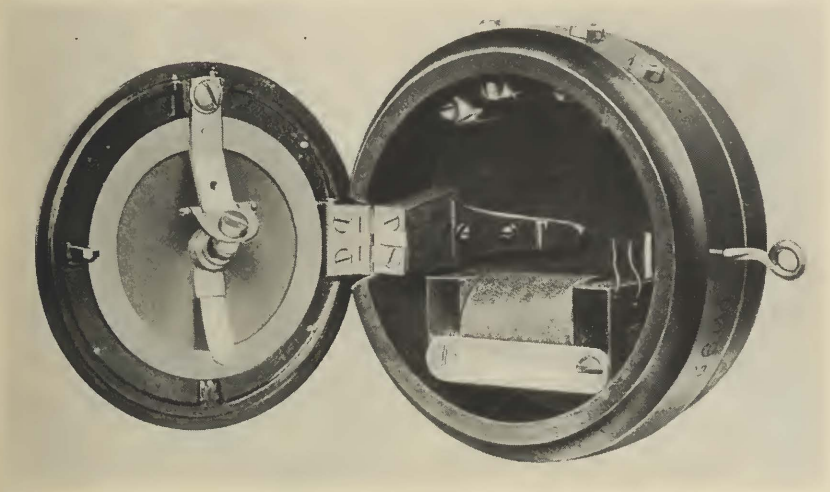


Fig. 4—Carbon-to-carbon single contact transmitter brought out in 1879 by Berliner.

more current than the Blake transmitter but it was liable to "pack" and become insensitive.

This difficulty was overcome in the design invented by White in 1890, called the solid back type (Fig. 9). Millions of these are used today in the ordinary desk-stand instrument. In this, carbon granules

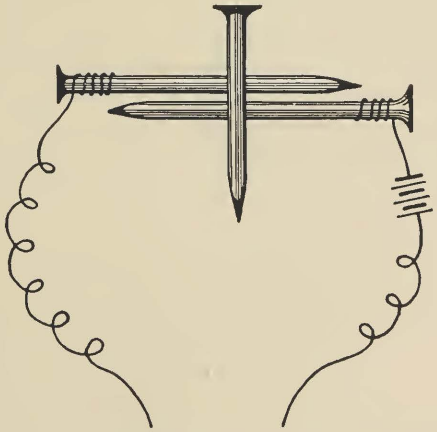


Fig. 5—Nail contacts used by Professor Hughes in 1878 to demonstrate their microphonic properties.

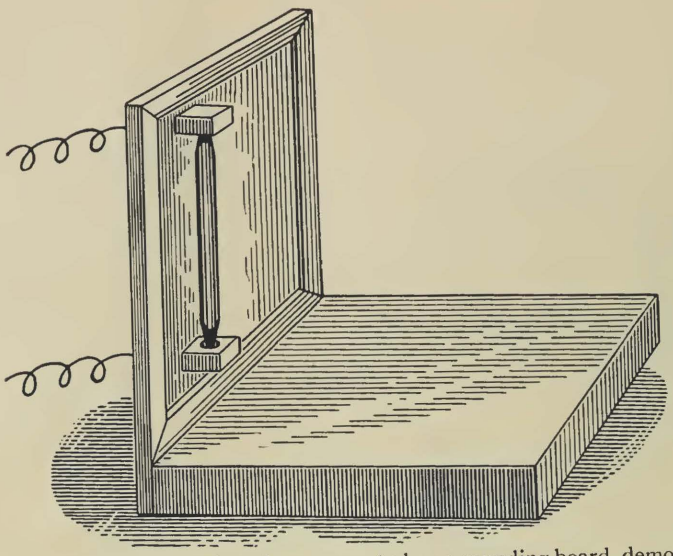


Fig. 6—Carbon pencil type microphone, mounted on a sounding board, demonstrated by Hughes in 1878.

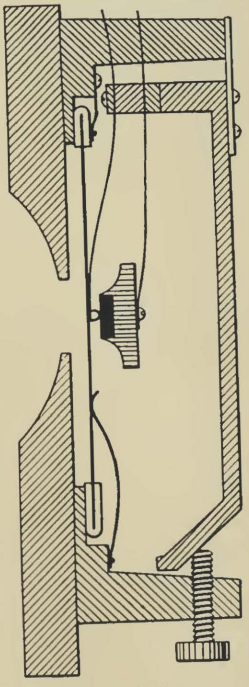


Fig. 7—The Blake transmitter using a platinum contact pressed against a carbon block.

are compressed between two polished carbon electrodes which are immersed in the granular mass in such a way that the particles have more freedom of movement than in the Hunnings instrument. This relieves excess pressure without undue packing.

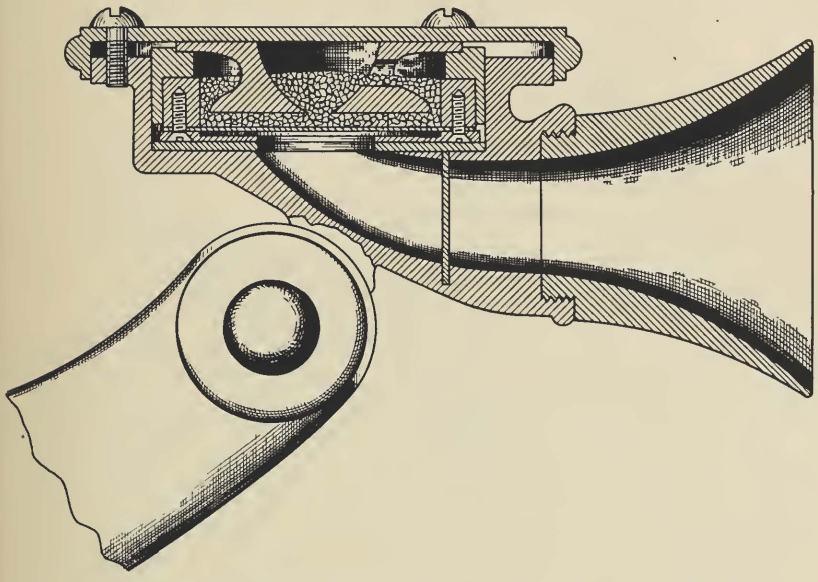


Fig. 8—Commercial model of the early Hunnings transmitter in which granular material was first used.

In Fig. 10 we have a cross-sectional view of a modern handset transmitter. This instrument, which is designed to operate in a wide variety of positions, follows the Hunnings' type in that the granular mass rests against the diaphragm but it differs from it in that the diaphragm does not act as an electrode. Both electrodes, separated by an insulating barrier, form part of the containing walls of the cell holding the carbon. This is the type which has recently been studied in detail and of which a two dimensional model is shown in Fig. 26.

The carbon used in these instruments is made by a heat treatment of anthracite coal. The particles are about 0.01 inch in size and when magnified they look just like lumps of coal taken from the domestic pile (Fig. 11).

Speculations of the Early Inventors

Part of the difficulty in elucidating the microphonic action of the "loose contact" arises because so many effects can be observed or are

associated with the action that it is hard to determine which of them is essential. It is therefore not surprising that there was great diversity of opinion amongst the early inventors.

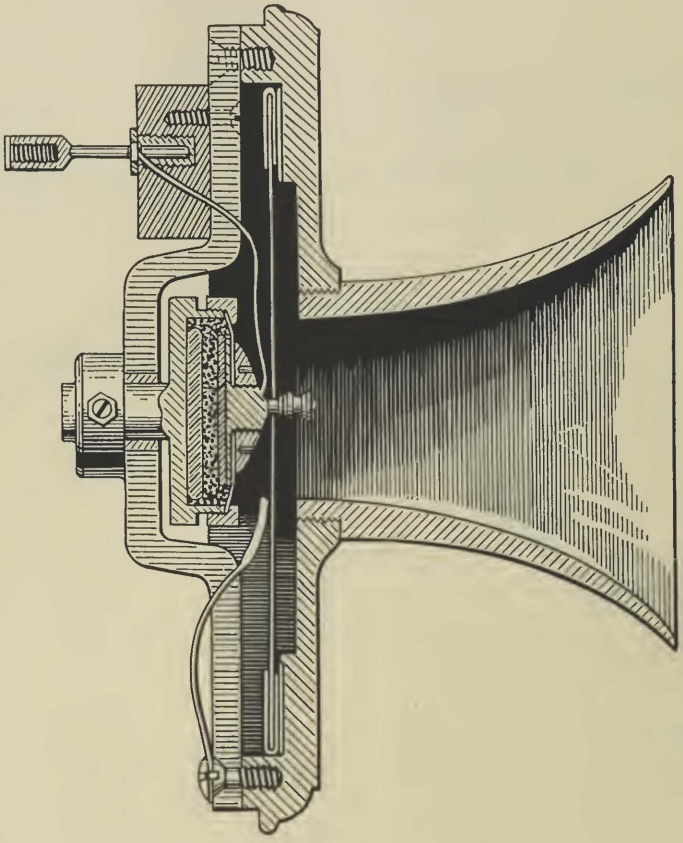


Fig. 9—The solid back transmitter invented by White in 1890.

For instance, experiment shows that contacts tend to move apart when in the act of transmitting sound. This led many, amongst them Berliner, to hold the view that an air film is necessary for microphonic action, that the current somehow passes through the film, and that the variation of the current is due to the variation of the thickness of the film. This view, however, was partly discredited by experiments showing that the moving apart was probably due to a heating of the contact through the passage of current and hence that it is not a necessary accompaniment of microphonic action.

Again, when one listens through a receiver placed in a circuit containing a "loose contact," noises are heard, especially when the

voltage across the contact or microphone is large. These noises are irregular like frying or crackling. Also, if a contact be viewed under a microscope, bright spots are sometimes seen. These facts have led many to think that small arcs are always present and are responsible for microphonic action. Hughes was very much inclined to this view.

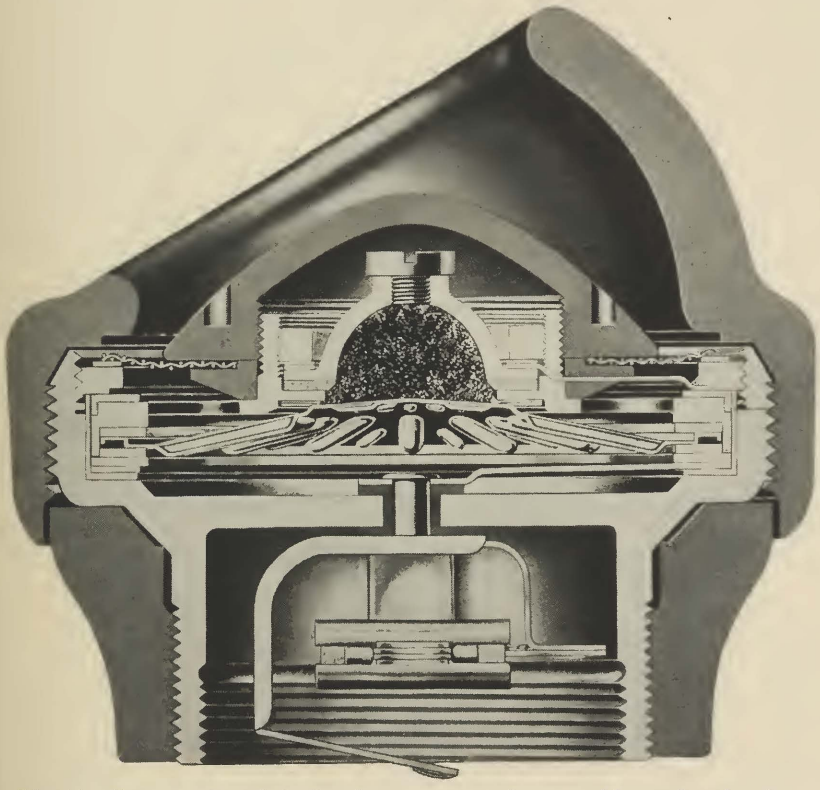


Fig. 10—Cross-section of the barrier type transmitter used in modern handset instruments.

There were reasons for supposing that the heating of the contact is a necessary factor in microphonic action. This point of view was supported by Preece, who wrote in 1893, "Indeed there are many phenomena such as hissing and humming that are clearly due to what is known as the Trevelyan effect, that is, the motion set up by expansion and contraction of bodies which are subjected to variation in temperature. This at least tends to favor the heat hypothesis as does also the fact that with continuous use some transmitters become essentially warm."

Another view was that microphonic action arises from change in resistivity of the solid carbon resulting from strain. This view was held by Edison who doubtless believed it because of the success of his microphone which was designed with the object of applying pressure variation to a solid carbon block. It failed of general



Fig. 11—Carbon granules made from anthracite coal (× 15).

acceptance because the effect of pressure on resistance, as shown by experiment, seemed definitely to be too small. It was generally considered that the Edison instrument was in fact a "loose contact" although Edison himself did not realize it.

Others of the early inventors considered the contact area to be the essential element—that is to say, the extent of surface or the number of molecules involved in intimate contact. As Professor Sylvanus Thompson expressed it in 1883, "An extremely minute motion of approach or recession may suffice to alter very greatly the number of molecules in contact. . . . Just as in a system of electric lamps in parallel arc the resistance of the system increases when the number of lamps is diminished and diminishes when the number of lamps connecting the parallel mains is increased, so it is with the molecules at the two surfaces of contact."

Recent Theories

The first attempt at a quantitative theory of microphonic action was made by Professor P. O. Pedersen in 1916.[2] He assumed that microphonic action is due to the variation of the contact area arising from the elastic deformation of the contact material by pressure. Considering the case of two elastic conducting spheres brought into contact, Pedersen assumed that the resistance is made up of two parts; viz., (1) the resistance of a conducting film having a specific resistivity differing from bulk carbon and independent of pressure, and (2) the so-called "spreading resistance" or that which is caused by the concentration of the current flow within the region of the contact area and which would exist independently of any film.

This theory results in a quantitative expression [3] for the dependence of the contact resistance on the force holding the contacts together. Pedersen tested it by experiments on carbon spheres and found reasonable agreement over a wide range of force. However a very similar expression can be obtained without postulating the existence of the high resistance film. We have merely to suppose that contact does not take place over the whole contact area owing to surface roughness (the existence of which can be observed under a microscope, especially in the case of carbon).

Dr. F. Gray of Bell Telephone Laboratories worked out an expression [4] based on this assumption which was so nearly like Pedersen's that it was difficult to discriminate between them experimentally. He assumed both that the number of microscopic hills in electrical contact increases as the contact force is increased and that the resistance per hill varies in accordance with the theory of spreading resistance as assumed by Pedersen. His equation was found to fit experimental curves remarkably well for contact forces which are relatively larger than those holding the granules together in a microphone. In the range of smaller forces, however, marked departures from theory were found, the measured value of resistance decreasing too rapidly with an increase of force. Although these departures were believed to be due at least in part to a plastic deformation of the contact material, it appeared possible that other factors come into play and may even be dominant in this region of small contact forces.

For instance, it had been demonstrated that adsorbed films of air are capable of producing a marked increase in the resistance of granular carbon contacts. This revived the air film theory as a possibility under the condition of small contact forces.

[2] *The Electrician*, Jan. 28–Feb. 4, 1916.

[3] $R = AF^{-2/3} + BF^{-1/3}$.

[4] $R = AF^{-7/9} + BF^{-1/3}$ (*Phys. Rev.*, **36**, 375, 1930).

Again there is a marked decrease in the resistance of granular carbon contacts with increase in voltage which had not been satisfactorily explained. This fact suggested amongst other possibilities that the conduction process may involve the passage of electrons across gaps of molecular dimensions in the manner of a cold point discharge. Field gradients of sufficient magnitude to extract electrons from a solid must exist in these gaps with only a fraction of a volt across the contacts. If this is the main process by which current passes between contacts, microphonic action might well be associated with a variation of the gap dimensions under strain.

Again recent work on the theoretical strength of solids had led to experimental results showing that under certain conditions solids may, without fracture, be subjected to strains greatly exceeding those heretofore obtained. This suggested the possibility that the microphonic effect of contacts might after all be associated with the straining of small junctions welded under pressure and current.

In view of the speculative nature of the situation it was clear that a new experimental attack on the problem was necessary. We have been making such an attack during the last few years and I now turn attention to some of the experimental results and the main conclusions to be drawn from them.

Recent Experimental Work

Statement of the Problem

Since the essential element in the carbon microphone is the so-called "loose contact," the first and most fundamental step toward the understanding of the physics of microphones is the solution of the problem of the "loose contact" when in its sensitive or microphonic state.

Measurements on microphones such as the handset have enabled us to specify pretty accurately the conditions under which any two granules within the structure operate when the microphone is transmitting speech or sound.

In addition to the voltage, which is limited to one volt per contact, these conditions may be stated briefly either in terms of contact forces or in terms of movements between centres of granules. When you realize how small these are—particularly the movements between centres of granules—you will, I think, not be surprised that the solution of the problem of the "loose contact" has been so long delayed.

For the condition of reasonably loud speech the diaphragm motion is about

$$1 \times 10^{-5} \text{ cm.},$$

which is just on the limit of resolution of the highest-power microscopes. It follows from a consideration of the number of granules in series that the movement between centres of granules would not be greater than 1/10th of this, viz.,

$$1 \times 10^{-6} \text{ cm.},$$

which is in the submicroscopic range. We must, therefore, be able to control and measure movements at least as small as 10^{-7} cm.; not an easy thing to do with a "loose contact."

The contact forces are on the average somewhat less than 10 dynes when the aggregate is in the unagitated state. In the presence of acoustic waves, variable forces of several dynes are superimposed on these fixed forces. The variable forces are smaller than the fixed forces, so that the granules will on the average remain in contact throughout any reversible cycle. We have reason to believe that 10 dynes is about the maximum force which is attained at any one contact during a stress cycle. We must therefore be able to control contact forces within the range 1 to 10 dynes.

Apparatus and technique have now been developed for studying single contacts within the prescribed range of forces and displacements, and significant measurements have been made which I will now endeavor to describe to you somewhat in detail.

Single Contact Studies

Figure 12 shows the construction of one of the contact tubes used in this study.

Its essential features are shown diagrammatically in Fig. 13. The contact pieces C_1 and C_2 are fastened respectively to a movable base M and to the lower end of a helical spring made of fused quartz. The base is supported from a fixed frame by two vertical platinum wires P and two stretched springs as shown. The lower contact piece is moved by heating or cooling the platinum wires through the passage of current. In this way the contacts may be made or broken and any desired contact force applied, the measure of the force being the compression of the helical spring. The temperature of the contact is varied by surrounding the contact region with a metal cylinder S which may be heated by means of radiation from a coil of platinum wire H, the temperature within the cylinder being measured by means of a thermocouple placed near the contacts.

In practice the upper contact piece consists of a single granule fastened to the end of a platinum wire and the lower contact piece consists of a number of granules attached to a horizontal metal plate;

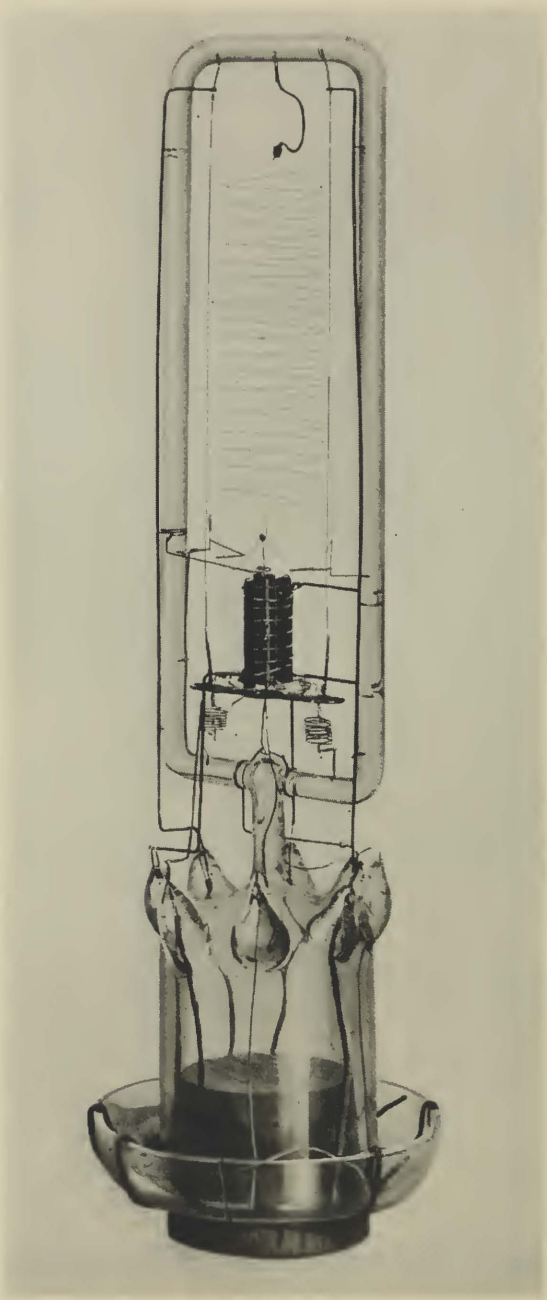


Fig. 12—Device for controlling force and temperature used in the study of single contacts.

in this way a variety of contacts can be studied with the same tube. A small hole in the metal cylinder permits of direct observation of the contacts during measurement. Figure 14 shows how the apparatus

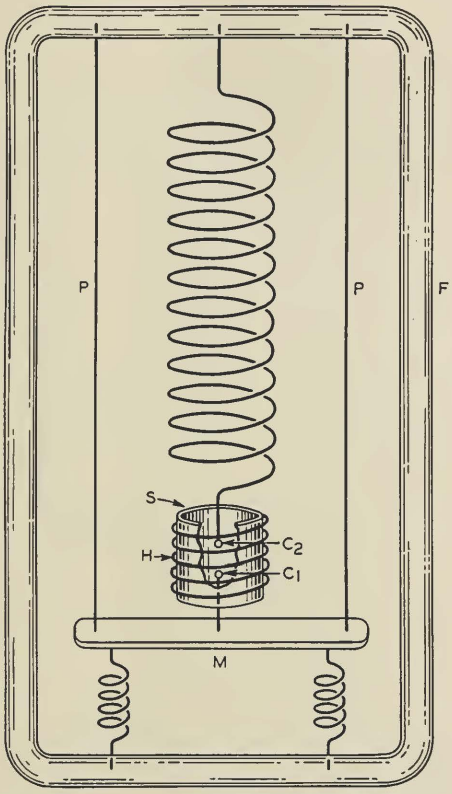


Fig. 13—Diagrammatic view of single contact device shown in Fig. 12.

was mounted in an iron cylinder on a damped suspension to protect it from acoustical and mechanical disturbance. The two microscopes were used to observe the compression of the silica spring.

We first studied the effect of voltage and temperature on contacts held together with constant forces. Reversible characteristics could in all cases be obtained for voltages up to 1 volt and for temperatures up to about 80° C.

Typical characteristics are shown on Fig. 15 in which the contact forces were of the order of 1 dyne. On the left are plotted the resistance-voltage characteristics and on the right the resistance-temperature characteristics. All of the variables are plotted for convenience on logarithmic scales.

The curves *I*, *II* and *III* illustrate the fact that Ohm's law is found to hold for all contacts up to about 0.1 volt and that above these values the contact resistance decreases with increase of voltage.

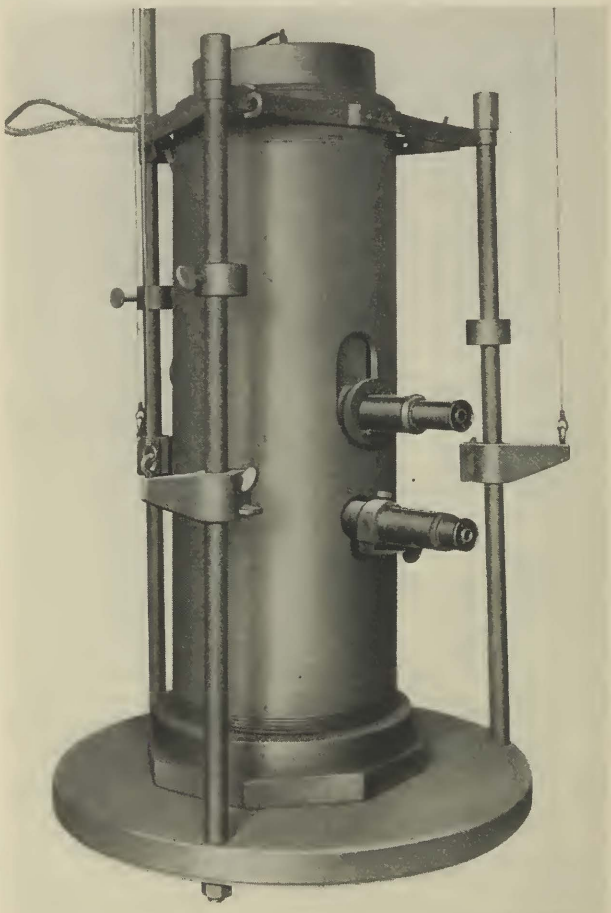


Fig. 14—The single contact device is mounted in a heavy container on a spring suspension to minimize acoustic and mechanical disturbance.

The fractional decrease in resistance with voltage above 0.1 volt is independent of the contact resistance and whether or not the measurements are made in air or vacuum.

In curves *I'*, *II'* and *III'* we have changed the voltage scale of the curves *I*, *II* and *III* to a temperature scale in accordance with the relation,

$$T = T_0 + 40V^2.$$

This relation has a theoretical basis in the Joule heating of the contacts due to the passage of current and contains the assumption of a value of Wiedemann Franz ratios characteristic of solid carbon.[5]

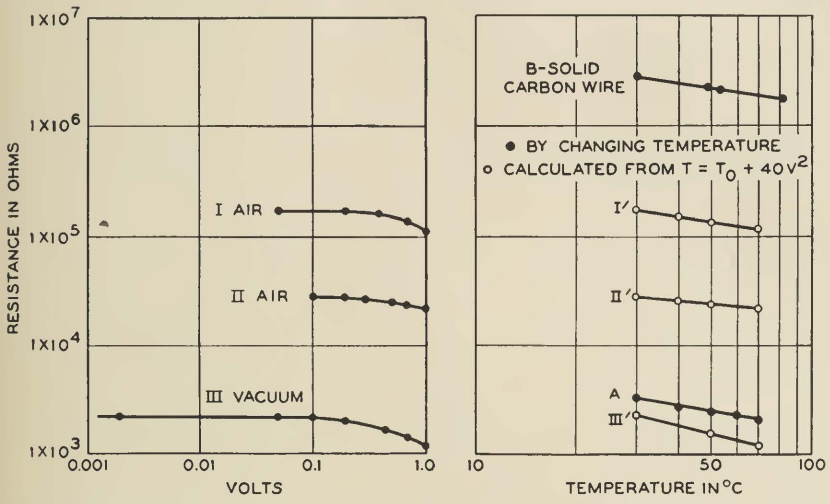


Fig. 15—Characteristics showing the effect of voltage and temperature on contact resistance.

These curves have substantially the same slope as A, which is a characteristic measured by heating a contact in the furnace, the contact voltage being sufficiently small to avoid appreciable heating of the contact due to this cause, and also with B, which was obtained with a solid carbon wire produced in a manner to simulate closely microphone carbon. We are able to conclude from measurements such as these that the nature of the conducting portions of contacts is that of solid carbon both for air and vacuum and that the departures from Ohm's law—at least up to 1 volt—are due to the Joule heating of the contacts.

From measurements similar to these in which we show that the admission of air has no effect on the temperature coefficient of resistance—although it produces a marked increase in the resistance at any particular temperature—we are also able to conclude that the presence of adsorbed air does not alter the nature of the conducting portions of the contacts but merely limits their areas.

[5] This theory, based on earlier work of Kohlrausch, was worked out in useful form independently in Bell Telephone Laboratories (unpublished work) and by R. Holm (*Zeit. Tech. Phys.*, 3, 1922). It gives the approximate relation, const. $\frac{V^2}{K_0/\sigma_0}$, as the increase in temperature above room temperature, V being the contact voltage, and K_0/σ_0 the Wiedemann Franz ratio for the contact material.

Turning now to the effect of contact force on contact resistance: we see (Fig. 16) that large and approximately reversible resistance changes are produced as the force is varied repeatedly between fixed limits. This shows that the effect is in the main elastic, though the

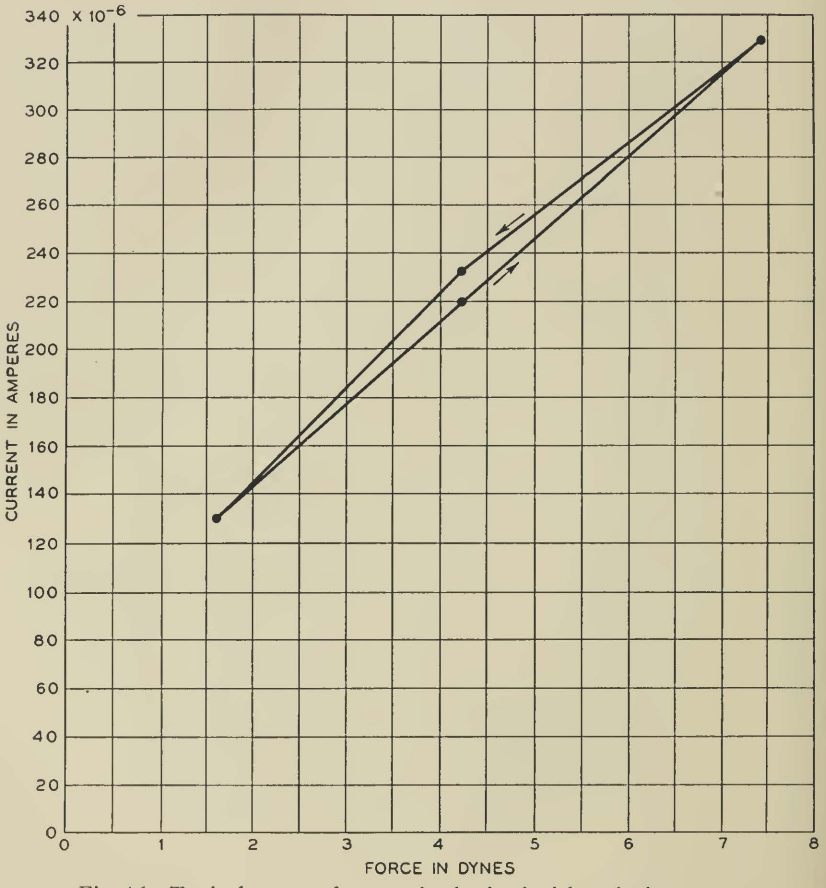


Fig. 16—Typical current-force cycle obtained with a single contact.

existence of a narrow loop indicates a small plastic or irreversible movement as a secondary effect.

We have thus established that the current is conducted through solid carbon and that the deformations are mainly elastic. These facts give strong support to the "elastic theory" of "loose contacts," i.e., the hypothesis that the change of resistance takes place because of a change in contact area under pressure. An extensive study of the resistance-force characteristics gave results which could not be

simply interpreted (just as Gray had found) and, because of the possibility that unknown cohesional or frictional forces were involved, the work was extended by a study of resistance-displacement characteristics. Through a comparison of the two sets of data we were led to the conclusions that the stress-strain characteristics are not so simple as those assumed in Pedersen's or Gray's analysis and, therefore, that a study of the elastic behavior of contacts offered the most promising line of attack on the problem.

Figure 17 shows the mechanical system developed for this purpose. With it known forces can be applied to a contact element and at the same time its movement can be measured.

The contact is made between a carbon granule and a polished carbon plate, the granule being attached to the end of a rod R suspended by springs S from a fixed frame and the plate being attached to the end of a micrometer screw M_2 capable of giving to it a translational motion without rotation.

The force is applied to the granule electrostatically by means of voltage applied between the condenser plates C_2, one of which is attached to the rod R and the other to the micrometer screw. This is in principle the attracted disc electrometer of Kelvin and it is capable of applying forces up to 15 dynes without using voltages greater than 200.

The motion of the granule with respect to the carbon plate is measured electrically through the variation of capacity of the condenser C_1, of which one plate is attached to the other end of the rod R. C_1 forms part of an oscillating circuit of natural frequency n_0 (about 2000 kc.) which is coupled to a wave-meter circuit adjusted for oscillation at a frequency n_1 slightly different from n_0. Changes in the frequency arising from the changes in capacity C_1 alter the energy picked up by the wave-meter circuit and this energy, which is recorded by means of a galvanometer, serves as a measure of the change of capacity or motion of the rod R. With this arrangement it is possible to measure motions as small as 1×10^{-7} cm. and under the best conditions as small as 1×10^{-8} cm. It is necessary to have good damping, which is obtained by means of immersing the drum D in polymerized castor oil. The accessory spring S_2 is used merely for calibrating purposes.

Figure 18 shows the appearance of the apparatus as set up for measurement. The condenser is contained in the lower housing at the left, the wave-meter in the upper housing. The whole apparatus including the galvanometer is supported on a delicate spring suspension within a second large lead container, the frame of which just appears at the edge of the photograph and which is also supported by springs.

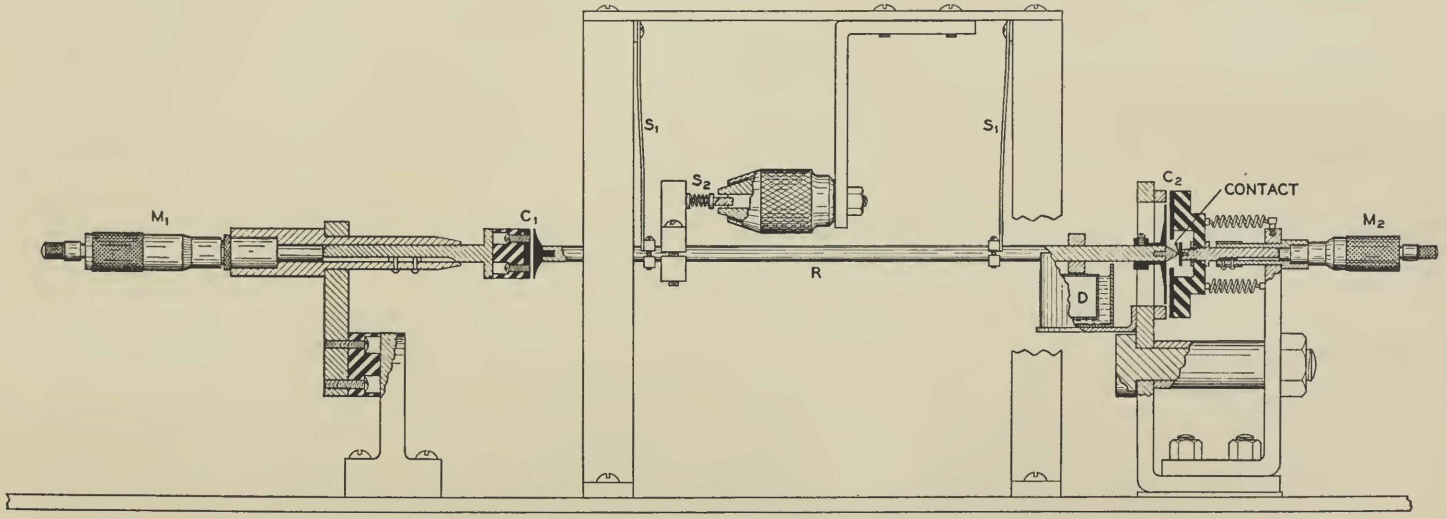


Fig. 17—Diagrammatic view of mechanical system used in the study of single contacts.

Figure 19 shows the appearance of the complete setup with the cover on the outside container. This begins to compete with cosmic ray apparatus from the point of view of the amount of lead involved, the outer container weighing about 600 lbs. Port-holes—one of

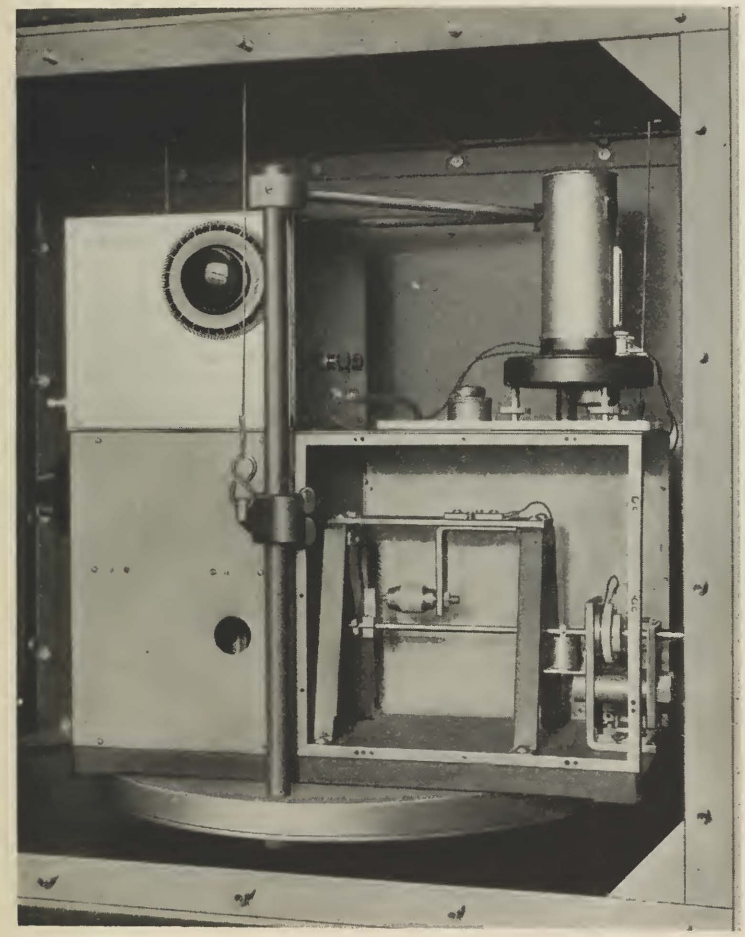


Fig. 18—Mechanical system and associated electrical apparatus as set up for single contact study.

which appears on the near end of the box—permit adjustments to be made on the apparatus within, thus eliminating the necessity for removing the large outer cover which, as you may surmise from the number of handles, requires the combined efforts of two men to

remove it. All of this protection is, of course, to shield the apparatus from mechanical vibrations and acoustic disturbances.

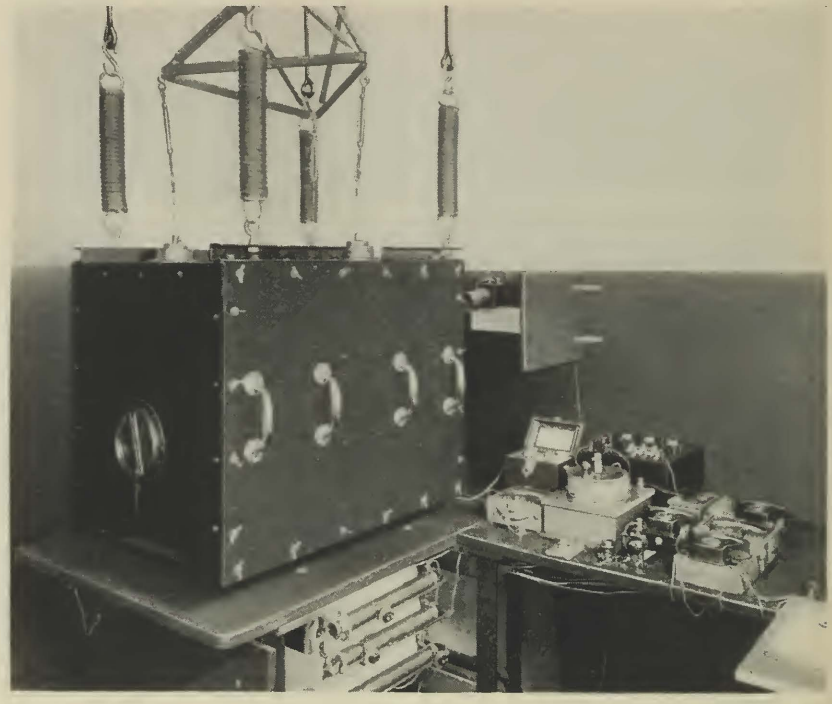


Fig. 19—Exterior view of complete experimental arrangement.

With this apparatus we investigated the variations in displacement and resistance when the forces are varied cyclically between fixed limits. Measurements on a large number of contacts are summarized in curves, Figs. 20 and 21. The cyclic characteristics, though somewhat irregular and having the form of narrow loops, approximate straight lines when the variables are plotted on logarithmic scales. Only one complete characteristic is shown in each set of curves, other typical measurements being represented by dotted straight lines joining the end points of their respective cycles. The full line in each figure represents the cycle of a typical contact, obtained by averaging, over the range in which the difference between the maximum and minimum force limits or maximum and minimum displacement limits is relatively large, in which case the slope is apparently constant.

If we let N'' and N represent the slopes of the typical force-displace-

ment and resistance-force characteristics and if F, D and R be the contact force, the contact displacement and the contact resistance,

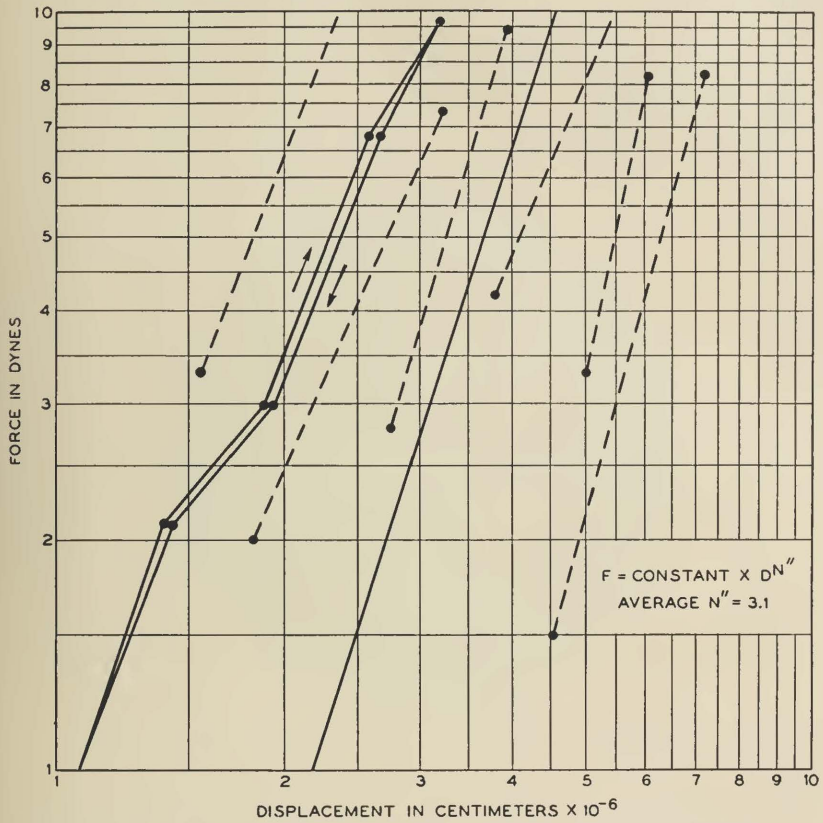


Fig. 20—Typical force-displacement characteristics of carbon granules pressed against a polished carbon plate.

respectively, we may express our results by the approximate relations:

$$F = \text{const.} \cdot D^{N''}, \tag{1}$$

$$R = \text{const.} \cdot F^{-N}. \tag{2}$$

The values N'' and N are not, however, independent of the force or displacement limits when these limits are relatively small. In Fig. 22 we have plotted values of N'' and N as functions of the difference between the maximum and minimum displacement (ΔD). We see that for relatively large values of ΔD, N'' and N approach the limiting values 3.1 and 0.47, respectively, but for smaller values of ΔD, N''

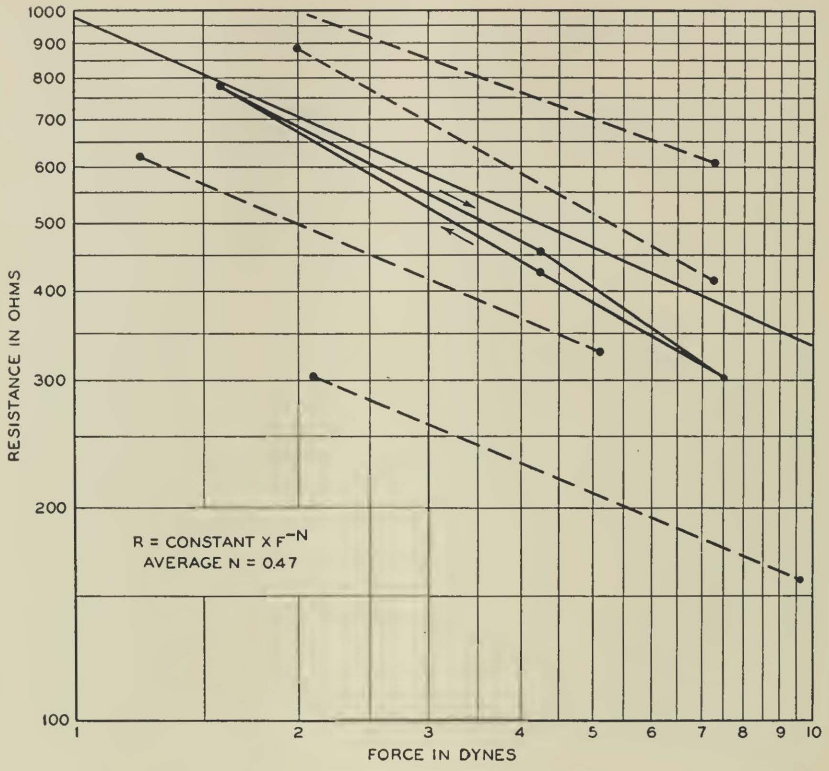


Fig. 21—Typical resistance-force characteristics of carbon granules pressed against a polished carbon plate.

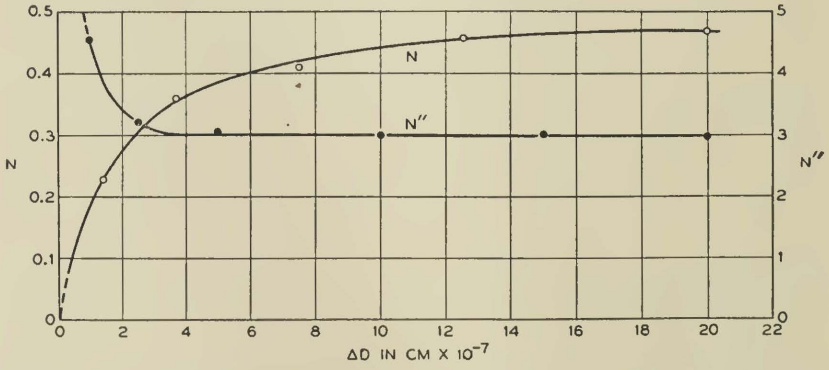


Fig. 22—Effect of the extent of contact motion (ΔD) on N and N''. (Average values.)

becomes greater than and N less than its limiting value. The limiting value of N'' is greater than that which would be obtained through the contact of hemispherical surfaces and represents a more rapid stiffening of the contact with compression.

We will first give our attention to the limiting value of N''.

A consideration of the nature of contact surfaces as revealed by the microscope furnished the clue to the interpretation of our results. A typical surface is shown in the photomicrograph (Fig. 23). Evi-

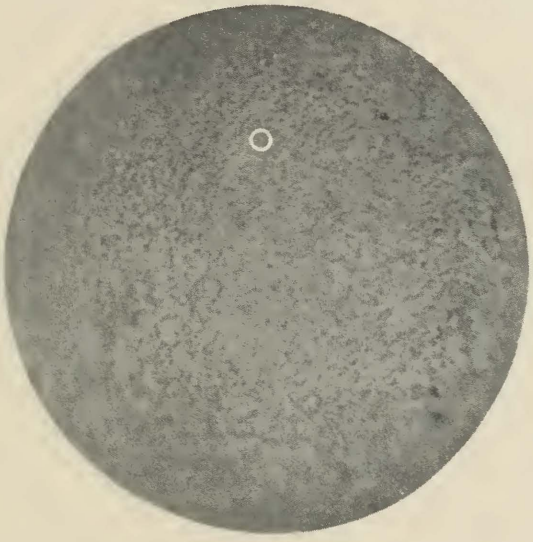


Fig. 23—Photomicrograph of the surface of a carbon granule ($\times$ 2400).

dently it is very hilly, the hills being much the same size and height. The magnification ($\times$ 2400) is such that the small white circle has a diameter of 8×10^{-5} cm. and it is clear that the circle encloses several hills.

From the theory of elasticity we may deduce that if two hemispherical hills of carbon having a radius of the order 1×10^{-5} cm. are brought together with forces of the order of 1 dyne the maximum stresses will probably not exceed the elastic limit of carbon and hence that the hills will deform elastically. The motion involved in such a deformation will be of the order of 1×10^{-6} cm. and if other hills are encountered, as is most probable with such a movement, the stresses will be shared and hence the stress per hill reduced. According to this view forces larger than one dyne can be applied without exceeding the elastic limit merely by virtue of the distribution of the hills which will come in to share the stresses. Furthermore, such a contact will

stiffen up more rapidly with compressional displacement than will a contact made on a single hill. This concept of a loose contact, therefore, seemed to offer possibilities in the way of an adequate explanation of the experimental results.

At first the problem seemed too complex for mathematical analysis and a study of the elastic behavior of contact surfaces having various arrangements of little hemispherical hills was made with the aid of large scale rubber models. Quarter inch rubber balls were cut in half for this purpose and arranged on bases of suitable material and shape.

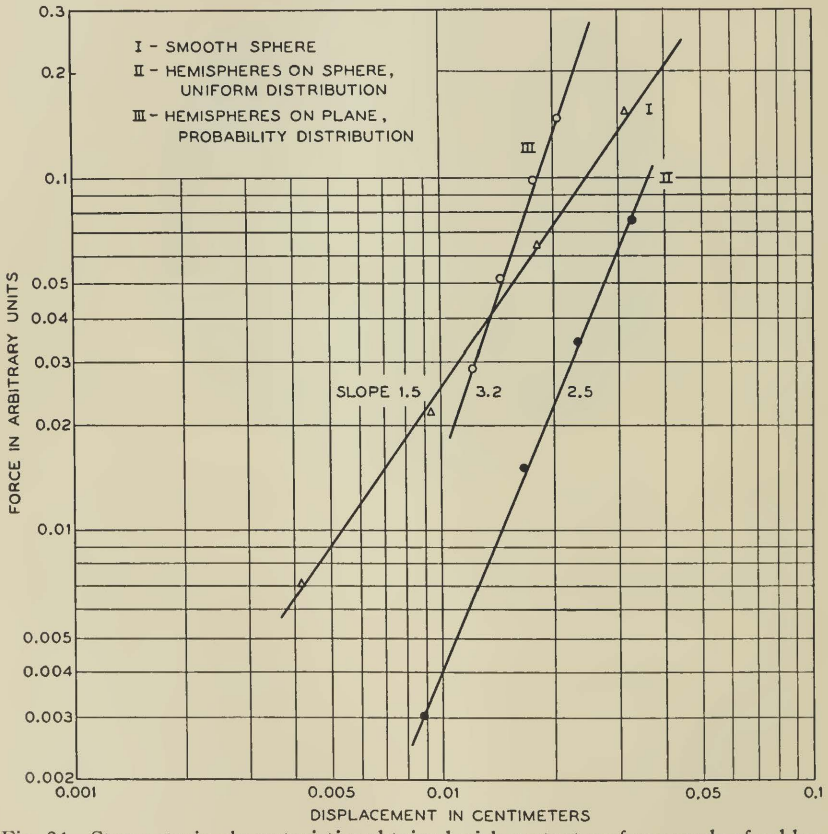


Fig. 24—Stress-strain characteristics obtained with contact surfaces made of rubber.

In Fig. 24 we have plotted the force-displacement characteristics of three different surfaces: *I*, that of a single smooth hemisphere; *II*, that of small hemispheres of equal height evenly distributed on a portion of a large 32 inch sphere made also with rubber; and *III*,

that of hemispherical surfaces of random height fastened to a flat plate, about 100 hemispheres being used.

We see from the slopes of these curves that the model made with hills of random height on a flat plate behaves most like the actual contacts, the slopes of the corresponding curves being 3.2 and 3.1, respectively. This arrangement is also the one which most nearly represents the carbon surfaces as viewed under the microscope. Here the hills have various heights and the radius of the underlying base (0.015 cm.) is so much larger (1000 fold) than that of the average hill that within the region of the contact area the surface of the former may be regarded as plane.

The slope of curve *I* is in accord with a formula derived from the theory of elasticity by Hertz connecting the force F pressing together two elastic spheres and the movement D between the centres of the spheres:

$$F = \text{const. } D^{3/2}. \tag{3}$$

The constant includes such factors as the elastic moduli of the contact materials and the radii of the spheres and need not concern us here. The case of a sphere pressed against a flat plate, as in our experiments, is a particular case of this general equation, the constant only being affected.[6]

The slopes of curves *II* and *III* are also in accord with theory, as we shall see, when one makes the simple assumption that the elastic deformation is confined to such a small region near the contact in each hill that the underlying base is not appreciably deformed. This assumption was tested in the case of the model having the spherical distribution of hemispheres by changing the stiffness of the rubber used in the underlying sphere. No effect was produced on the stress-strain characteristic (curve *II*). We may therefore consider that the elastic reactions produced in each hill are independent of each other and that the base is not deformed, so that with a given distribution of hills it becomes a simple matter to calculate their combined effect over a given compressional range. We may represent the conditions essential for our calculation by the diagram, Fig. 25, in which A represents the plane surface of the smooth contact element just making contact with the highest hill of the rough contact element. Under compression, A may be considered as moving in the direction of its normal x, compressing B and, with increasing motion, coming into

[6] Formula (3) is known to hold accurately for values of D not greater than about 1 per cent. of the radius of the sphere (J. P. Andrews, *Phys. Soc. Proc.*, Vol. 42, No. 236). This condition is fulfilled in the case of curve *I* but D is as great as 10 per cent. of the radius in the case of a few of the hills involved in the maximum compression shown in curves *II* and *III* (Fig. 24).

contact with other hills C and compressing them according to equation (3). The position of C is conveniently defined by its distance X from the plane A.

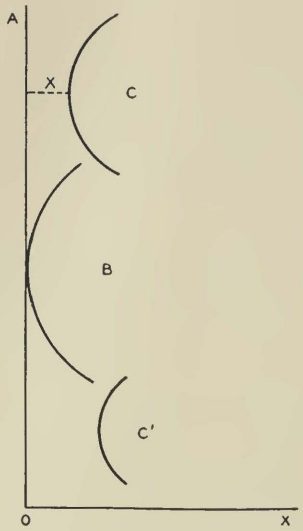


Fig. 25—Schematic representation of a rough surface used in mathematical analysis.

Any continuous distribution of hill positions, typified by C, which would be encountered through a small compressional movement, may be approximately represented by the expression,

$$N_x = \text{const. } x^n, \tag{4}$$

where N_x is defined as the number which multiplied by dx gives the number of hills coming into contact with the plane when it moves from x to $x + dx$. The exponent n is a constant which for convenience we may call the distribution constant.

For a total compression D the $N_x dx$ hills will be compressed an amount $D - x$, and hence the total force of reaction F is given by

$$F = \text{const.} \int_0^D x^n (D - x)^{3/2} dx,$$

which integrates to the form,

$$F = \text{const. } D^{n+5/2} = \text{const. } D^{N''}. \tag{5}$$

The constant here includes a summation of the individual constants of equation (3). It is clear that if the hills have different radii the constant only will be affected, so that equation (5) may be regarded as general in this respect.

For the case of uniform hills distributed on the surface of a sphere it may be shown that equal numbers of hills will be added for equal increments in x, in which case $N_x = \text{const.}$[7] From this it follows that $n = 0$ and $N'' = 2.5$ in agreement with the measured value, curve *II*.[8] For $N'' = 3.2$ as obtained with the hemispheres of random height on a plane, curve *III*, n would have the value 0.7. The corresponding distribution function N_x would approximate to that of the portion of an ordinary error curve near its maximum. A rough determination of the distribution of heights amongst the small rubber hemispheres showed in fact that they approximated closely to an error curve and that the displacement range covered that portion of the curve near the most probable height.

It would appear from this analysis that the elastic behavior of our carbon contacts under conditions of relatively large strain is adequately explained on the very simple assumption that the hills which we observe under the microscope have a random distribution of heights and behave like smooth spherical surfaces. We have, however, still to account for the hysteresis and the large values of N'' corresponding to small values of ΔD as well as the values of N (Fig. 22).

It is unlikely that the hills which we observe under the microscope are submicroscopically smooth, in which case we would expect a small plastic movement in these secondary hills arising from overstrain. We have direct evidence for this in the fact that contacts once established—even without the passage of current—require relatively small but finite forces to break them. Such junctions within the contact region could well account for hysteresis and a stiffening up of the contact in the region of small strains. Furthermore it is to be expected that they might affect the resistance behavior to a much greater extent than the elastic, and over a wider range of strain, since the junctions—though too weak to affect appreciably the contact stiffness—might well carry a relatively large proportion of current; in which case the value of N would be smaller than that calculated on the assumption of smooth spherical surfaces.

We will now derive an expression relating resistance and force for the type of contact considered in the derivation of equation (5), assuming smooth hills.

Classical theory[9] gives the following formula for the conductance

[7] This argument rests on the fact of geometry that if A is the area of contact between a sphere of radius r and a plane, $dA/dx = 2\pi r$.

[8] This agreement between theory and experiment shows that the compression of some of the hills by an amount in excess of 1 per cent. of their radii has not affected the applicability of equation (3) to our problem.

[9] Riemann Weber.

It is here assumed that the mechanical and electrical areas of contact are coincident, which according to the ideas of wave mechanics may not be the case.

$1/r$ of the contact formed by compressing, by an amount D, a single smooth conducting sphere against a flat conducting plate,

$$\frac{1}{r} = \text{const. } D^{1/2}. \tag{6}$$

It appears reasonable to assume that the hills which come into contact with compression act independently of each other as regards conduction. The conductances may therefore be added and we may write for the total conductance $(1/R)$ produced by a compression D involving many hills:

$$\frac{1}{R} = \text{const.} \int_0^D x^n (D - x)^{1/2} dx,$$

which integrates to the form,

$$\frac{1}{R} = \text{const. } D^{n+3/2},$$

which in combination with (5) gives

$$R = \text{const. } F^{\frac{2n+3}{2n+5}} = \text{const. } F^{-N}. \tag{7}$$

Using the value of n consistent with equation (5) through the measured value of N'', viz., $n = 0.6$, we get $N = 0.68$. The measured value of N (0.47) is, as we have surmised, too small though it is of the right order of magnitude.

We are, of course, investigating the factors which give rise to this discrepancy as they will play an important part in any complete theory of microphonic action, and we are extending our study to the behavior of granular aggregates in simple cells and microphone structures. We have shown that the value of N in a simple cell composed of parallel electrodes is quite consistent with our simple theory for single contacts, which therefore indicates that the behavior of an aggregate of contacts is determined by the behavior of the individual contact. Furthermore, we have shown, through static measurements on the handset instrument, that the granular aggregate within this irregularly shaped structure behaves like the aggregate in a simple cell. We are therefore confident that the behavior of the microphone will be explained in terms of the behavior of the single contact.

The behavior of the two dimensional model of the handset microphone (Fig. 26) is most convincing in this connection. Although this model was set up originally to study the distribution of stresses in this type of structure it has proved most useful in other phases of our work. Quarter inch rubber balls represent the granular particles of

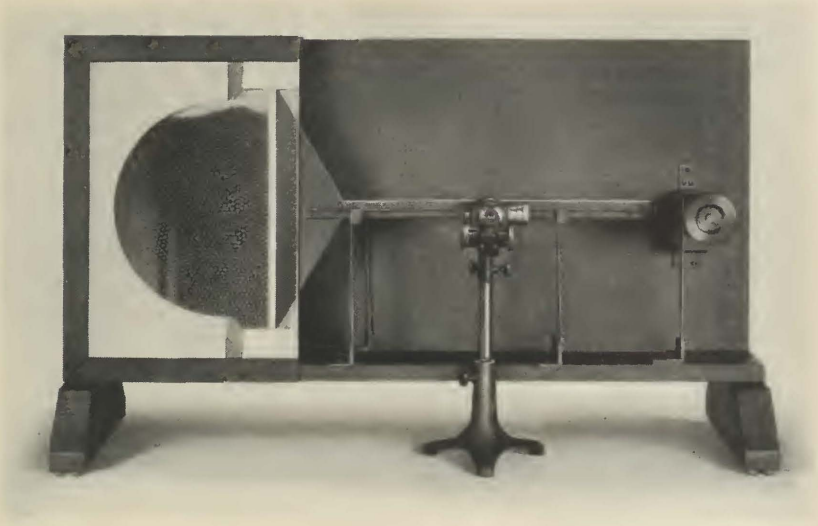


Fig. 26—Model of handset transmitter cell.

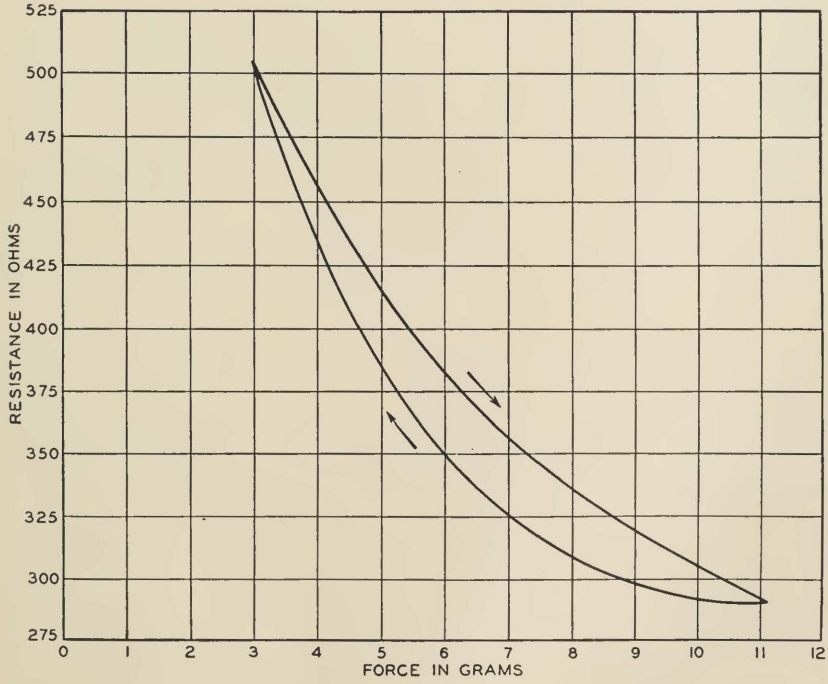


Fig. 27—Resistance-force cycle obtained with transmitter model.

the actual microphone and by coating these with a conducting layer of graphite and lacquer we are able to make them behave electrically as well as elastically in accordance with our simple theory. When placed in the model the aggregate is compressed cyclically by means of the piston which acts as a diaphragm, producing a change of resistance in the current path around the insulating barrier. The curves shown in Figs. 27 and 28 show typical resistance-force cycles, obtained with the model and the actual instrument under conditions wherein the reactive forces are mainly elastic. The similarity of these characteristics is striking. The existence of the loops indicates that the reactive forces are not entirely elastic and that the behavior is modified by friction, as in the case of single contacts.

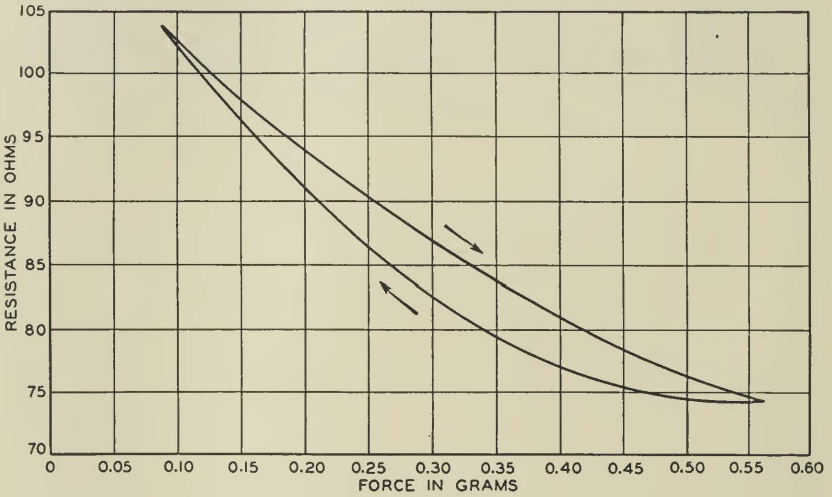


Fig. 28—Resistance-force cycle obtained with a standard transmitter.

In conclusion it seems fair to say that our experiments on "loose contacts" under conditions which are equivalent to those under which they operate in actual microphones have given a satisfactory picture of the essential nature of such contacts, and their mode of operation when strained, both from the elastic and the electrical point of view. The electrical current is carried through regions in intimate contact and changes in resistance under strain are due both to a variation in the number of microscopic hills which form the carbon surface and to area changes at the junctions of these hills arising from their elastic deformation in accordance with the well known laws of elasticity.

Open-Wire Crosstalk *

By A. G. CHAPMAN

Effect of Constructional Irregularities

IF the cross-sectional dimensions of an open-wire line were exactly the same at all points and if the transpositions were located at exactly the theoretical points, the crosstalk could be reduced by huge ratios by choosing a suitable transposition arrangement and interval between the transposition poles.

Practically, however, the crosstalk reduction is limited by unavoidable irregularities in the spacing of the wires and of the transposition poles. There is no point in reducing the type unbalances by transposition design beyond the point where the constructional irregularities control the crosstalk.

Transposition Pole Spacing Irregularities

The following discussion covers the method of estimating the crosstalk due to irregularities in the spacing of transposition poles and the derivation of rules for limiting such irregularities. With practical methods of locating transposition poles, the effect of the pole spacing irregularities may ordinarily be calculated by considering only the transverse crosstalk. Special conditions for which attention must be paid to interaction crosstalk are discussed later. The simplest case, that of transverse far-end crosstalk due to pole spacing irregularities, will be discussed first.

A transposition section is divided into segments by transposition poles which in practice vary in number from four to 128. Each segment causes an element of crosstalk current at a circuit terminal and this element is about proportional to the segment length. For far-end crosstalk between similar circuits all these crosstalk current elements would add almost directly if there were no transpositions. The function of the transpositions is to reverse the phase of half the current elements. The segments corresponding to the reversed current elements may be called the minus segments. If the other half of the

* This is the second half of a paper which was begun in the January 1934 issue of the *Technical Journal*, giving a comprehensive discussion of the fundamental principles of crosstalk between open-wire circuits and their application to the transposition design theory and technique which have been developed over a period of years.

segments are called the plus segments, the far-end crosstalk is proportional to the difference of the sum of the plus segments and the sum of the minus segments. This difference may be called the unbalanced length and the output-to-output far-end crosstalk is this length multiplied by the far-end coefficient and by the frequency.

If the sum of the plus segments equals the sum of the minus segments, the unbalanced length will be zero. The poles of a line are necessarily spaced somewhat irregularly but for a single circuit combination the unbalanced length could be made very small by carefully picking the transposition poles so as to keep the sums of the plus and minus segments about equal. This procedure is impractical, however, because many circuit combinations must be considered and because necessary line changes would prevent the maintenance of very low initial unbalanced lengths.

In practice, therefore, the segment lengths are allowed to deviate in a chance fashion from the mean segment length. The unbalanced length varies among the various circuit combinations depending on the arrangement of the transpositions which determines the order in which plus and minus segments occur. For any particular combination, the unbalanced length has a wide range of possible values and its sign is equally likely to be plus or minus.

In any transposition section, the length of any segment may deviate from the average segment length for that section. If the sum of the squares of all the deviations in each transposition section is known, the unbalanced length for a succession of transposition sections may be estimated, that is, the chance of the total unbalanced length lying in any range of values may be estimated.

Letting S_1^2 be the sum of the squares of the deviations for the first transposition section, etc., and letting R be the r.m.s. of all the possible values of the total unbalanced length in all the sections, the following approximate relation may be written:

$$R^2 = S_1^2 + S_2^2 + \cdots \text{etc.}$$

The chance of exceeding the value R may then be computed. For example, there is about a one per cent chance that the total unbalanced length will exceed $2.6R$.

In making rules for locating transposition poles the first step is to determine a value for R. For example, if consideration of tolerable crosstalk coupling indicated that there should not be more than one per cent chance that the total unbalanced length in a 100-mile line would exceed one mile, then R, the r.m.s. of all possible values of the total unbalanced length, should not exceed 1/2.6 miles. Since R is

calculated from the values of S for the individual transposition sections, a given permissible value of R may be obtained with various sets of values of S. It seems reasonable to determine individual values of S on the principle that a transposition section of length L_s should have the same probability of exceeding a given unbalanced length as any other section of the same length and that a section of length $2L_s$ should have the same probability as two sections of length L_s, etc. On this basis, the value of S^2 for any transposition section should be proportional to the section length L_s. This leads to the rule used in practice that for any transposition section S^2 should not exceed kL_s. If L_s and S are expressed in feet, a value of three for k is found suitable for practical use. The choice of a value for k will depend, of course, upon the cost of locating and maintaining transposition poles with various degrees of accuracy and upon the effect on the crosstalk of varying the value of k.

The above rule permits a large deviation at one point in a transposition section if it is compensated by small deviations in the rest of the segments. For example, with 128 segments and a mean segment length of 260 feet, one long segment of 575 feet is permissible if the rest of the segments are 258 feet. The expression for the total unbalanced length in a succession of transposition sections assumed that the deviations varied from segment to segment in a truly random manner. The above example involves an unusual arrangement of the deviations. When there are a number of transposition sections in a line, such unusual arrangements of deviations in various sections do not have much effect on the probability that the total unbalanced length will exceed a given value.

The computation of near-end crosstalk due to pole spacing irregularities is a more complicated problem since the crosstalk elements resulting from the various segments vary in their magnitudes and phase relations because the various segments involve different propagation distances. It may be concluded, however, that the r.m.s. value of the total unbalanced length in all the sections may be expressed as follows:

$$R^2 = S_1^2 + S_2^2 A_1^2 + S_3^2 A_2^2 + \cdots.$$

This differs from the expression for far-end crosstalk in that the values of S^2 for the second and succeeding transposition sections are multiplied by attenuation factors. The attenuation factor A_1^2 corresponds to propagation through the first section to the second section and back again. The other attenuation factors are similarly defined. The above expression neglects attenuation within any particular

transposition section since this is ordinarily small. It also assumes that the rule for locating transposition poles, that is, that S^2 should not exceed kL_2, is applied for lengths having only negligible attenuation.

In making estimates of R in connection with transposition design work, it is assumed that all the segments are nominally the same length, D, and that r is the r.m.s. value of the deviations of the segments. Since r^2 equals S^2 divided by the number of segments in length L_s, r^2 should not exceed kD. R^2 may be expressed approximately in terms of r^2 as follows:

$$R^2 = r^2 \frac{1 - \epsilon^{-4\alpha L}}{1 - \epsilon^{-4\alpha D}},$$

where R and r are expressed in the same units, L is the length of the line in miles, α is the attenuation constant per mile, and D is the segment length in miles. If the line loss is 6 db or more the expression is nearly equal to:

$$R^2 = \frac{r^2}{.46Da},$$

where a is the line loss in db per mile and D is the segment length in miles. This assumes $4\alpha D$ is small compared to unity which is usually the case.

The chance that the total unbalanced length will exceed about $2.1R$ is estimated at 1 per cent.

For far-end crosstalk (output-to-output) the same assumption as to nominal segment length leads to the expression:

$$\frac{R}{r} = \sqrt{\frac{L}{D}}.$$

The general expressions given for R^2 suggest that a very long segment might be permitted at some point in the line if the deviations of the segments were properly restricted in other parts of the line. The expressions given for far-end and near-end values of R^2 were

$$R^2 = S_1^2 + S_2^2 + S_3^2 + \cdots,$$
$$R^2 = S_1^2 + S_2^2 A_1^2 + S_3^2 A_2^2 + \cdots.$$

If a very long segment at some point, such as a river crossing, were permitted, this would increase the sum of the squares of the deviations for some transposition section. For example S_3 might be abnormally large. R^2 could be kept at some assigned value by limiting S_1^2, S_2^2, etc. This procedure is not considered good practice because of the difficulty of maintaining some parts of the line with very small deviations of the segments from their nominal lengths.

A very long segment has another effect on near-end crosstalk not indicated by the above discussion. If there were no deviations in any of the segments, the near-end crosstalk would be the vector sum of a number of current elements of various magnitudes and phase angles and the sum would be small due to a proper choice of these magnitudes and angles in designing the transpositions. If a segment deviates from its normal length, the magnitude of the crosstalk due to the segment changes and the phase angle also changes. The phase angles of the crosstalk values due to succeeding segments are also changed since they must be propagated through the segment in question. For ordinary deviations in segment lengths these effects on the phase angles may be neglected.

Since transverse crosstalk is independent of transpositions occurring in both circuits at the same point, it would appear from the above discussion that the location of such transpositions need not be accurate. This is not ordinarily a question of practical importance. If some circuit combinations have both circuits transposed at a certain transposition pole there will usually be other combinations which have relative transpositions at this pole. The transposition pole is of importance, therefore, in connection with the latter combinations and the same accuracy of location is required for all transposition poles. A question of practical importance, however, is whether the above rules for locating transposition poles properly limit the interaction crosstalk. This is affected by transpositions in both circuits at the same pole as well as by relative transpositions. In the following discussion of this matter it is concluded that the effect of transposition pole spacing irregularities on interaction crosstalk may be ignored at frequencies now used for carrier operation.

The effect of deviations in segment length on interaction crosstalk is indicated by Fig. 18. This figure indicates a short part of a parallel between two long circuits a and b. A representative tertiary circuit c is also shown. The transposition arrangements are like those of Fig. 9B. In connection with the latter figure it was shown that the interaction crosstalk would be very small if all segments had the same length d. On Fig. 18, D is used to indicate the normal segment length and the deviation of two segments from D is indicated by d. Since the length AC equals the length CF, these deviations have no effect on the transverse crosstalk which is controlled by the transposition at C. The deviations affect the interaction crosstalk between the length CF and length AC.

The circuit a has near-end crosstalk coupling with circuit c in the length CF. This effect is normally practically suppressed by the

transposition in a at E. Due to the deviation d of segment CE, the near-end crosstalk between a and c in length CF will not be suppressed but will be proportional to d. There will likewise be near-end crosstalk between c and b in the length AC proportional to d. The two deviations, therefore, introduce interaction crosstalk practically proportional to d^2.

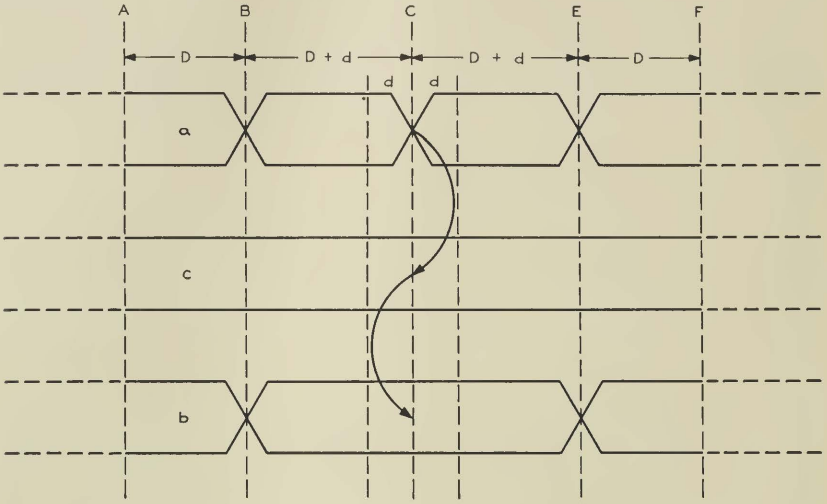


Fig. 18—The effect of deviations in segment length on interaction crosstalk.

Since there will be small deviations in numerous other segments of circuit b, the deviation d in circuit a will introduce numerous other interaction crosstalk paths similar to that discussed above. The r.m.s. value of the total interaction crosstalk caused by deviations in segment lengths may be roughly estimated as follows:

$$\frac{2FK\gamma r^2\sqrt{\frac{L}{D}}}{\sqrt{.46aD}} = \frac{.1FK^2r^2\sqrt{\frac{L}{D}}}{\sqrt{aD}},$$

where r is the r.m.s. deviation, L is the line length, D is the nominal segment length and a is the line loss in db per mile, all distances being expressed in miles. The above expression varies about as the 1.75 power of frequency and as the square of r. The corresponding expression for transverse crosstalk, i.e., $FKr\sqrt{\frac{L}{D}}$ varies as the first power of frequency and of r. It follows that, if the rules for accuracy of transposition pole spacing are relaxed or the maximum frequency is raised, the effect of pole spacing on interaction crosstalk increases more rapidly than the effect on the transverse crosstalk.

For the range of frequencies and accuracy of pole spacing used in practice, it has been found that the effect of pole spacing irregularity on interaction crosstalk is not controlling. This is indicated by Fig. 19

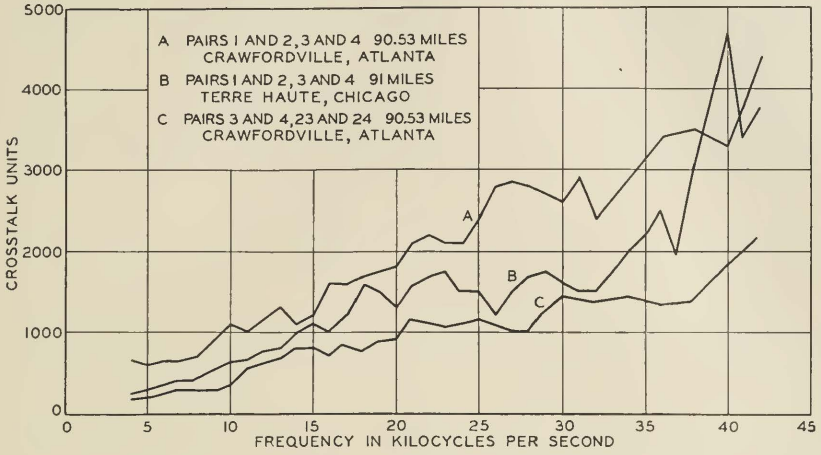


Fig. 19—Far-end crosstalk caused by pole spacing irregularities.

which shows some measurements of output-to-output far-end crosstalk between long circuits having transposition arrangements designed to make the crosstalk due to type unbalance small compared to that due to irregularities. The curves are about linear with frequency as would be predicted if the effect of the pole spacing irregularities (and wire spacing irregularities) on the interaction crosstalk is neglected. For these particular curves, a knowledge of the pole spacing indicated that pole spacing rather than wire spacing irregularities were controlling in causing crosstalk.

The above discussion assumes that a transposition section is divided by the transposition poles into segments all of the same nominal length. It is sometimes economical to use segments of different nominal lengths in the same transposition section. If the variation among the segment lengths is consistent rather than accidental it may be allowed for in the design of the transpositions.

In practice, segments of different lengths are used in the same transposition section when it is desired to adapt for multi-channel carrier frequency operation a few pairs on a line already having many pairs transposed for voice-frequency operation. Such lines often have existing transposition poles nominally spaced ten spans apart while for the pairs retransposed for carrier operation it is necessary to space the transposition poles about two spans apart. In such cases the cost

of the carrier channels is appreciably increased if uniform spacing between the new transposition poles is used. The transpositions in the pairs retransposed for carrier operation must be coordinated with the transpositions in the other circuits and it is necessary, therefore, either to divide the ten spans into four approximately equal parts with consequent expense in setting new poles at the quarter points or to retranspose all the circuits on the line.

To avoid either of these expensive procedures, the new transposition poles are nominally located in the manner indicated by Fig. 20. This

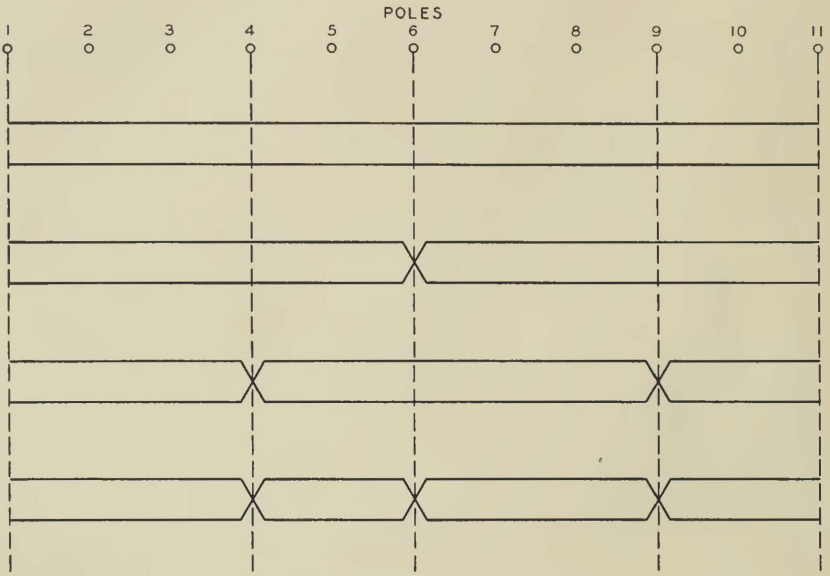


Fig. 20—Location of extra transpositions in a ten-span segment of line.

figure shows ten-pole spans subdivided into four parts in order to create three additional transposition poles. The figure indicates the location of the new transposition poles and the possible methods of transposing at these new poles. For some of the circuit combinations the crosstalk within the ten-span interval is considerably greater than if the four segments were equal in length. In each other ten-span interval the crosstalk is likewise increased by a similar inequality in segment length. Since all ten-span intervals are nominally alike, considerable crosstalk reduction may be obtained by properly designed transpositions located at the junctions of these intervals.

The use of segments of different lengths inherently decreases the effectiveness of the transpositions in reducing crosstalk and adds to

the complexity of the transposition design problem. Uniform segments are therefore used except in special circumstances.

Wire Spacing Irregularities

In the past there has been a tendency to permit wire spacing irregularities in order to reduce the cost of construction and maintenance. For example, "H fixture" crossarms formerly had special wire spacing to permit the two poles to pass between pairs of wires and thus reduce the length of the arms. Another example is that of resetting a pole with a rotted base and reducing the spacing between crossarms to get clearance between wires and ground. The development of repeatered circuits and carrier current operation has increased the seriousness of the crosstalk resulting from such irregularities and made such practices generally undesirable.

There are, of course, unavoidable irregularities in wire spacing due to variations in dimensions of crossarms, insulators and pole line hardware and warping of crossarms. Corners and hills are other causes since the crossarms at a corner and the poles on a hill are not at right-angles to the direction of the wires. The most important unavoidable spacing irregularity is, however, due to variations in wire sag. Of recent years, limits have been set on wire sag deviations to insure that this effect is properly limited during construction. The main criterion adopted has been the difference in sag of the two wires of a pair. This difference is a rough measure of the crosstalk increment due to variations of the sag from normal. The crosstalk between two pairs in a given span will be abnormal if the two pairs have different sags even if there is no difference in sag for the two wires of a pair. The crosstalk is usually more nearly normal, however, than in the case of two pairs having the same average sag but different sags for the two wires of a pair. As far as practicable, all pairs are sagged alike in a given span.

The crosstalk between two pairs due to sag differences is computed much like that due to pole spacing irregularities. The change in crosstalk due to a known pole spacing deviation may, however, be computed from the crosstalk coefficient while the change in crosstalk due to a sag deviation is not related to the crosstalk coefficient in any simple way. Two methods have been used to obtain constants for calculation.

With the first method, crosstalk measurements were made on a long line (about 100 miles) having small pole spacing and type unbalance crosstalk. The r.m.s. of a number of crosstalk measurements was determined for each particular type of pair combination, for

example, for horizontally adjacent pairs. The r.m.s. of the sag differences in a representative number of spans was also determined for the two pairs of each type of combination. The two r.m.s. values for any particular type of pair combination were called R and r. The ratio of R to r gave a constant k for estimating R from a known value of r and for L_0, the particular length of line tested. For other line lengths, R is estimated from the expression $R = kr\sqrt{\frac{L}{L_0}}$. Having computed R, the chance of the crosstalk for any pair combination in a long line lying in a given range may be estimated by probability methods.

The second method of studying sag differences is more precise although much more laborious. The change in crosstalk due to introducing sag differences in but two spans is determined. The poles are specially guyed to make it possible to adjust all the wires in these spans to have practically the same sag. Turnbuckles are installed at the ends of the two-span interval for this purpose. At the center pole the wires were supported so as to slip readily and equalize the sag in the two spans.

The phase and magnitude of the crosstalk is first measured for all pair combinations with all wires at normal sag. The wires are terminated in the same way as in the measurements of crosstalk coefficients. From sag measurements on actual lines, a set of unequal sag values for *all* the wires is then selected by probability methods and the crosstalk remeasured. The vector difference between the values of crosstalk before and after introducing unequal sags is then determined. This process is repeated a large number of times in order to cover the range of sag conditions encountered in practice. An r.m.s. value of the change in crosstalk due to sag difference is then determined for each pair combination and related to the r.m.s. sag difference per pair. This permits the probable crosstalk in a long line to be estimated and the importance of sag difference crosstalk to be determined. The two methods of study were found to be in general agreement. The second method has been extensively used to study proposed new wire configurations.

Drop Bracket Transpositions

An ideal transposition would cross the two sides of a circuit in an infinitesimally small distance, there being no displacement of the wires from their normal positions on either side of the transposition. The point-type transposition indicated by Fig. 21 is close enough to the ideal for practical purposes. Its deviation from the ideal requires little consideration in transposition design. To avoid cutting the

wires, one wire is raised about 3/4 inch and the other lowered this amount at the transposition point. The drop bracket transposition

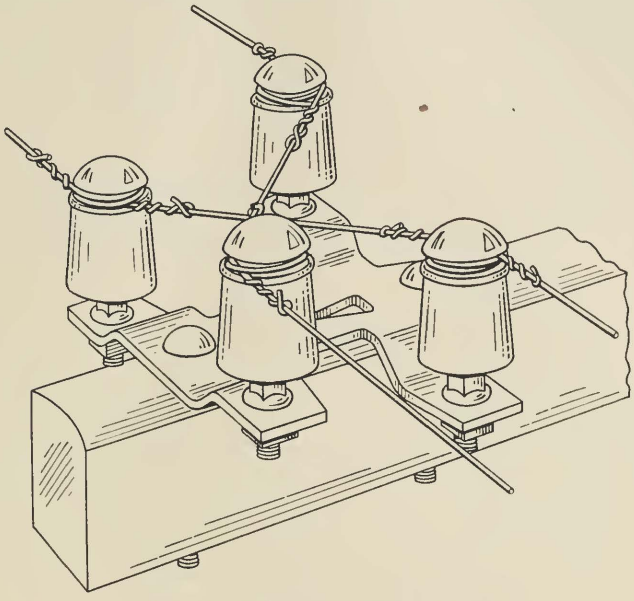


Fig. 21—Point-type transposition.

illustrated by Fig. 22 is considerably cheaper but the displacement of the wires is much greater. The effect of this displacement is important and must be especially considered in transposition design.

If all the spans adjacent to a drop bracket were of the same length

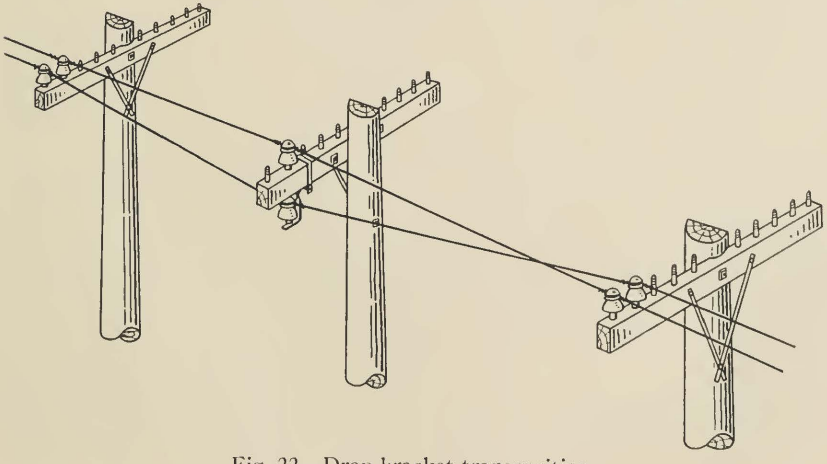


Fig. 22—Drop-bracket transposition.

and all wires could be kept under the same tension, the effect of drop brackets on crosstalk would be consistent and could, theoretically, be made negligible by a suitable transposition design.

There is, however, an accidental crosstalk effect. This effect is partly due to the fact that it is more difficult to avoid deviations from normal sag in the spans adjacent to drop brackets than in normal spans. The main effect, however, is thought to be due to inequalities in the lengths of the spans adjacent to drop brackets.

The crosstalk in such a span is very nearly proportional to the length of the span times a constant or "equivalent crosstalk coefficient." The usual crosstalk coefficient can not be used because the wires are not parallel.

Fig. 23-A indicates two long circuits, one circuit being transposed on drop brackets at the first and third quarter points of the short length D. The lengths of the spans adjacent to the drop bracket transpositions are indicated by d_1 to d_4. The equivalent far-end crosstalk coefficient for the span preceding a transposition bracket is F_1 and that for the span following the bracket is F_2. (F_1 and F_2 are usually quite different.) The total far-end crosstalk (output-to-output) due to the four spans is (very nearly):

$$K(F_1d_1 - F_2d_2 - F_1d_3 + F_2d_4),$$

where K is the frequency in kilocycles.

If the four spans were equal the crosstalk would be zero (very nearly). The actual value of the crosstalk is a matter of chance since the deviations of the four spans from the normal length are a matter of chance. These deviations cause a chance increase in the near-end crosstalk as well as in the far-end crosstalk.

This effect has been studied experimentally by using transposition designs which suppressed the consistent effect. The pole spacing effect was minimized by using very accurate spacing. The wire sag effect was allowed for by comparing similar pair combinations transposed alike except that dead-ended point transpositions were compared with drop bracket transposition. Due to the great number of transpositions necessary at carrier frequencies it was found that the accidental drop bracket effect was important at these frequencies. In recent years, point-type transpositions have been extensively used on lines transposed for long-haul carrier systems.

When, for economic reasons, a transposition system is designed for use with drop bracket transpositions, the consistent crosstalk effect must be considered in the transposition design. The equivalent crosstalk per mile for a span adjacent to a drop bracket must be

determined for each pair combination. Approximate methods of computation have been worked out for doing this and checked against measurements. The computations are involved in connection with far-end crosstalk since the "tertiary effect" is controlling. Since the

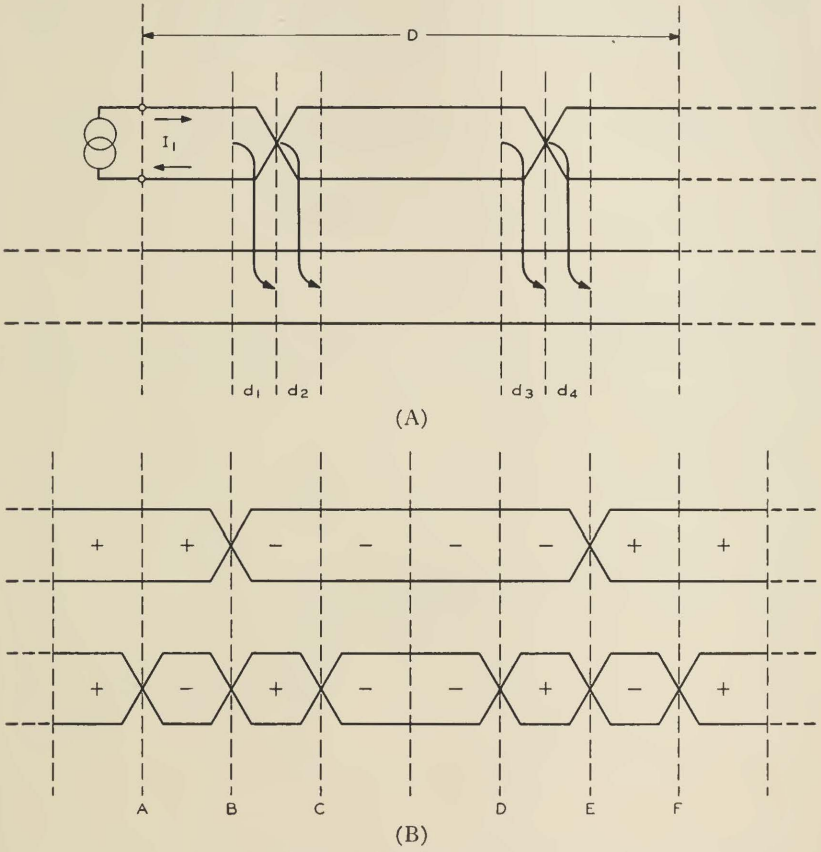


Fig. 23—Effect of drop brackets on crosstalk.

summation of crosstalk due to drop brackets is a consistent effect, "drop bracket type unbalances" can be worked out and used in transposition design. This matter is so complicated, however, that the practical method of design is to first practically ignore the drop bracket effect and then check the design to determine whether this effect has been properly suppressed.

Certain rules are adopted, however, to ensure that the transposition arrangements are properly chosen to avoid the larger drop bracket effects. Fig. 23-B indicates an arrangement of transpositions for two

pairs in a short length of line which, with point transpositions, would have very low crosstalk. At points B and E both circuits are transposed alike. With point transpositions the near-end crosstalk in the two spans adjacent to one of these pairs of transpositions would be $NK2d$, where d is the span length, N the near-end crosstalk coefficient and K the frequency in kilocycles. For drop bracket transposition the crosstalk would be $K(N_1 + N_2)d$ or a change of $K(N_1 + N_2 - 2N)d$.

The transpositions are so arranged that the crosstalk in the two spans at B tends to add to that in the two spans at E. With drop brackets at B and E the major crosstalk in this length of line would be twice the above change since the crosstalk with point transpositions is very small.

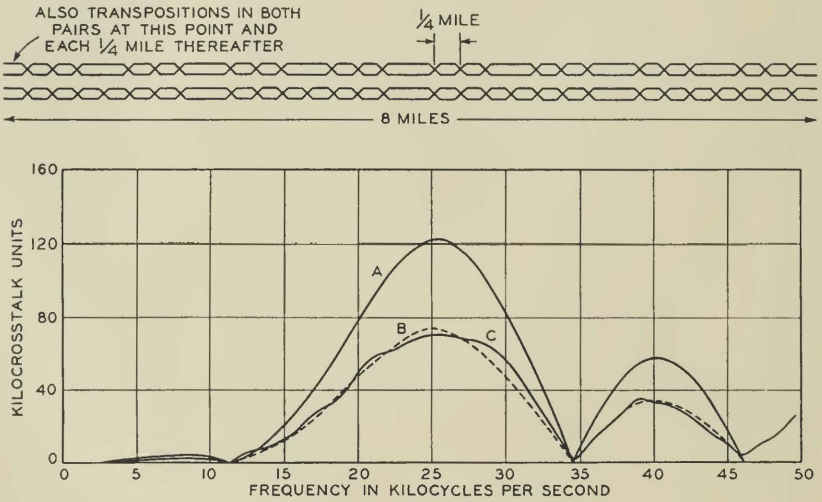


Fig. 24—Near-end crosstalk with and without drop brackets.

If the arrangement of Fig. 23-B is reiterated in a long line, the total increase in the crosstalk due to drop brackets at such points as B and E may be marked. It may be noted that the crosstalk in the two spans at A tends to cancel the crosstalk in the two spans at C and likewise there is cancellation at D and F. Drop brackets may, therefore, be used at points A, C, D and F without a consistent increase in crosstalk. Arrangements like those at B and E of Fig. 23B should be avoided in transposition design involving drop brackets.

The change in the crosstalk due to drop brackets is not necessarily an increase. Fig. 24 shows an arrangement of transpositions in an eight-mile line and three crosstalk frequency curves. Curve A shows

the calculated near-end crosstalk for ideal point transpositions. Curves *B* and *C* show the calculated and observed near-end crosstalk for drop bracket transpositions. The curves show that the drop bracket effect can be calculated quite accurately and that it may reduce the total crosstalk. In the general case, it is impractical to take much advantage of this reduction effect because a marked reduction for one combination of circuits is likely to result in an increase for some other combination and because a reduction of crosstalk in one part of the line may increase the vector sum of crosstalk elements from all parts of the line.

Wire Configurations

The crosstalk coefficients for the various pair combinations may be altered by changing the configuration of the wires. Therefore, the crosstalk for a given transposition design and a given accuracy of transposition pole spacing irregularity may also be altered. The crosstalk due to sag differences also depends on the wire configuration. It is important, therefore, to choose a configuration most desirable from the crosstalk standpoint. Such an optimum configuration requires the fewest transpositions and least accuracy of pole spacing for a given maximum frequency and given permissible values of crosstalk coupling.

Various "non-inductive" arrangements of wire configurations have been suggested and tested. Such arrangements may appear to have possibilities but their study to date has indicated that they are impracticable for more than a few pairs on a line.

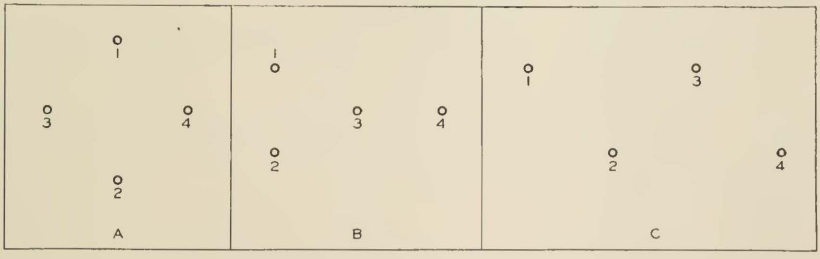


Fig. 25—"Non-inductive" arrangements for two pairs of wires.

Fig. 25 illustrates several suggested arrangements for two pairs. Arrangement *A* is often called a square phantom. If pair 1–2 is the disturber and there are equal and opposite currents in wires 1 and 2 there will be no voltages induced in either wire 3 or wire 4 because either of these wires is equally distant from wires 1 and 2. Since wires 1 and 2 are not equally distant from the ground, the currents

in these wires may be not quite equal and opposite. As a result, voltages will be induced in wires 3 and 4 but these will be equal and there will be no crosstalk current in pair 3–4. By the reciprocal theorem the crosstalk between the two pairs will also be zero when pair 3–4 is the disturber.

Arrangement *B* is nearly non-inductive. In this case if pair 1–2 is the disturber and the currents in the two wires are not quite equal and opposite due to the presence of the ground, unequal voltages will be induced in wires 3 and 4 and there will be a crosstalk current in this pair. This effect could be minimized by transposing both pairs at the same points. They would not require relative transpositions since equal and opposite currents in pair 1–2 will induce no voltage in either wire 3 or wire 4.

With pair 3–4 as the disturber, equal and opposite currents will result in equal voltages induced in wires 1 and 2. These voltages cause a phantom current in phantom 1–2/3–4. This phantom current will divide between wires 3 and 4 but can not induce unequal voltages in wires 1 and 2 because 1 and 2 are equally distant from either 3 or 4. The crosstalk coefficient is, therefore, zero both for the direct effect and for the indirect effect of the phantom. However, the indirect effect of the ground or other conductors is not zero and may require transpositions.

Arrangement *C* is non-inductive for direct crosstalk. It is not non-inductive in regard to the indirect effect of the phantom 1–2/3–4. Equal and opposite currents in pair 1–2 induce equal voltages in wires 3 and 4. The resulting equal phantom currents in wires 1 and 2 of phantom 1–2/3–4 will induce unequal voltages in pair 3–4.

When there are many pairs on a line it is not possible to make all combinations strictly non-inductive even for direct crosstalk. With perfect wire spacing the larger values of direct crosstalk per mile could be greatly reduced, however, and appreciable reductions could be obtained in the indirect effect which is usually controlling in far-end crosstalk.

Wire sag deviations must be considered, however. If a given number of "non-inductive" pairs are placed in the pole head area normally occupied by the same number of pairs with conventional configuration, the crosstalk due to sag deviations is likely to be more serious with the "non-inductive" pairs than with conventional pairs. For the same pole head area, the number of transpositions and, therefore, the "pole spacing" crosstalk could be reduced if non-inductive arrangements were used. The tests to date indicate, however, that the total crosstalk would not be reduced because of increased "sag difference" crosstalk.

The mechanical problem of supporting the wires of the "non-inductive" arrangements is considerable if serious increases in crossarm and hardware costs are not to be incurred. This objection seems at present to override the possible advantages of (1) fewer transpositions for a given pole head area and crosstalk result, or (2) fewer transpositions and lower crosstalk with a greater pole head area.

Another possibility is the use of non-parallel wires. It is possible to arrange two pairs of wires in such a way that they have a certain direct crosstalk per mile at one end of a span and the value at the other end of the span is about equal and opposite. The net direct crosstalk per mile integrated over the span is zero or small. An example of this is the barreled square formerly used abroad. Fig. 26 illustrates this arrange-

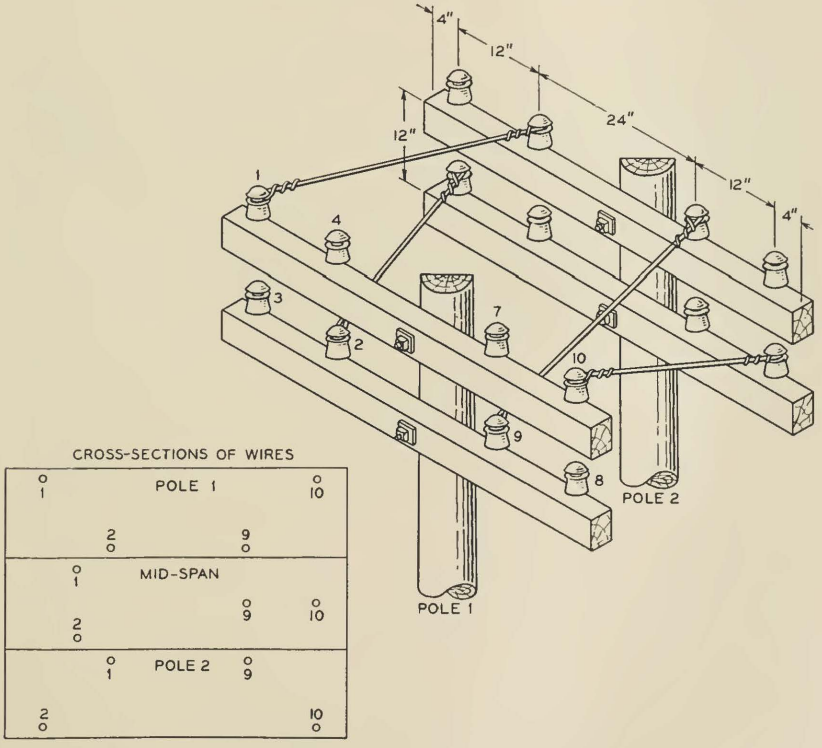


Fig. 26—Two pairs of wires in different barreled squares.

ment. The wires are arranged in groups of four, each four being arranged on the corners of a square. The two wires of a pair are on diagonally opposite corners of a square. Each pair is given a quarter turn in each span. For simplicity only two pairs in different four-wire

groups and one span are shown. The two pairs shown are nearly "non-inductive" for direct crosstalk in this span.

Consideration has been given to applying this principle to a number of pairs in order to reduce the crosstalk coefficients. Since all the crosstalk coefficients could not be made very small, transpositions would be needed. The experience to date indicates that this method does not look attractive because it is not very effective in reducing the indirect crosstalk, the mechanics of transposing are difficult, the variations in sag are likely to be abnormal and the system is complicated.

There remains the simple method of improving the configuration of the wires in a given pole head area by reducing the spacing between the wires of a pair and increasing the spacing between wires of different pairs.

The crosstalk per mile between pairs is evidently reduced by this procedure since the two wires of a pair are approaching the ideal of being equally distant from every other conductor. The "sag difference crosstalk" is also reduced and higher frequencies may be used for a given crosstalk result. Fig. 27-A and Fig. 27-B indicate a 20-wire line with the wire spacing used in the past and also the configuration commonly used today on lines where heavy carrier development is involved. The spacing between the two wires of a pair has been reduced from 12 inches to 8 inches and the spacing between pairs correspondingly increased.

It was not possible to reduce the spacing of the pole pairs and for this reason they are unsuited for the higher carrier frequencies and it is sometimes uneconomic to string them. For such cases the crossarm indicated by Fig. 27-C may be considered. The 8-inch spacing of pairs is retained but the distance between pairs is further increased. With this last crossarm, phantom circuits are not superposed on the 8-inch pairs since their use results in greater crosstalk between the pairs and restricts the possibilities of multi-channel carrier operation.

The crossarm with 8-inch pairs and pole pairs may be used on lines where multi-channel carrier operation is not employed. In such cases, the 8-inch pairs may be phantomed. Since the average spacing between the side circuits of such a phantom is not reduced by the 8-inch spacing, the crosstalk between the phantom circuits is about the same as with the 12-inch pairs. The crosstalk from a side circuit into a phantom is somewhat reduced because of the reduced spacing of the pairs. For a given pole head area it does not appear practicable to devise a configuration which will result in marked reductions in the susceptibility of both phantoms and side circuits to crosstalk and

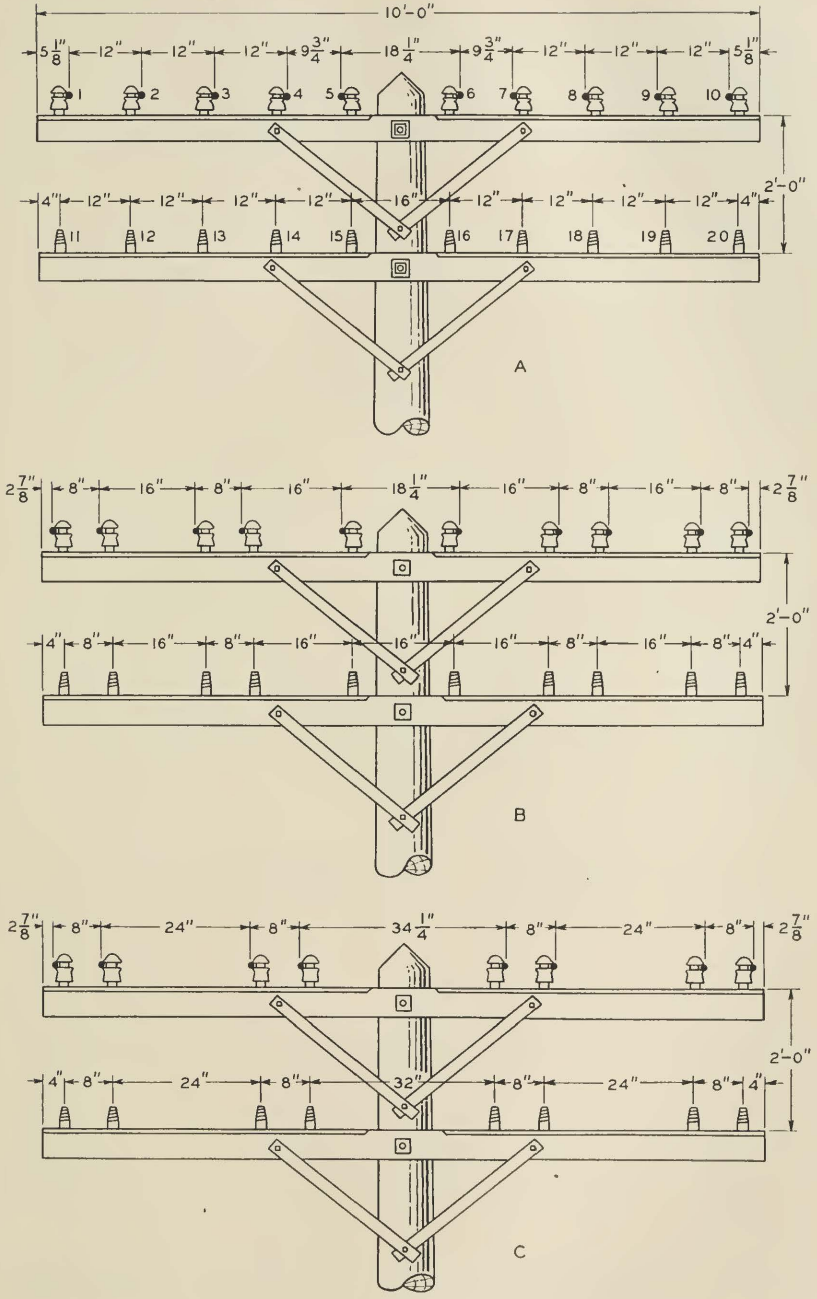


Fig. 27—Configurations of open-wire lines.

noise The "square phantom" indicated by *A* of Fig. 25 has theoretical possibilities but studies of the effect of wire spacing deviations make this arrangement appear impracticable.

The proposal to reduce the spacing of the wires of a pair from the historic value of 12 inches naturally raised the question of swinging contacts. However, extensive experience with 8-inch spacing has shown no appreciable increase in the number of wire contacts. This applies to lines where ordinarily the span length did not exceed about 150 feet. With long span crossings, crossarms were supported from steel strand at intervals of 260 feet or less.

The effectiveness of the reduction in wire spacing is indicated by the following table. The table shows the measured near-end and far-end crosstalk coefficients for important circuit combinations and for the two-pole head diagrams of Figs. 27-A and 27-B.

Crosstalk Per Mile Per Kilocycle—104-Mil Conductors

Pair Combination	Near-End Crosstalk		Far-End Crosstalk	
	12-Inch	8-Inch	12-Inch	8-Inch
1–2 to 3–4	974	439	74	34
3–4 to 7–8	133	47	77	15
1–2 to 11–12	653	326	66	30
1–2 to 13–14	40	18	58	24
3–4 to 13–14	549	288	155	69
1–2 to 21–22	163	78	35	16
1–2 to 23–24	55	28	43	17
3–4 to 23–24	107	55	75	36

General Transposition Design Methods

The preceding discussion will indicate that transposition design involves much more than consideration of the locations of the transpositions.

In practical design, the first step is to estimate the crosstalk due to unavoidable pole spacing and wire spacing irregularities for the configuration of wires under consideration and for a wide frequency range. This crosstalk represents the best that can be done with an ideal transposition design. It must be kept in mind that great precision is impracticable. The pole spacing of a line may change from time to time due to minor reroutings caused by highway changes, etc. The wire sag differences change with temperature and are affected by sleet.

If two long circuits are on adjacent or nearby pairs in one repeater section, they should, as far as practicable, be routed over non-adjacent

pairs in other repeater sections in order to minimize the overall crosstalk between these two circuits. This crosstalk will usually be largely due to those parts of the parallel where the circuits are on adjacent or nearby pairs, since the pole line seldom has enough pairs to make it practicable to keep any two circuits far apart for a large proportion of the total parallel. It is important, therefore, to strive for the lowest possible crosstalk between adjacent or nearby pairs even though this requires permitting higher crosstalk between widely separated pairs than would otherwise be necessary.

For the adjacent or nearby pairs with naturally high crosstalk, limits on the type unbalance crosstalk are set which make this type of crosstalk small compared with that due to irregularities. Since the type unbalance crosstalk varies with frequency and, in general, increases with frequency, these limits are imposed only for the range of frequency which the line will be required to transmit. It is not advisable to go beyond this, since more severe limits require closer spacing of transpositions and the increased number of transpositions would make the "pole spacing" irregularity crosstalk larger. For the well-separated pairs with naturally lower crosstalk, the type unbalance crosstalk rather than the irregularity crosstalk may be allowed to control with the same idea in mind of requiring a minimum number of transposition points.

Fig. 28 indicates the method used generally in the Bell System for arranging transpositions with 32 transposition poles. The arrangements shown are called fundamental types. They are iterative, i.e., if the first two-interval length is transposed at the center, each following two-interval length is likewise transposed, etc. Various other arrangements called hybrid types are possible but in the long run there appears to be no advantage from their use except in the case of side circuits of phantoms. In this case the transposition pattern may change when the side circuit changes pin positions at a phantom transposition.

The fundamental types may be extended to involve 64, 128, 256, etc., transposition poles. Types involving 128 transposition poles are often used.

A long line, say 100 miles, is divided into short lengths called transposition sections. With the latest transposition designs, sections having 128, 64, 32, 16 and 8 transposition poles are provided. The nominal lengths of these sections vary from 6.4 to .25 mile. The purpose of these sections is to provide an approximate balance against crosstalk (and induction from power circuits) in short lengths and thus to allow for unavoidable discontinuities in the exposure between

circuits such, for example, as points where circuits branch off the line. Transposition arrangements must be chosen for each circuit in each type of section to ensure this approximate crosstalk balance.

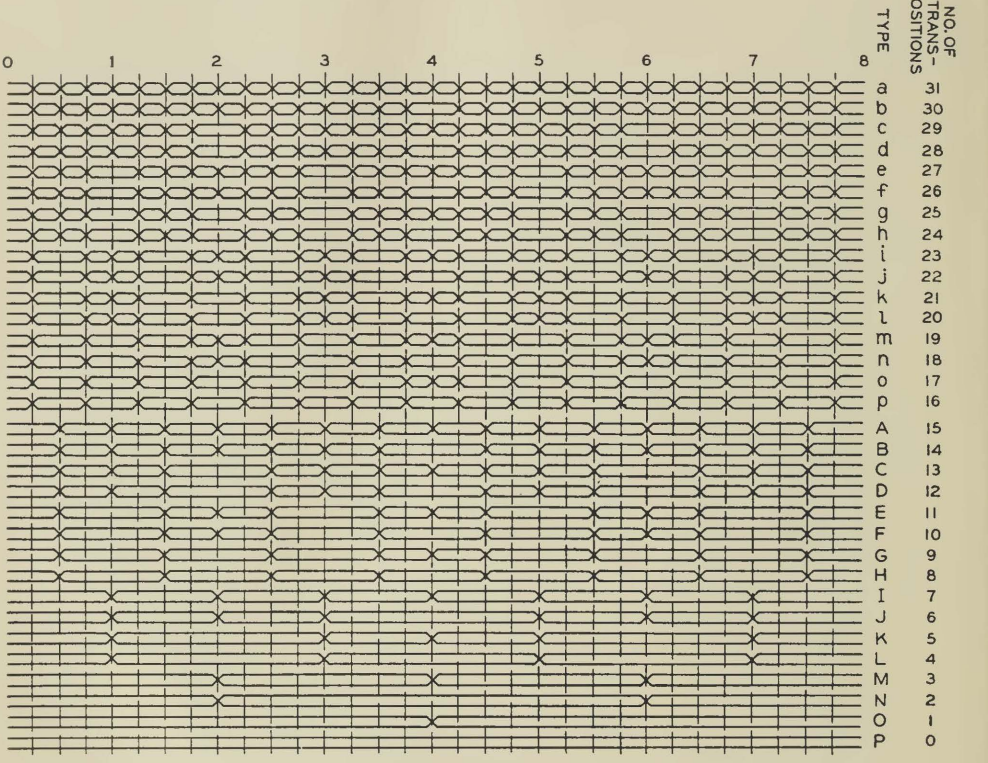


Fig. 28—Fundamental types for 32 transposition poles.

Certain lines have few, if any, discontinuities and a succession of the longest type of section is used. To improve the effectiveness of the transpositions, junction transpositions are used at the junctions of successive similar sections. For such lines it would be more effective to use longer transposition sections and not require that all circuits be approximately balanced in a short length. Such a special design would be impracticable, however, since it would be too inflexible in regard to circuit changes, etc.

In choosing the transposition arrangements for a section it must be kept in mind that the object is to meet certain crosstalk limits for a succession of sections considering both type unbalance and irregularity crosstalk. The method of procedure is discussed below.

Evolution of Transposition Designs

In designing transposition systems it must be kept in mind that much of the crosstalk is due to irregularities and is a matter of chance. Theoretically the crosstalk elements due to all of the various irregularities might chance to add directly. This is highly improbable and if the design were based on making this limiting condition satisfactory, the expense would be very great. Practically, therefore, the designs are based on exceeding a tolerable value a small percentage of the time. If, in practice, the tolerable value happens to be exceeded and this is not found to be due to an error in construction, the unfortunate adding up of crosstalk elements can be broken up by a different connection of circuits at the offices.

The tolerable values commonly chosen are 1000 crosstalk units (60 db) for open-wire carrier circuits and 1500 units (56 db) for voice-frequency open-wire circuits, which tend to have more line noise than cable or carrier circuits. These limits apply to the crosstalk between terminating test boards with the circuits worked at net losses of about 9 db.

Before proceeding with the design of the individual transposition sections which are but a few miles long, it is evidently necessary to determine what part of the overall limit can properly be assigned to an individual section. Assumptions must first be made as to typical and limiting lengths in which circuits are on the same pole line and in which adjacent or nearby circuits continue in this relation. A representative repeater layout must then be chosen. The repeater layout is very important, since the crosstalk in each repeater section is propagated to the circuit terminal and amplified or attenuated, depending on the arrangement of the repeaters. As a matter of fact, the layout of repeaters must be governed to a considerable extent by crosstalk considerations.

On the assumption that the relative magnitudes and phase relations of the crosstalk couplings in the various repeater sections are a matter of chance the tolerable crosstalk in a single repeater section can be estimated by the use of probability laws. Similarly the tolerable value for any part of the repeater section can be estimated. These probability methods apply very well to crosstalk due to irregularities. Type unbalance crosstalk is systematic, however, and in assigning tolerable values of type unbalance crosstalk in a transposition section, it is necessary to consider how the crosstalk values for various transposition sections may add up.

It is not likely that there will be systematic building up of type unbalance crosstalk in successive repeater sections and, therefore, the

tolerable crosstalk per repeater section may be estimated by probability methods. The total of the irregularity crosstalk and the type unbalance crosstalk in a repeater section is a matter of chance and may be estimated from probability theory. Conversely, the part of the tolerable crosstalk which may be assigned to type unbalance crosstalk may be estimated. As noted above, the allowance for type unbalance crosstalk for adjacent or nearby pairs is usually made so small that irregularity crosstalk controls the total. The maximum permissible carrier frequency is, then, the frequency at which the irregularity crosstalk just reaches the tolerable value. Having determined tolerable values of type unbalance crosstalk for a repeater section for the various pair combinations, tolerable values for the individual transposition sections must be determined.

If a repeater section involves a number of different types of transposition sections it is not likely that there will be a systematic building up of type unbalance crosstalk. Factors are, therefore, worked out to relate the crosstalk in a succession of similar transposition sections to that in one section. Numerous factors are required since they depend upon the transpositions at the junctions of the sections. A study of such factors indicates values which it is reasonable to assign to an individual transposition section in order to avoid excessive type unbalance crosstalk in a complete repeater section.

In the case of a voice-frequency transposition system, both near-end and far-end type unbalance limits must be set. The far-end limits are usually easily met. In the case of a transposition system for carrier systems, far-end crosstalk is controlling and the far-end type unbalance limits are important. The "reflection crosstalk" previously discussed depends, however, on both the magnitude of the near-end crosstalk and on the impedance mismatches. Information on the degree to which it is practicable to reduce these mismatches must be available in order to set limits on near-end type unbalances at carrier frequencies.

Pairs used for carrier systems are usually also used for voice-frequency telephone systems and in designing transpositions for these pairs crosstalk limits suitable for both types of systems must be met. In practice, an existing line may have only a part of the pairs retransposed for carrier operation and in designing a system of transpositions for such retransposed pairs limits must be set for the crosstalk at voice frequencies between the retransposed pairs and the pairs not retransposed.

It has been the practice to transmit certain carrier telegraph frequencies in the opposite directions used for these frequencies in

connection with carrier telephone, or, in some cases, program transmission circuits. At these frequencies near-end crosstalk limits must be set so as to limit the induced noise from the carrier telegraph.

When the type unbalance crosstalk limits are finally determined, the transposition designer must attempt to meet the requirements for all circuit combinations and all the transposition sections. It may be that the requirements can not be met and consideration must be given to modifications in the nature of the transmission systems. A vast amount of such preliminary transposition design work has been necessary in order to evolve the present transposition systems and transmission systems.

Such studies led to the development of non-phantomed circuits with 8-inch spacing since they indicated that multi-channel long-haul carrier operation on all pairs on a line was, in general, impracticable from the crosstalk standpoint with 12-inch phantomed pairs.

It may be noted that there are also difficulties in the crosstalk problem when 12-inch phantomed pairs are used for voice-frequency repeatered circuits. These circuits have a crosstalk advantage over carrier circuits in that the frequency is lower but they have an offsetting disadvantage in that they use the same frequency range in both directions. This makes the near-end crosstalk directly audible to the subscriber. As previously discussed the near-end crosstalk is inherently greater than the far-end crosstalk and, for this reason, practicable designs of multi-channel carrier systems do not allow near-end crosstalk to pass to the subscriber, the path being blocked by one-way amplifiers. While it takes fewer transpositions to control the type unbalance effects with voice-frequency transposition designs, for a given length of parallel the difficulties with crosstalk due to irregularities are about as great as with designs for multi-channel carrier operation.

The simple example of Fig. 29 illustrates the reasons for the difficulties with near-end crosstalk with the voice-frequency designs for 12-inch spaced pairs. It also illustrates the method of deducing the permissible crosstalk per repeater section as discussed above.

This figure indicates two paralleling repeatered circuits, each having six repeater sections of 10 db loss and five repeaters of 10 db gain. The net loss of each circuit is, therefore, 10 db. The near-end crosstalk values in the six sections are indicated by n_1 to n_6. The crosstalk coupling at A due to n_2 is just equal to n_2 since there is no net loss or gain in either circuit between A and B. There is also no net loss or gain between A and C, A and D, A and E or A and F. The total crosstalk coupling at A is, therefore, the vector sum of the six values

n_1 to n_6. If the crosstalk is due to irregularities the exact values of n can not be calculated but from the data collected on the crosstalk due to irregularities, the r.m.s. of all possible values may be estimated.

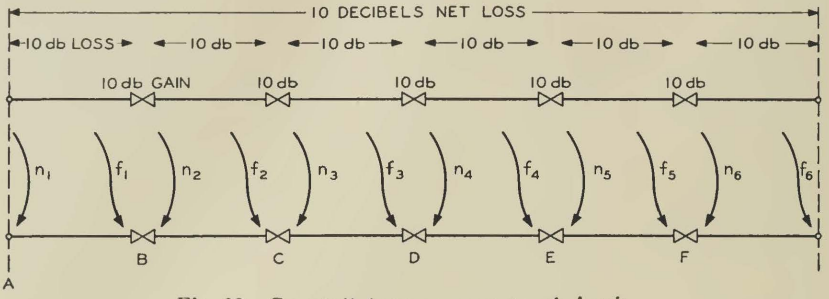


Fig. 29—Crosstalk between repeatered circuits.

Letting r_1 equal the r.m.s. value of n_1, etc., and using probability theory we may write:

$$R^2 = r_1^2 + r_2^2 + \cdots + r_6^2,$$

where R is the r.m.s. of all possible values of the near-end crosstalk at A. If $r_1 = r_2$, etc.

$$R = r\sqrt{6}.$$

The chance of the overall crosstalk deviating from R by any specified amount may be estimated by probability methods. It will be noted that the crosstalk in six repeater sections tends to be more severe than that in one section by $\sqrt{6}$ or, in other words, that the crosstalk varies as the square root of the length. If the use of repeaters were avoided by using more copper, for the same overall loss the crosstalk would be practically the same as with the repeatered circuits. *With the arrangement of repeaters shown* it is not the use of repeaters which causes the increase in crosstalk but rather the increase in circuit length without corresponding increase in circuit loss. For a given circuit length, circuit loss and wire size, other arrangements of repeaters may cause greater or less crosstalk.

If the repeaters of Fig. 29 are spaced farther apart, say 15 db instead of 10, there will be three line repeaters of 15 db gain each and terminal repeaters will be necessary to supply a terminal gain of 5 db in order to obtain a net loss of 10 db. The near-end crosstalk would be reduced by about $\sqrt{4} \div \sqrt{6}$ or 1.8 db because there are only four repeater sections but the terminal repeaters would amplify the near-end crosstalk by 5 db. The net increase would be 3.2 db. From the standpoint of near-end crosstalk, it is thus seen that close spacing between repeaters is very desirable.

In Fig. 29 the output-to-output far-end crosstalk in each repeater section is indicated by f_1 to f_6. The transmission path through any one of these crosstalk couplings is (for like circuits) a loss 10 db greater than the value of the coupling expressed as a db loss. With the repeater arrangement of the figure, the far-end crosstalk paths are attenuated by 10 db while the near-end crosstalk paths are not attenuated. Furthermore, the far-end crosstalk paths ordinarily introduce greater losses than the near-end paths. With greater spacing between repeaters, the near-end crosstalk is amplified but the far-end crosstalk (for like circuits) is still attenuated by the net loss of the circuits. At a given frequency the near-end crosstalk between such "two-wire" circuits is, therefore, much greater than the far-end crosstalk.

Review

Evidently the problem of keeping crosstalk between open-wire circuits within tolerable bounds is by no means a simple one. As we have seen, the work begins with consideration of complete circuits (telephone, program transmission or carrier telegraph) which may be hundreds or even thousands of miles long. The total crosstalk allowance for such long circuits must first be broken down into allowances for the various sections of line between repeaters and then into allowances for the individual transposition section, these individual sections ranging from less than 1/4 to about 6 miles in length.

Then bearing in mind that irregularities in pole spacing and in wire configuration set limits to crosstalk reduction which it is not practicable to overcome by transpositions, the crosstalk designer determines by computation whether, when considering these irregularity effects alone, the crosstalk requirements for the individual transposition sections can be met. If these requirements can not be met he must either have the general circuit layout altered so that, for example, the repeater gains will be more favorably disposed from the standpoint of crosstalk, or he must alter the pole head configuration so that the electrical separation between the circuits will be increased.

Having obtained an overall circuit layout and a configuration of the wires which makes it possible to attain the desired overall crosstalk results, the design of the transpositions proper is undertaken. In this work the transposition designer makes every effort to keep the number of transpositions at a minimum. He does this partly to save money but more particularly because he recognizes that more than enough transpositions do harm rather than good by increasing the number of pole spacing irregularities.

In dealing with the problems of crosstalk coupling between open-wire circuits, consideration must be given not only to the direct effect of one circuit on another but also to the indirect effect of the other circuits on the line. What happens is that the disturbing circuit crosstalks not only directly into the circuit under consideration but also into the group of other circuits and thence into the disturbed circuit. The name "tertiary" circuit has been given to this group of circuits although it is not in reality one circuit but rather any or all of the possible circuits which may be formed of the different wires. The system of transpositions must, therefore, not only substantially balance out the direct couplings between disturbing and disturbed circuits but must also substantially balance the couplings from the disturbing circuit into the "tertiary" circuit and from this "tertiary" circuit into the disturbed circuit.

Reflections of the electrical waves also add interest and complexity to the problem. Such reflections tend to increase crosstalk because the electrical waves which are changed in direction as a result of reflections crosstalk differently, and in many cases more severely, into neighboring circuits than do the waves traveling in the normal direction. The most important reflections occur at junctions between lines and office apparatus. The possibility of other reflections must also be considered, however, at intermediate points in the line which might be caused by inserted lengths of cable, change in spacing of wires, etc.

In working out the transposition designs, the fact that crosstalk between two paralleling circuits tends to manifest itself at both ends is of great importance. At the "near end" crosstalk coming from the disturbed circuit in a direction opposite to the transmission in the disturbing circuit must be considered. At the "far end" crosstalk coming in the same direction as the transmission in the disturbing circuit must be considered.

For telephone circuits which use the same path for transmission in both directions, the "near-end" crosstalk is considerably more severe than the "far-end" for two reasons: (1) The crosstalk per unit length of the paralleling circuits is greater; (2) the gains of the repeaters especially augment the "near-end" crosstalk. Voice-frequency open-wire telephone circuits have always been worked on this "one-path" basis and are good examples of circuits in which "near-end" crosstalk is controlling and must be given principal consideration in working out transposition designs.

In the case of carrier circuits, it was found early in the development that if these circuits were worked on a one-path basis, the crosstalk

would be prohibitively great. Consequently, carrier circuits are now designed to operate on a two-path basis. Two separate bands of frequencies are set aside, each being restricted, by means of one-way amplifiers and electrical filters, to transmission in one direction only. Each telephone circuit is then made up of two oppositely directed channels, one in each frequency band. Thus, direct "near-end" crosstalk is kept from passing to the telephone subscribers. Consequently, the "near-end" type of crosstalk needs to be considered only with respect to that portion which arises from electrical waves reflected at discontinuities in the circuits, which effects have already been mentioned.

In practice a pole line may have some of the pairs very frequently transposed to make them suitable for carrier frequency operation and other pairs less frequently transposed and suitable only for voice-frequency operation. A system of transpositions must permit any arrangement of the two types of pairs which may be found economical for a given line and layout of circuits. Each pair must meet limits for near-end and far-end crosstalk to any other pair which may crosstalk into it in its frequency range. Pairs used only for voice frequencies are usually phantomed and transpositions must, of course, be designed for the phantom circuits as well as the side circuits. The design of a transposition system is, therefore, extraordinarily complicated and tedious and, to paraphrase the Gilbert and Sullivan policeman, "A transposer's lot is not a happy one."

Bibliography

The published material on the matter of open-wire crosstalk and transposition design appears to be very limited. The following papers are of interest:

The Design of Transpositions for Parallel Power and Telephone Circuits, H. S. Osborne. *Trans. of A. I. E. E.*, Vol. XXXVII, Part II, 1918.

Telephone Circuits with Zero Mutual Induction, Wm. W. Crawford, *Trans. of A. I. E. E.*, Vol. XXXVIII, Part I, 1919.

Measurement of Direct Capacities, G. A. Campbell. *Bell System Technical Journal*, July, 1922.

Propagation of Periodic Currents Over a System of Parallel Wires, John R. Carson and Ray S. Hoyt. *Bell System Technical Journal*, July, 1927.

On Crosstalk Between Telephone Lines, M. Vos. L. M. Ericsson Review, English Edition, 1930, Vol. 7.

Application of High Frequencies to Telephone Lines, M. K. Kupfmuller. Presented Before International Electrical Congress, July, 1932.

Probability Theory and Telephone Transmission Engineering, Ray S. Hoyt. *Bell System Technical Journal*, Jan., 1933.

APPENDIX

Calculation of Crosstalk Coefficients

This appendix will first cover methods of calculating the coefficients of transverse crosstalk coupling. It is necessary to calculate both near-end and far-end crosstalk coefficients which involve both direct and indirect components of transverse crosstalk coupling. Coefficients for the direct and for the indirect components will be derived separately and then combined to obtain the total coefficients.

Ordinarily, the indirect effect cannot be readily computed with good accuracy and the total coefficients are usually measured. As previously noted, the method of computing the indirect effect can be used with fair accuracy, however, and it is useful in cases where measurements are impracticable.

The crosstalk between frequently transposed circuits may be calculated with the aid of the above coefficients of transverse coupling and in addition an "interaction crosstalk coefficient" relating to interaction crosstalk coupling of the most important type. The relation of this interaction coefficient to the far-end coefficient of transverse coupling is also discussed herein.

Direct Crosstalk Coefficients

Figure 30 indicates the definitions of the direct crosstalk coefficients used in computing the direct component of the transverse crosstalk coupling. This figure shows a thin transverse slice in a parallel between two long circuits a and b, the thickness of the slice being the

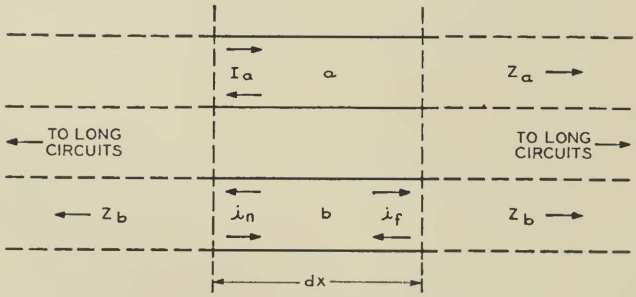


Fig. 30—Crosstalk in a single infinitesimal length.

infinitesimal length dx. Circuit a is energized from the left, the current entering dx being I_a. Propagation of I_a through dx results in near-end and far-end currents i_n and i_f in circuit b at the ends of dx. Since the coefficients are the crosstalk per mile per kilocycle, the near-end coefficient N and the far-end coefficient F may be expressed

as follows:

$$N = \text{limit of } \frac{i_n}{I_a} \cdot \frac{10^6}{Kdx} \text{ as } dx \text{ approaches zero.}$$

$$F = \text{limit of } \frac{i_f}{I_a} \cdot \frac{10^6}{Kdx} \text{ as } dx \text{ approaches zero.}$$

where K is the frequency in kilocycles. For circuits of different characteristic impedances Z_a and Z_b the above current ratios should be multiplied by the square root of the ratio of the real parts of Z_b and Z_a. This correction is not included in the expressions for N and F derived below.

Figure 31 indicates the equivalent electromotive forces which, if impressed on the disturbed circuit b, would cause the same direct crosstalk currents as the electric and magnetic fields of the disturbing circuit. The series and shunt electromotive forces V_m and V_e corre-

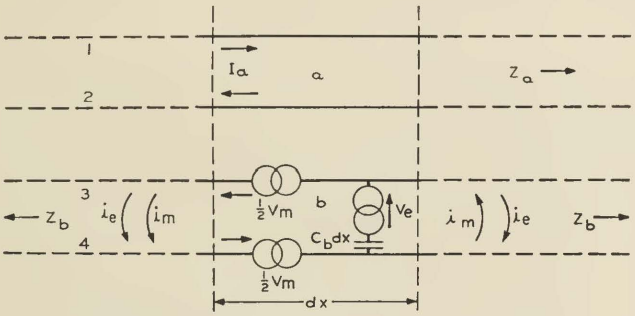


Fig. 31—Equivalent e.m.f.'s in a disturbed circuit.

spond to the magnetic and electric components of the field and cause crosstalk current i_m and i_e. These currents are about equal in magnitude and they add almost directly at the near end of the length dx and subtract almost directly at the far end. The near-end coefficient is, therefore, inherently much greater than the far-end coefficient.

To calculate i_e, the crosstalk current due to the electric field of circuit a, it is necessary to know the shunt voltage V_e. This depends on the charges on the wires of circuit a in the length dx. These charges are due to a voltage V impressed on the left-hand end of circuit a which may be remote from the length dx. Since it is desired to transmit on the metallic circuit a and not on the circuit composed of its wires with ground return, care is taken to "balance" the impressed voltage, i.e., this sending circuit has equal and opposite voltages between its two sides and ground with circuit a disconnected.

The impressed voltage V is propagated to the left-hand end of dx. Letting V_a be the voltage across circuit a at this point, it will be shown that V_a would be balanced except for the effect of interaction crosstalk which is excluded from consideration for the present. Designating the wires of circuit a as 1 and 2, the balanced voltage V_a causes charges Q_1 and Q_2 per unit length on these wires in the length dx. These charges are affected by the presence of other wires in the length dx and they are usually unbalanced. There will be equal and opposite or balanced charges $\pm \frac{Q_1 - Q_2}{2}$ on each wire and unbalanced equal charges $\frac{Q_1 + Q_2}{2}$ on each wire. Since the direct crosstalk is defined as the effect of balanced charges and currents, only the balanced charges should be considered in computing V_e. Letting $Q_a = \frac{Q_1 - Q_2}{2}$ or the balanced charge on wire one per unit length, then:

$$Q_a = V_a C_a = I_a Z_a C_a,$$

where C_a is equal to the "transmission capacitance" per mile, i.e., the capacitance used in calculating α_a the attenuation constant and Z_a the characteristic impedance of circuit a.

The above expression for Q_a includes the reaction of charges in the disturbed circuit. This reaction should not theoretically, be included at this time, since, for convenience in calculation, the disturbed circuit is assumed to have the impressed voltages V_m and V_e but no crosstalk currents or charges as yet. The effect on Q_a of charges in the disturbed circuit, is, however, usually small compared with the effect of charges in various tertiary circuits.

Designating the conductors of circuit b as 3 and 4, V_e is the difference of the potentials of the electric field at 3 and 4 caused by the balanced charges per unit length on 1 and 2. Therefore:

$$V_e = V_3 - V_4 = (Q_a p_{13} - Q_a p_{23}) - (Q_a p_{14} - Q_a p_{24}),$$

where p_{13}, etc., are the potential coefficients.

For c.g.s. elst. units, $p_{13} = 2 \log \frac{s_{13}}{r_{13}}$ where s_{13} and r_{13} are the distances indicated by Fig. 32. Therefore:

$$V_e = V_a C_a (p_{13} - p_{23} - p_{14} + p_{24}) = V_a C_a p_{ab}.$$

The capacitance C_a may be obtained from measurements on a short length of a multi-wire line. Its value is, however, only a few per cent

greater than C_a' the value for a single pair line (without capacitance at the insulators). For a single pair having like wires in a horizontal

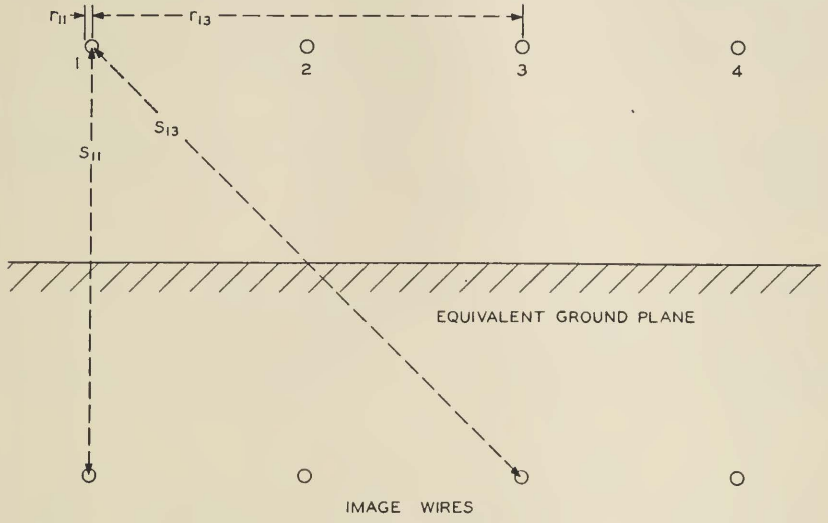


Fig. 32—Distances used in computing potential coefficients.

plane, C_a' is readily calculated as follows:

$$C_a' = \frac{1}{2(p_{11} - p_{12})},$$

where p_{11} in c.g.s. elst. units is:

$$2 \log_e \frac{s_{11}}{r_{11}}.$$

The distances s_{11} and r_{11} are indicated on Fig. 32.

The expression for V_e may be written:

$$V_e = V_a C_a p_{ab} = V_a C_a' p_{ab} \frac{C_a}{C_a'} = V_a T_{ab} \frac{C_a}{C_a'}.$$

The coefficient T_{ab} is called the "voltage transfer coefficient." It is readily computed since it is a function of potential coefficients and it is independent of the system of units used in computation. Since C_a is about equal to C_a', T_{ab} is about equal to the ratio of V_e to V_a.

The shunt voltage V_e drives a current through the shunt admittance of circuit b in the infinitesimal length dx of Fig. 31. This shunt admittance is $(G_b + j\omega C_b)dx$ which is very nearly equal to $j\omega C_b dx$ where C_b is the transmission capacitance per mile of circuit b and

$\omega = 2\pi f$ where f is the frequency in cycles per second. This current divides equally between the two ends of circuit b. The near-end current is:

$$i_e = -\frac{1}{2}\frac{V_e}{\frac{1}{j\omega C_b dx} + \frac{Z_b}{2}}.$$

The near-end direct crosstalk coefficient due to the electric field of circuit a may be called N_e and is the limiting value of the following expression as dx approaches zero:

$$\frac{i_e}{I_a}\cdot\frac{10^6}{Kdx} = -\frac{1}{2KI_a}\cdot\frac{V_e 10^6}{\frac{1}{j\omega C_b} + \frac{Z_b}{2}dx},$$

where K is the frequency in kilocycles. The near-end direct crosstalk coefficient for the electric field is, therefore:

$$(1)\quad N_e = -\frac{V_e 10^6 j\omega C_b}{2KI_a} = -\frac{Z_a T_{ab} j\omega C_b 10^6}{2K}\cdot\frac{C_a}{C_a'}$$

$$= -j\pi Z_a T_{ab} C_b 10^9 \frac{C_a}{C_a'}.$$

The far-end current due to V_e of Fig. 31 is $-i_e$ and, therefore, the far-end coefficient due to the electric field is $-N_e$.

The near-end and far-end crosstalk currents of Fig. 31 due to the magnetic field are alike and are designated i_m which may be calculated as follows:

$$i_m = \frac{V_m}{2Z_b} = -\frac{I_a j\omega M_{ab} dx}{2Z_b}.$$

The near-end or far-end crosstalk coefficients for the magnetic field may be called N_m and F_m. They are alike and equal to the limit of:

$$\frac{i_m}{I_a}\cdot\frac{10^6}{Kdx} \text{ as } dx \text{ approaches zero.}$$

Therefore:

$$N_m = F_m = -\frac{j\omega M_{ab}}{2Z_b K}10^6 = -\frac{j\pi M_{ab}}{Z_b}10^9.$$

In the above, M_{ab} is the mutual inductance per unit length between circuits a and b. It is calculated in the same manner as p_{ab} used in computing V_e. These methods of computing V_e and V_m from the distances r_{13}, r_{14}, s_{13}, etc., of Fig. 32 are not precise but are sufficiently accurate for open-wire circuits since the diameters of the wires are

small compared with their interaxial distances. The "image" wires of Fig. 32 should, theoretically, be located farther below the equivalent ground plane for calculations of mutual inductance. This alters s_{14}, etc. Since the distances between wires are small compared to those between wires and images, the values of s are all about equal and have practically no effect on the value of p_{ab}.

Therefore, M_{ab} in c.g.s. elmg. units may be assumed numerically equal to p_{ab} or:

$$M_{ab} = p_{ab} = \frac{T_{ab}}{C_a'}.$$

In c.g.s. elst. units $C_a' = \dfrac{1}{2(p_{11} - p_{12})}$ which is also the expression for $1/L_a'$ in c.g.s. elmg. units where L_a' is the *external* inductance of circuit a, i.e., the inductance due to the magnetic field external to the wires of circuit a. Therefore:

$$M_{ab} = T_{ab}L_a',$$

where M_{ab} and L_a' may be expressed in any system of units.

The near-end or far-end direct coefficient for the magnetic field may, therefore, be written:

$$N_m = F_m = -\frac{j\pi T_{ab}L_a'}{Z_b}10^9. \tag{2}$$

The above expression is almost equal to N_e, the near-end coefficient for the electric field.

It may be written:

$$N_m = N_e\left[\frac{j\omega L_a' j\omega C_a'}{Z_a j\omega C_a Z_b j\omega C_b}\right].$$

Now $Z_a j\omega C_a$ is very nearly equal to $Z_a(G_a + j\omega C_a)$ which is γ_a. Likewise $Z_b j\omega C_b$ is very nearly equal to γ_b. If the circuits had no resistance or leakance the propagation constant would be $\gamma_0 = j\omega\sqrt{L_a'C_a'}$ or $j\omega/v$ where v is the speed of light in miles per second. Therefore:

$$N_m = \frac{N_e\gamma_0^2}{\gamma_a\gamma_b} \text{ very nearly.}$$

The total direct crosstalk coefficients are:

$$N_d = N_m + N_e = N_e\left(1 + \frac{\gamma_0^2}{\gamma_a\gamma_b}\right) = 2N_e \text{ approx.} \tag{3}$$

At carrier frequencies the ratio of γ_0 to γ_a (or γ_b) is about equal to

the ratio of the actual speed of propagation to the speed of light, i.e., 180,000 to 186,000 or about .97. Therefore $\left(1 + \frac{\gamma_0^2}{\gamma_a\gamma_b}\right)$ is about 1.94.

$$(4) \quad F_d = N_m - N_e = N_e\left(\frac{\gamma_0^2}{\gamma_a\gamma_b} - 1\right)$$

$$= N_e\left(-.06 + j\frac{90}{\pi}\frac{\alpha_a + \alpha_b}{K}\right) \text{approx.}$$

The attenuation of the disturbing and disturbed circuits may not be neglected in evaluating the expression $\left(\frac{\gamma_0^2}{\gamma_a\gamma_b} - 1\right)$.

The expression given for N_e in equation (1) above may be written:

$$N_e = -\frac{\gamma_a T_{ab} 10^6}{2K}\frac{C_a}{C_a'}\frac{C_b}{C_a}.$$

This assumes $Z_a j\omega C_a$ equals γ_a. At carrier frequencies γ_a is about equal to $j\beta_a$ which is about $j\pi K/90$ since the speed of propagation is about 180,000 miles per second. The expression for N_e may, therefore, be written in the following simple approximate form:

$$N_e = -j\frac{\pi T_{ab} 10^6}{180}\frac{C_a}{C_a'}\cdot\frac{C_b}{C_a}.$$

The ratio of C_a to C_a' does not ordinarily exceed 1.02. For like circuits, therefore:

$$N_e = \frac{-j\pi T_{ab} 10^6}{180} \text{ approx.}$$

On Fig. 33 the magnitudes of N_d and F_d are plotted against frequency for 8-inch spaced conductors .128-inch in diameter. Both F_d and N_d are divided by T_{ab} to make the curves applicable to any circuit combination. These curves show that N_d is practically independent of frequency (above a few hundred cycles) but F_d decreases rapidly with frequency for several thousand cycles.

Indirect Crosstalk Coefficients

Expressions for the indirect crosstalk coefficients used in computing the indirect component of transverse crosstalk coupling will now be derived. The derivation first covers the case of a single representative tertiary circuit. Fig. 34 shows a thin transverse slice of the parallel of the three circuits, the thickness of the slice being the infinitesimal length dx. The only tertiary circuit to be considered for the present is the metallic circuit composed of wires 5 and 6 and designated as c. There are other possible tertiary circuits in the system of 6 wires,

for example, the phantom circuit composed of wires 1 and 2 as one side and 5 and 6 as the other side. The method of estimating the total effect of all possible tertiary circuits will be discussed later.

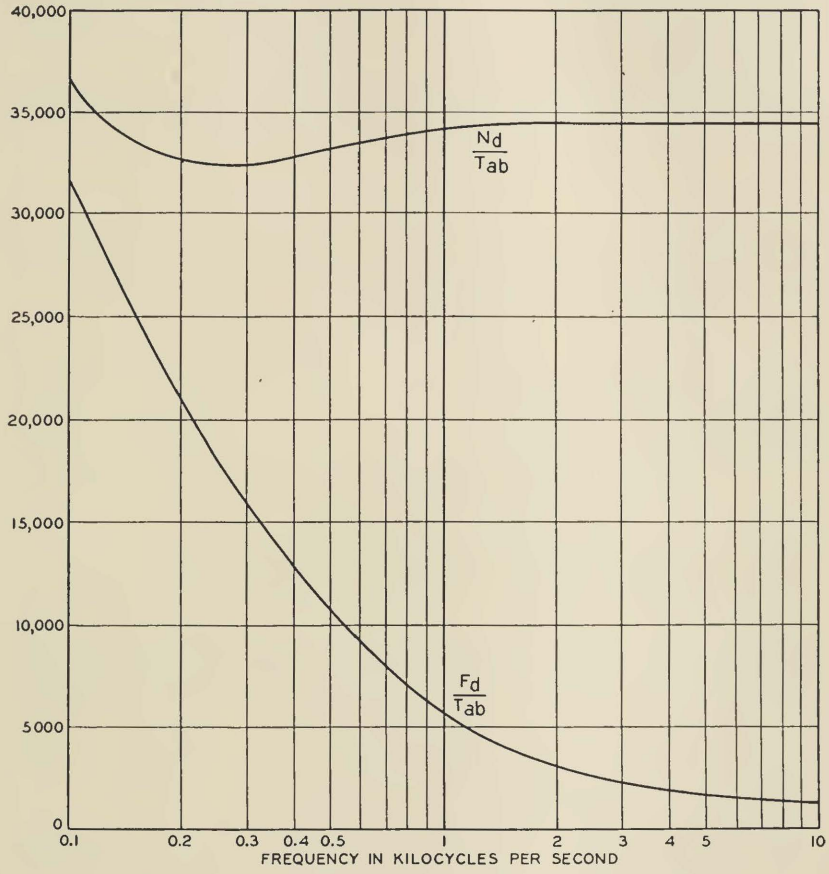


Fig. 33—Variation of direct crosstalk coefficients with frequency.

The immediate problem is to compute the crosstalk currents in circuit *b* at the ends of the length *dx* due to currents and charges in circuit *c* in this length and caused by transmission over circuit *a* through *dx*.

The crosstalk currents in circuit *b* due to currents and charges in circuit *a* were computed by determining the equivalent series and shunt e.m.f.'s in circuit *b*. The effect of currents and charges in circuit *c* on crosstalk currents in circuit *b* may be computed in a similar manner. The series e.m.f. in circuit *b* proportional to the current in

circuit c will, however, be negligible compared with the series e.m.f. proportional to current in circuit a. This is evident since the current in circuit c is a crosstalk current which approaches zero as dx approaches zero while the current in circuit a does not vary with dx.

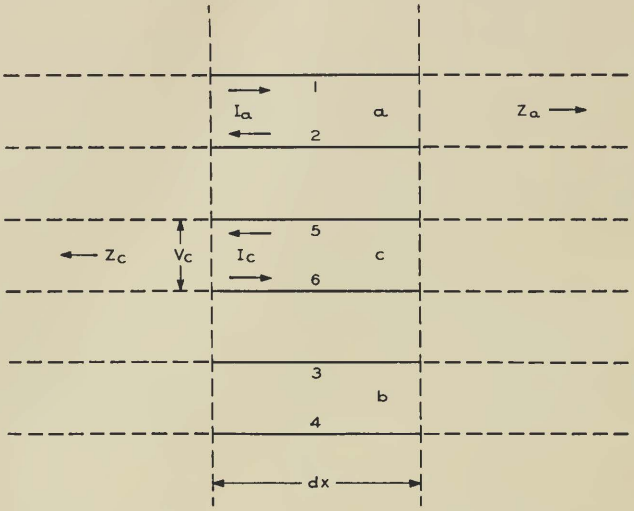


Fig. 34—Schematic used in deriving formulas for indirect crosstalk coefficients.

The shunt e.m.f. in circuit b dependent on the charges of circuit c is not, however, negligible compared with the shunt e.m.f. in circuit b due to charges in circuit a since the charges in both a and c approach zero as dx decreases. In other words the magnetic field of circuit c may be neglected but the electric field must be considered. (Both fields must be considered in computing interaction crosstalk.)

To determine the equivalent shunt e.m.f. in circuit b which depends upon the electric field of circuit c the voltage between the wires of circuit c must be determined. If circuit c did not exist, the electric field of circuit a would cause a difference of potential between the points actually occupied by wires 5 and 6 at the left-hand end of dx in Fig. 34. This difference of potential would be:

$$V_{ac} = V_a C_a' p_{ac} = V_a T_{ac}.$$

With circuit c present, this difference of potential is changed to V_c, the actual voltage across circuit c. The voltage could not change from V_{ac} to V_c without charges on circuit c and the charge per wire per unit length is proportional to the change in voltage from V_{ac} to V_c which may be designated U_c. The equivalent shunt e.m.f. in

circuit b due to the presence of charges in circuit c is, therefore, proportional to U_c.

By definition:

$$V_{ac} + U_c = V_c \quad \text{or} \quad U_c = V_c - V_{ac}.$$

Since the crosstalk current in circuit c approaches zero as dx approaches zero, V_c must also approach zero and U_c approaches $-V_{ac}$. The shunt e.m.f. in circuit b due to charges on circuit a was computed as:

$$V_e = V_a T_{ab} \frac{C_a}{C_a'}.$$

To allow for the *electric* field of circuit c, V_e must be augmented by:

$$V_e' = U_c T_{cb} \frac{C_c}{C_c'} = -V_{ac} T_{cb} \frac{C_c}{C_c'} = -V_a T_{ac} T_{cb} \frac{C_c}{C_c'}.$$

Since the part of the direct near-end crosstalk coefficient resulting from V_e was found to be $N_e = -j\pi Z_a T_{ab} C_b 10^9 \frac{C_a}{C_a'}$, by proportion the indirect near-end coefficient resulting from V_e' will be:

$$N_i = j\pi Z_a T_{ac} T_{cb} C_b 10^9 \frac{C_c}{C_c'} = \frac{j\pi T_{ac} T_{cb} 10^6}{180} \text{ approx.} \tag{5}$$

Since the far-end crosstalk current resulting from a shunt voltage in circuit b is opposite in sign to the near-end current, the indirect far-end coefficient will be:

$$F_i = -N_i. \tag{6}$$

Total Crosstalk Coefficients

The total near-end and far-end crosstalk coefficients used in computing transverse crosstalk coupling will be the sum of the direct and indirect coefficients or:

$$N = N_d + N_i. \tag{7}$$

$$F = F_d + F_i = F_d - N_i. \tag{8}$$

The expressions for F_i and N_i are about independent of frequency in the carrier-frequency range because Z_a does not depend much on frequency above a few thousand cycles, C_b is about independent of frequency and $T_{ac}T_{cb}$ depends only on the cross-sectional dimensions of the wire configuration.

Since, as indicated by Fig. 33, N_d is usually about independent of

frequency and since $N = N_d + N_i$ is largely determined by N_d, the near-end coefficient N is about independent of frequency above a few hundred cycles. The far-end coefficient F is about independent of frequency above a few thousand cycles where it is largely determined by F_i.

The preceding discussion of indirect crosstalk coefficients covered only the effect of charges in the single metallic tertiary circuit c of Fig. 34. The indirect coefficient in a practical case may be estimated with fair accuracy by considering all the more important tertiary circuits in a similar manner. It was shown that the final voltage of tertiary circuit c was zero. Similarly, the final voltage of each tertiary circuit is zero. This includes any tertiary circuits involving the two wires of the disturbing circuit in multiple. The average voltage of the two wires of the disturbing circuit is zero and the voltage across the disturbing circuit is balanced. As previously stated, this voltage does not become unbalanced as a result of transverse crosstalk in any infinitesimal length but it may become unbalanced due to interaction crosstalk.

The charges per unit length on the various tertiary circuits are the same as those which would be caused by impressing a system of voltages equal and opposite to those induced by the balanced charges per unit length which would be on the two wires of the disturbing circuit if this circuit were the only pair on the line. Assuming such a system of impressed voltages, it is not practicable to accurately compute the charges in any tertiary circuit since this depends on the voltages impressed on all the tertiary circuits and the couplings between the various tertiary circuits. Advantage may be taken, however, of the fact that the charge on a tertiary circuit will depend mostly on the voltage impressed on that circuit provided it is not heavily coupled with other circuits.

It is possible to divide the various voltages impressed on the tertiary circuits into components such that (1) equal voltages are impressed on wires of a "ghost" circuit composed of all the wires on the line with ground return, (2) balanced voltages are impressed on each pair used for transmission purposes (except the disturbed and disturbing circuits) and (3) balanced voltages are impressed on each possible phantom of two pairs used for transmission purposes.

Such a system of impressed voltages and tertiary circuits is convenient for computation since the charge on any tertiary circuit largely depends on the voltage impressed on that circuit. If accurate calculations of the charges were practicable, a simpler system of tertiary circuits could be used to obtain the same final result, i.e.,

single-wire tertiary circuits with ground return could be used. Computation with such tertiary circuits is impracticable because of the large coupling between them.

In the elaborate system of tertiary circuits described above, the ghost circuit may be neglected. The voltage impressed across this circuit is the average of all the voltages impressed on the various wires. These voltages may be plus or minus and the average tends to be small. Also, the charge per pair per impressed volt is usually much less for the ghost circuit than for a phantom circuit due to the relatively small capacitance between a pair and ground as compared with that between two pairs.

The pairs used for transmission purposes may usually be disregarded, also, since their coupling with the disturbing and disturbed circuits is much smaller than that of the phantom tertiary circuits.

The practical method of computing the indirect crosstalk coefficient is, therefore, to consider as tertiary circuits a considerable number of phantoms composed of pairs used for transmission purposes including the disturbed and disturbing pairs. In calculating the charge in any tertiary circuit, the voltages impressed on other tertiary circuits are disregarded.

In calculating the effect of a single tertiary circuit c, the expression for the indirect coefficient contained the factor $T_{ac}T_{cb}$. To estimate the effect of all the tertiary circuits, this factor should be replaced by:

$$\frac{2}{m}\sum T_{ap}T_{pb}.$$

This expression assumes that there are n pairs on the line and that $m - 2$ of these pairs are close enough to the disturbing and disturbed pairs to appreciably affect the indirect crosstalk between them. The subscript p indicates any phantom of the m pairs including the disturbed and disturbing pairs. The summation is for all possible phantoms each consisting of two of the m pairs. If the voltages induced by the balanced charges Q_a' of pair a are V_r and V_s for the two sides of a phantom, the balanced voltage assumed to be impressed across the phantom is $\frac{2}{m}(V_s - V_r)$. Other parts of V_r and V_s are used in the "ghost" voltage and in balanced voltages across other phantoms.

T_{ap} and T_{pb} are voltage transfer coefficients relating balanced impressed voltage on the disturbing circuit to induced voltage on the disturbed circuit. T_{ap} involves C_a' the transmission capacitance of circuit a on a single pair line. T_{pb} involves the transmission capaci-

tance of a particular phantom on a line having only that phantom present. This capacitance is the ratio of balanced charge (on each side of the phantom) to the balanced impressed voltage. The phantom capacitance may be readily estimated from the potential coefficients. For example, if the phantom involves pairs 1–2 and 5–6 the phantom capacitance is very nearly:

$$C_p = \frac{4}{p_{11} + p_{22} + p_{55} + p_{66} + 2p_{12} + 2p_{56} - 2p_{15} - 2p_{25} - 2p_{16} - 2p_{26}}.$$

If the disturbing circuit is pair 1–2 and the disturbed circuit is pair 3–4:

$$T_{ap} = \frac{p_{15} + p_{16} - p_{25} - p_{26}}{2(p_{11} - p_{12})}$$

$$T_{pb} = \frac{C_p}{2}(p_{13} + p_{23} + p_{45} + p_{46} - p_{14} - p_{24} - p_{35} - p_{36}).$$

These computations of indirect coefficients are necessarily laborious. They can be simplified to some extent by ignoring phantoms for which either T_{ap} or T_{pb} is zero or small. For example, the voltage transfer coefficient is zero for pair 1–2 to such phantoms as 1–2 and 11–12, 11–12 and 21–22, etc.

In the following table are given comparisons of far-end crosstalk coefficients as measured in a 40-wire line and as computed by the methods discussed above. The spacing of the various wires and crossarms is indicated by Fig. 27A. The measured values are for 40 wires and the computed values are for 10, 20 and 30 wires. It will be seen that a considerable number of wires must be taken into account in the computations in order to obtain a fair check with the coefficient measured for a heavy line.

Far-End Crosstalk Per Mile Per Kilocycle

	Combination		
	1–2 to 3–4	1–2 to 11–12	1–2 to 9–10
Computed for 10 wires	45	28	
Computed for 20 wires	63	47	11
Computed for 30 wires	69	58	22
Measured for 40 wires	74	70	21

Interaction Crosstalk Coefficient

It was assumed in the discussion of crosstalk coefficients that the "interaction crosstalk coefficient" $N_{ac}N_{cb}10^{-6}$ was nearly equal to $-2F_i\gamma_c/K$. This relation is deduced below, for a representative tertiary circuit c, from the expressions for F_i and N_d given by equations (3) and (6) above. N_{ac} may be obtained by using the expressions for N_e and N_d given by equations (1) and (3) above. In these equations, subscript c should be substituted for subscript b. The expression for N_{ac} becomes:

$$N_{ac} = -j\pi Z_a T_{ac} C_c 10^9 \frac{C_a}{C_a'}\left[1 + \frac{\gamma_0^2}{\gamma_a\gamma_c}\right].$$

Deriving a similar approximate expression for N_{cb}:

$$N_{ac}N_{cb}10^{-6} = -\frac{F_i}{2K}\gamma_c\frac{C_a}{C_a'}\left[1 + \frac{\gamma_0^2}{\gamma_a\gamma_c}\right]\left[1 + \frac{\gamma_0^2}{\gamma_c\gamma_b}\right].$$

This assumes $Z_c j\omega C_c = Z_c(G_c + j\pi C_c)$ which is γ_c.

Crosstalk measurements indicate that the ratio of γ_0 to γ_a or γ_b or γ_c is about .97 at carrier frequencies and C_a/C_a' about 1.02. Therefore:

$$N_{ac}N_{cb}10^{-6} = -\frac{F_i\gamma_c}{2K}1.02(1.94)^2 = -2F_i\gamma_c/K \text{ approx.}$$

The above discussion covers the case of a single tertiary circuit c. In the practical case the known crosstalk coefficient F_i includes the effect of a large number of tertiary circuits which have various values of γ_c. Obtaining the interaction crosstalk coefficient from the expression $-2F_i\gamma_c/K$ involves assuming a representative value of γ_c. At carrier frequencies γ_c is about equal to $j\beta_c$ which for all important tertiary circuits is in the neighborhood of $j\pi K/90$.

A known value of the crosstalk coefficient F_i expresses the effects of the electric fields of many tertiary circuits in one infinitesimal slice and includes the alteration of the field of any one tertiary circuit due to the presence of the others. The interaction crosstalk coefficient involves consideration of both electric and magnetic fields. In any one slice, the electric field of a tertiary circuit determines its crosstalk into the disturbed circuit but the current in the tertiary circuit at the end of the slice is determined by both the electric and magnetic fields of the disturbing circuit. The tertiary circuit current is transmitted into another slice and sets up electric and magnetic fields which both contribute to the crosstalk current in the disturbed circuit.

In the case of a single tertiary circuit, the electric and magnetic effects expressed by the interaction coefficient are simply related and the interaction coefficient is simply related to F_i. With many tertiary circuits the relation between the electric and magnetic fields of any one tertiary circuit is altered by the presence of the others and the relative importance of electric and magnetic fields in the interaction effect varies among the tertiary circuits (except for the ideal case of "non-dissipative" circuits). In a practical case, therefore, the interaction coefficient is only approximately proportional to F_i. Measurements of the interaction coefficients by indirect methods have, however, indicated that the approximation is satisfactory for purposes of practical transposition design.

Symposium on Wire Transmission of Symphonic Music and Its Reproduction in Auditory Perspective

Basic Requirements*

By HARVEY FLETCHER

The fundamental requirements involved in a system capable of picking up orchestral music, transmitting it a long distance, and reproducing it in a large hall, are discussed in this paper.

IN THIS electrical era one is not surprised to hear that orchestral music can be picked up in one city, transmitted a long distance, and reproduced in another. Indeed, most people think such things are commonplace. They are heard every night on the radio. However, anyone who appreciates good music would not admit that listening even to the best radio gives the emotional thrill experienced in the concert hall. Nor is it evident that a listener in a small room ever will be able to get the same effect as that experienced in a large hall, although it must be admitted that such a question is debatable. The proper answer will involve more than a consideration of only the physical factors.

This symposium describes principles and apparatus involved in the reproduction of music in large halls, the reproduction being of a character that may give even greater emotional thrills to music lovers than those experienced from the original music. This statement is based upon the testimony of those who have heard some of the few concerts reproduced by the apparatus which will be described in the papers of this symposium.

It is well known that when an orchestra plays, vibrations which are continually changing in form are produced in the air of the concert hall where the orchestra is located. An ideal transmission and reproducing system may be considered as one that produces a similar set of vibrations in a distant concert hall in which is executed the same time-sequence of changes that takes place in the original hall. Since such changes are different at different positions in the hall, the use of such an ideal system implies that at corresponding positions in the

* First paper in the Symposium. Presented at Winter Convention of A. I. E. E., New York City, Jan. 23–26, 1934. Published in *Electrical Engineering*, January, 1934.

two halls this time-sequence should be the same. Obviously, this never can be true at every position unless the halls are the same size and shape; corresponding positions would not otherwise exist. Let us consider the case where the two halls are the same size and shape and also have the same acoustical properties. Let us designate the first hall in which the music originates by *O*, and the second one in which the music is reproduced by *R*. What requirements are necessary to obtain perfect reproduction from *O* into *R* such that any listener in any part of *R* will receive the same sound effects as if he were in the corresponding position in *O*?

Suppose there were interposed between the orchestra and the audience a flexible curtain of such a nature that it did not interfere with a free passage of the sound, and which at the same time had scattered uniformly over it microphones which would pick up the sound waves and produce a faithful electrical copy of them. Assume each microphone to be connected with a perfect transmission line which terminates in a projector occupying a corresponding position on a similar curtain in hall *R*. By a perfect transmission line is meant one that delivers to the projector electrical energy equal both in form and magnitude to that which it receives from the microphone. If these sound projectors faithfully transform the electrical vibrations into sound vibrations, the audience in hall *R* should obtain the same effect as those listening to the original music in hall *O*.

Theoretically, there should be an infinite number of such ideal sets of microphones and sound projectors, and each one should be infinitesimally small. Practically, however, when the audience is at a considerable distance from the orchestra, as usually is the case, only a few of these sets are needed to give good auditory perspective; that is, to give depth and a sense of extensiveness to the source of the music. The arrangement of some of these simple systems, together with their effect upon listeners in various parts of the hall, is described in the paper by Steinberg and Snow (page 245).

In any practical system it is important to know how near these ideal requirements one must approach before the listener will be aware that there has been any degradation from the ideal. For example, it is well known that whenever a sound is suddenly stopped or started, the frequency band required to transmit the change faithfully is infinitely wide. Theoretically, then, in order to fulfill these ideal requirements for transmitting such sounds, all three elements in the transmission system should transmit all possible frequencies without change. Practically, because of the limitations of hearing, this is not necessary. If the intensities of some of the component frequencies required to represent

such a change are below the threshold of audibility it is obvious that their elimination will not be detected by the average normal ear. Consequently, for highgrade reproduction of sounds it is obvious that, except in very special cases, the range of frequencies that the system must transmit is determined by the range of hearing rather than by the kind of sound that is being reproduced.

Tests have indicated that, for those having normal hearing, pure tones ranging in frequency from 20 to 20,000 cycles per second can be heard. In order to sense the sounds at either of these extreme limits, they must have very high intensity. In music these frequencies usually are at such low intensities that the elimination of frequencies below 40 c.p.s. and those above 15,000 c.p.s. produces no detectable difference in the reproduction of symphonic music. These same tests also indicated that the further elimination of frequencies beyond either of these limits did begin to produce noticeable effects, particularly on a certain class of sounds produced in the orchestra. For example, the elimination of all frequencies above 13,000 c.p.s. produced a detectable change in the reproduced sound of the snare drum, cymbals, and castanets. Also, the elimination of frequencies below 40 c.p.s. produced detectable differences in reproduced music of the base viol, the bass tuba, and particularly of the organ.

Within this range of frequencies the system (the combination of the microphone, transmission line, and loud speaker) should reproduce the various frequencies with the same efficiency. Such a general statement sounds correct, but a careful analysis of it would reveal that when any one tried to build such a system or tried to meet such a requirement he would have great difficulty in understanding what it meant.

For example, for reproducing all the frequencies within this band, a certain system may be said to have a uniform efficiency when it operates between two rooms under the condition that the pressure variation at a certain distance away from the sound projector is the same as the pressure variation at a certain position in front of the microphone. It is obvious, however, that in other positions in the two rooms this relation would not in general hold. Also, if the system were transferred into another pair of rooms the situation would be entirely changed. These difficulties and the way they were met are discussed in the papers of this symposium that deal with loud speakers and microphones (p. 259) and with methods of applying the reproducing system to the concert hall (p. 301). It will be obvious from these papers that the criterion for determining the ideal frequency characteristics to be given to the system is arbitrary within certain limits. However, solving the problem according to criteria adopted produced a system that gave very satisfactory results.

Besides the requirement on frequency response just discussed, the system also must be capable of handling sound powers that vary through a very wide range. If this discussion were limited to the type of symphonic music that now is produced by the large orchestras, this range would be about 10,000,000 to 1, or 70 decibels. To reproduce such music then, the system should be capable of handling the smallest amount of power without introducing extraneous noises approaching it in intensity, and also reproduce the most intense sounds without overloading any part of the transmission system. However, this range is determined by the capacities of the musical instruments now available and the man power that conveniently can be grouped together under one conductor. As soon as a system was built that was capable of handling a much wider range, the musicians immediately took advantage of it to produce certain effects that they previously had tried to obtain with the orchestra alone, but without success because of the limited power of the instruments themselves. For these reasons it seems clear that the desirable requirements for intensity range, as well as those for frequency range, are determined largely by the ear rather than by the physical characteristic of any sound. An ideal transmission should, without introducing an extraneous audible sound, be capable of reproducing a sound as faintly as the ear can hear and as loudly as the ear can tolerate. Such a range has been determined with the average normal ear when using pure tones. The results of recent tests are shown in Fig. 1.

The ordinates are given in decibels above the reference intensity which is 10^{-16} watts per square centimeter. The values are for field intensities existing in an air space free from reflecting walls. The most intense peaks in music come in the range between 200 and 1000 c.p.s. Taking an average for this range it may be seen that there is approximately a 100-db range in intensity for the music, provided about 10 db is allowed for the masking of sound in the concert hall even when the audience is quietest.

The music from the largest orchestra utilizes only 70 db of this range when it plays in a concert hall of usual size. To utilize the full capabilities of the hearing range the ideal transmission system should add about 10 db on the *pp* side and 20 db on the *ff* side of the range. The capacity of the sound projectors necessary to reach the maximum allowable sound that the ear can tolerate varies with the size of the room. A good estimate can be obtained by the following consideration.

If T is the time of reverberation of the hall in seconds, E the power of the sound source in watts, I the maximum energy density per cubic

centimeter in joules, and V the volume of the hall in cubic centimeters, then it is well known that

$$I = \frac{1}{6 \log_e 10} \cdot \frac{ET}{V}. \tag{1}$$

Measurements have shown that when the sound intensity in a free field reaches about 10^{-4} watts per square centimeter, the average person begins to *feel* the sound. This maximum value is approxi-

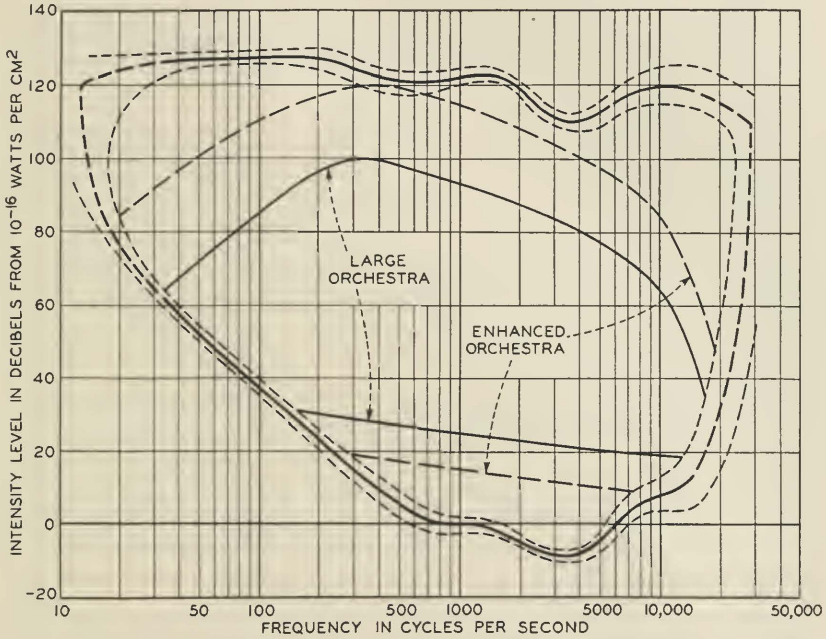


Fig. 1—Limits of audible sound as determined by recent tests.

mately the same for all frequencies in the important audible range. Any higher intensities, and for some persons somewhat lower intensities, become painful and may injure the hearing mechanism. This intensity corresponds to an energy density I of 3×10^{-9} joules. Using this figure as the upper limit to be tolerated by the human ear, then, the maximum power of the sound source must be given by

$$E = 4.1 \times 10^{-8} \frac{V}{T}. \tag{2}$$

For halls like the Academy of Music in Philadelphia and Carnegie Hall in New York City, in which the volume V is approximately 2×10^{10} cubic centimeters and the reverberation time about 2 seconds,

E, the power of the sound source, is approximately 400 watts. For other halls it may be seen that the power required for this source is proportional to the volume of the hall and inversely proportional to the reverberation time. A person would experience the sense of *feeling* when closer than about 10 meters to such a source of 400 watts power, even in free open space. Hence it would be unwise to have seats closer than 10 or 15 meters from the stage when such powers are to be used.

These, then, are the general fundamental requirements for an ideal transmission system. How near they can be realized with apparatus that we now know how to build will be discussed in the papers included in this symposium.

A system approximately fulfilling these requirements was constructed and used to reproduce the music played by the Philadelphia Orchestra. The first public demonstration was given in Constitution Hall, Washington, D. C., on the evening of April 27, 1933, under the auspices of the National Academy of Sciences. At that time, Dr. Stokowski, Director of the Philadelphia Orchestra, manipulated the electric controls from a box in the rear of Constitution Hall while the orchestra, led by Associate Conductor Smallens, played in the Academy of Music in Philadelphia.

Three microphones of the type described in the paper by Wente and Thuras (p. 259) were placed before the orchestra in Philadelphia, one on each side and one in the center at about 20 feet in front of and 10 feet above the first row of instruments in the orchestra. The electrical vibrations generated in each of these microphones were amplied by voltage amplifiers and then fed into a transmission line which was extended to Washington by means of telephone cable. The construction of these lines, the equipment used with them, and their electrical properties, are described in the paper by Affel, Chesnut, and Mills (p. 285). In Constitution Hall at Washington, D. C., these transmission lines were connected to power amplifiers. The type of power amplifiers and voltage amplifiers used are described in the paper by Scriven (p. 278). The output of these amplifiers fed three sets of loud speakers like those described in the paper by Wente and Thuras. They were placed on the stage in Constitution Hall in positions corresponding to the microphones in the Academy of Music, Philadelphia.

Judging from the expression of those who heard this concert, the development of this system has opened many new possibilities for the reproduction and transmission of music that will create even a greater emotional appeal than that obtained when listening to the music coming directly from the orchestra through the air.

Physical Factors*

By J. C. STEINBERG and W. B. SNOW

In considering the physical factors affecting it, auditory perspective is defined in this paper as being reproduction which preserves the spatial relationships of the original sounds. Ideally, this would require an infinite number of separate microphone-to-speaker channels; practically, it is shown that good auditory perspective can be obtained with only 2 or 3 channels.

ABILITY to localize the direction, and to form some judgment of the distance from a sound source under ordinary conditions of listening, are matters of common experience. Because of this faculty an audience, when listening directly to an orchestral production, senses the spatial relations of the instruments of the orchestra. This spatial character of the sounds gives to the music a sense of depth and of extensiveness, and for perfect reproduction should be preserved. In other words, the sounds should be reproduced in true *auditory perspective.*

In the ordinary methods of reproduction, where only a single loud speaking system is used, the spatial character of the original sound is imperfectly preserved. Some of the depth properties of the original sound may be conveyed by such a system,[1] but the directional properties are lost because the audience tends to localize the sound as coming from the direction of a single source, the loud speaker. Ideally, there are two ways of reproducing sounds in true auditory perspective. One is binaural reproduction which aims to reproduce in a distant listener's ears, by means of head receivers, exact copies of the sound vibrations that would exist in his ears if he were listening directly. The other method, which was described in the first paper of this series, uses loud speakers and aims to reproduce in a distant hall an exact copy of the pattern of sound vibration that exists in the original hall. In the limit, an infinite number of microphones and loud speakers of infinitesimal dimensions would be needed.

Far less ideal arrangements, consisting of as few as two microphone-loudspeaker sets, have been found to give good auditory perspective. Hence, it is not necessary to reproduce in the distant hall an exact copy of the vibrations existing in the original hall. What physical

* Second paper in the Symposium on Wire Transmission of Symphonic Music and Its Reproduction in Auditory Perspective. Presented at Winter Convention of A. I. E. E., New York City, Jan. 23–26, 1934. Published in *Electrical Engineering*, January, 1934.

properties of the waves must be preserved then, and how are these properties preserved by various arrangements of 2- and 3-channel loudspeaker reproducing systems? To answer these questions, some very simple localization tests have been made with such systems. Perhaps attention can be focused more easily on their important properties by considering briefly the results of these tests.

Localization Afforded by Multichannel Systems

In Fig. 1 is shown a diagram of the experimental set-up that was used. The microphones, designated as *LM* (left), *CM* (center), and *RM* (right), were set on a "pick-up" stage that was marked out on the floor of an acoustically treated room. The loud speakers, designated as *LS*, *CS*, and *RS*, were placed in the front end of the auditorium at the Bell Telephone Laboratories and were concealed from view by a curtain of theatrical gauze. The average position of a group of twelve observers is indicated by the cross in the rear center part of the auditorium.

The object of the tests was to determine how a caller's position on the pick-up stage compared with his apparent position as judged by the group of observers in the auditorium listening to the reproduced speech. Words were uttered from some 15 positions on the pick-up stage in random order. The 9 positions shown in Fig. 1 were always included in the 15, the remaining positions being introduced to minimize memory effects. The reproducing system was switched off while the caller moved from one position to the other.

In the first series of tests, the majority of the observers had no previous experience with the set-up. They simply were given a sheet of coordinate paper with a single line ruled on it to indicate the line of the gauze curtain and asked to locate the apparent position of the caller with respect to this line. Following these tests, the observers were permitted to listen to speech from various announced positions on the pick-up stage. This gave them some notion of the approximate outline of what might be called the "virtual" stage. These tests then were repeated. As there was no significant difference in results, the data from both tests have been averaged and are shown in Fig. 1.

The small diagram at the top of Fig. 1 shows the caller's positions with respect to the microphone positions on the pick-up stage. The corresponding average apparent positions when reproduced are shown with respect to the curtain line and the loudspeaker positions. The type of reproduction is indicated symbolically to the right of the apparent position diagrams.

With 3-channel reproduction there is a reasonably good corre-

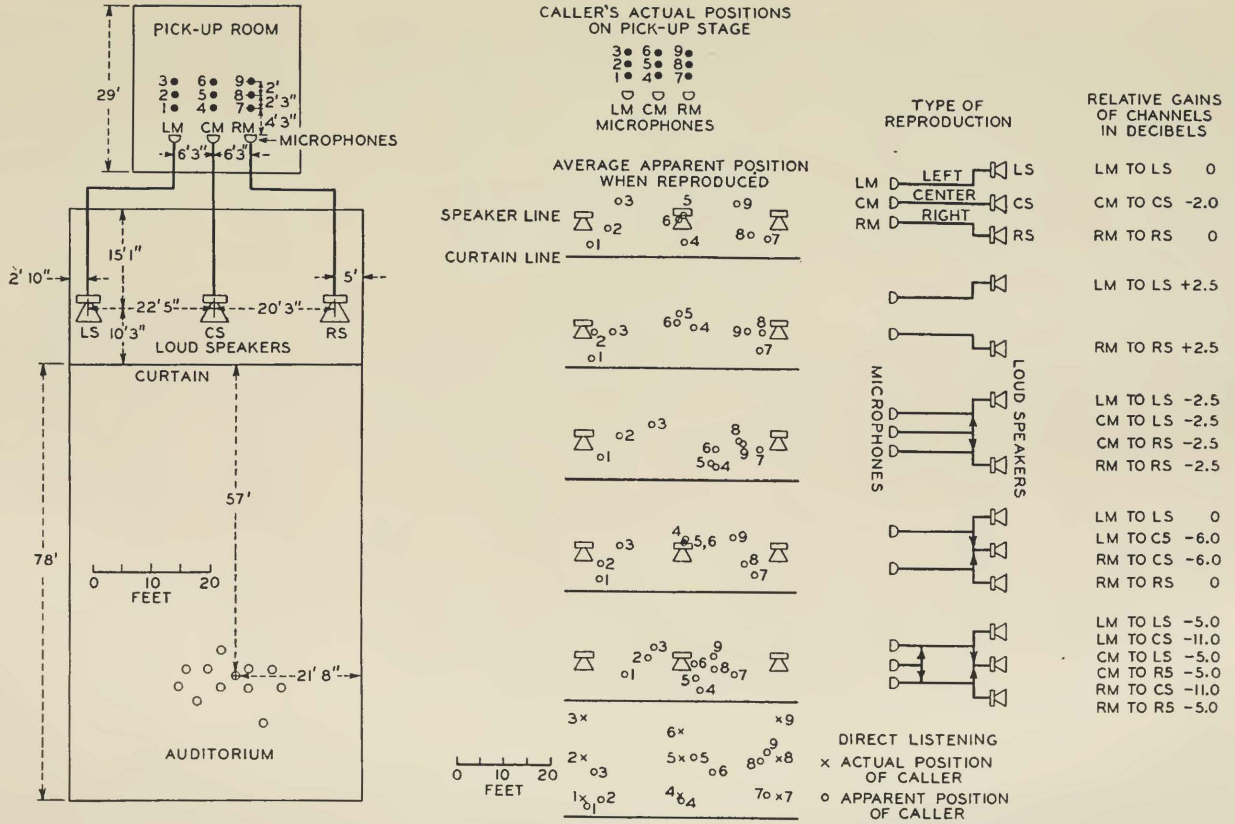


Fig. 1—Diagram of arrangement (left) for sound localization tests and (right) the results obtained.

spondence between the caller's actual position on the pick-up stage and his apparent position on the virtual stage. Apparent positions to the right or left correspond with actual positions to the right or left, and apparent front and rear positions correspond with actual front and rear positions. Thus the system afforded lateral or "angular" localization as well as fore and aft or "depth" localization. For comparison, there is shown in the last diagram the localization afforded by direct listening. The crosses indicate a caller's position in back of the gauze curtain and the circles indicate his apparent position as judged by the observers listening to his speech directly. In both cases, as the caller moved back in a straight line on the left or right side of the stage, he appeared to follow a curved path pulling in toward the rear center; e.g., compare the caller positions 1, 2, 3, with the apparent positions 1, 2, 3. This distortion was somewhat greater for 3-channel reproduction than for direct listening.

The results obtained with the 2-channel system show two marked differences from those obtained with 3-channel reproduction. Positions on the center line of the pick-up stage (i.e., 4, 5, 6) all appear in the rear center of the virtual stage, and the virtual stage depth for all positions is reduced. The virtual stage width, however, is somewhat greater than that obtained with 3-channel reproduction.

Bridging a third microphone across the 2-channel system had the effect of pulling the center line positions 4, 5, 6, forward, but the virtual stage depth remained substantially that afforded by 2-channel reproduction, while the virtual stage width was decreased somewhat. In this and the other bridged arrangements the bridging circuits employed amplifiers, as represented by the arrows in Fig. 1, in such a way that there was a path for speech current only in the indicated direction.

Bridging a third loud speaker across the 2-channel system had the effect of increasing the virtual stage depth and decreasing the virtual stage width, but positions on the center line of the pick-up stage appeared in the rear center of the virtual stage as in 2-channel reproduction.

Bridging both a third microphone and a third loud speaker across the 2-channel system had the effect of reducing greatly the virtual stage width. The width could be restored by reducing the bridging gains, but fading the bridged microphone out caused the front line of the virtual stage to recede at the center, whereas fading the bridged loud speaker out reduced the virtual stage depth. No fixed set of bridging gains was found that would enable the arrangement to create the virtual stage created by three independent channels. The gains used in

obtaining the data shown in Fig. 1 are indicated at the right of the symbolic circuit diagrams.

Factors Affecting Depth Localization

Before attempting to explain the results that have been given in the foregoing, it may be of interest to consider certain additional observations that bear more specifically upon the factors that enter into the "depth" and "angular" localization of sounds. The microphones on the pick-up stage receive both direct and reverberant sound, the latter being sound waves that have been reflected about the room in which the pick-up stage is located. Similarly, the observer receives the reproduced sounds directly and also as reverberant sound caused by reflections about the room in which he listens. To determine the effects of these factors, the following three tests were made:

1. Caller remained stationary on the pick-up stage and close to microphone, but the loudness of the sound received by the observer was reduced by gain control. This was loudness change without a change in ratio of direct to reverberant sound intensity.

2. Caller moved back from microphone, but gain was increased to keep constant the loudness of the sound received by the observer. This was a change in the ratio of direct to reverberant sound intensity without a loudness change.

3. Caller moved back from microphone, but no changes were made in the gain of the reproducing system. This changed both the ratio and the loudness.

All of the observers agreed that the caller appeared definitely to recede in all three cases. That is, either a reduction in loudness or a decrease in ratio of direct to reverberant sound intensity, or both, caused the sound to appear to move away from the observer. Position tests using variable reverberation with a given pick-up stage outline showed that increasing the reverberation moved the front line of the virtual stage toward the rear, but had slight effect upon the rear line. When the microphones were placed outdoors to eliminate reverberation, reducing the loudness either by changing circuit gains or by increasing the distance between caller and microphone moved the whole virtual stage farther away. It is because of these effects that all center line positions on the pick-up stage appeared at the rear of the virtual stage for 2-channel reproduction.

It has not been found possible to put these relationships on a quantitative basis. Probably a given loudness change, or a given change in ratio of direct to reverberant sound intensity, causes different sensations of depth depending upon the character of the reproduced sound

and upon the observer's familiarity with the acoustic conditions surrounding the reproduction. Since the depth localization is inaccurate even when listening directly, it is difficult to obtain sufficiently accurate data to be of much use in a quantitative way. Because of this inaccuracy, good auditory perspective may be obtained with reproduced sounds even though the properties controlling depth localization depart materially from those of the original sound.

Angular Localization

Fortunately, the properties entering into lateral or angular localization permit more quantitative treatment. In dealing with angular localization, it has been found convenient to neglect entirely the effects of reverberant sound and to deal only with the properties of the sound waves reaching the observer's ears without reflections. The reflected waves or reverberant sounds do appear to have a small effect on angular localization, but it has not been found possible to deal with such sound in a quantitative way. One of the difficulties is that, because of differences in the build-up times of the direct and reflected sound waves, the amount of direct sound relative to reverberant sound reaching the observer's ears for impulsive sounds such as speech and music is much greater than would be expected from steady state methods of dealing with reverberant sound.

For the case of a plane progressive wave from a single sound source, and where the observer's head is held in a fixed position, there are apparently only three factors that can assist in angular localization: namely, phase difference, loudness difference, and quality difference between the sounds received by the two ears.

In applying these factors to the localization of sounds from more than one source, as in the present case, the effects of phase differences have been neglected. It is difficult to see how phase differences in this case can assist in localization in the ordinary way. The two remaining factors, loudness and quality differences, both arise from the directivity of hearing. This directivity probably is due in part to the shadow and diffraction effects of the head and to the differences in the angle subtended by the ear openings. Measurements of the directivity with a source of pure tone located in various positions around the head in a horizontal plane have been reported by Sivian and White.[2] From these measurements, the loudness level differences between near and far ears have been determined for various frequencies. These differences are shown in Fig. 2 from which, using the pure tone data given, similar loudness level differences for complex tones may be calculated. Such calculated differences for speech are shown in Fig. 3.

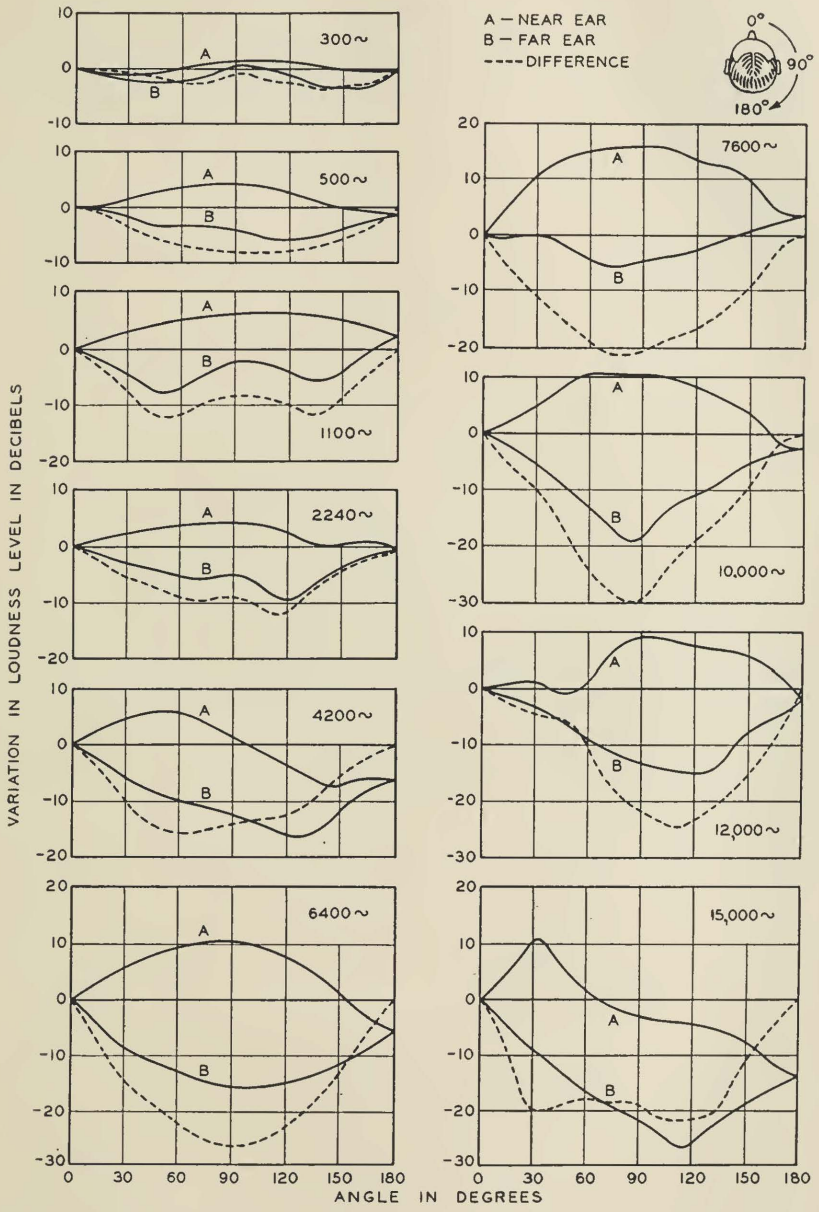


Fig. 2—Variation in loudness level as a sound source is rotated in a horizontal plane around the head.

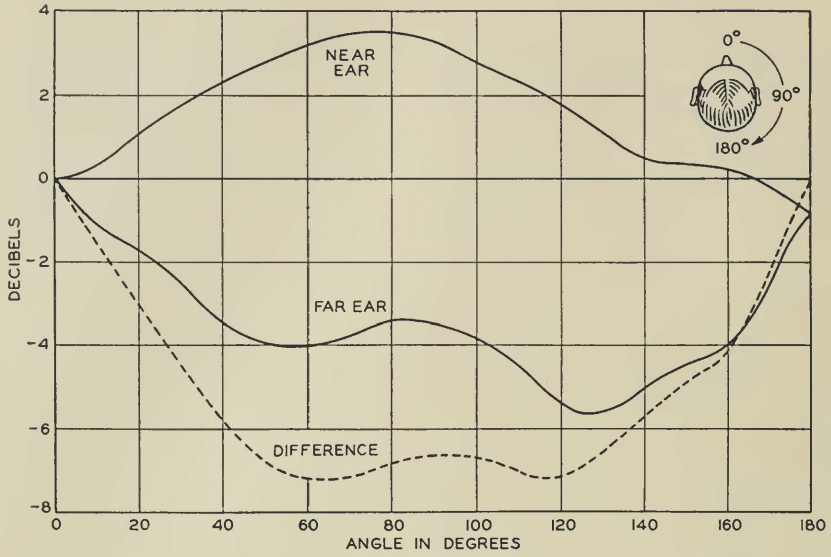


Fig. 3—Variation in loudness as a speech source is rotated in a horizontal plane around the head.

As may be inferred from the varying shapes of the curves of Fig. 2, the directive effects of hearing introduce a frequency distortion more or less characteristic of the direction from which the sound comes. Thus the character or quality of complex sounds varies with the angle of the source. There are quality differences at each ear for various angles of source, and quality differences between the two ears for a given angle of source. In Fig. 4 is shown the frequency distortion at the right ear when a source of sound is moved from a position on the right to one on the left of an observer. It is a graph of the "difference" values of Fig. 2 for an angle of 90 degrees. Frequencies above 4,000 cycles per second are reduced by as much as 15 to 30 decibels. This amount of distortion is sufficient to affect materially the quality of speech, particularly as regards the loudness of the sibilant sounds.

Reference to the difference curve of Fig. 3 shows that if, for example, a source of speech is 20 degrees to the right of the median plane the speech heard by the right ear is 3 db louder than that heard by the left ear. A similar difference exists when the angle is 167 degrees. Presumably, when the right ear hears speech 3 db louder than the left, the observer localizes the sound as coming from a position 20 degrees or 167 degrees to the right, depending upon the quality of the speech. If this be assumed to be true, even though the difference is caused by the combination of sounds of similar quality from several sources, it should be possible to calculate the apparent angle.

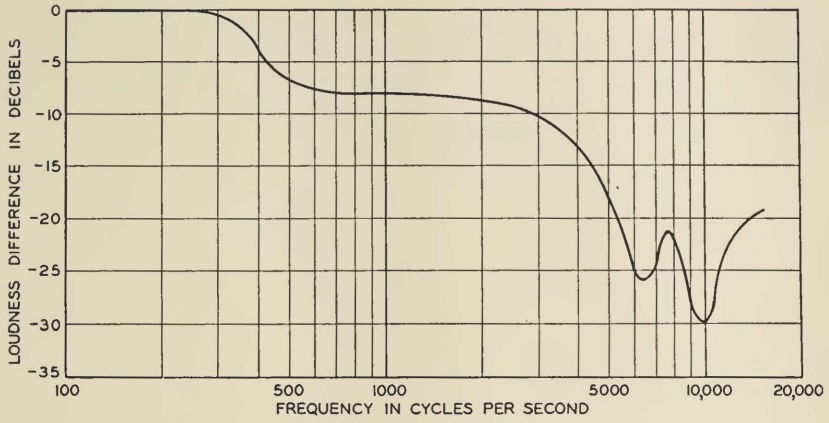


Fig. 4—Loudness difference produced in the right ear when a source of pure tone is moved from the right to the left of an observer.

Loudness Theory of Localization

Upon this assumption the apparent angle of the source as a function of the difference in decibels between the speech levels emitted by the loud speakers of the 2- and 3-channel systems has been calculated. Each loud speaker contributes an amount of direct sound loudness to each ear, depending upon its distance from, and its angular position with respect to, the observer. These contributions were combined on a power basis to give a resultant loudness of direct sound at each ear, from which the difference in loudness between the two ears was determined. The calculated results for the 2- and 3-channel systems are shown by the solid lines in Fig. 5. The y axis shows the apparent angle, positive angle being measured in a clockwise direction. The x axis shows the difference in decibels between the speech levels from the right and left loud speakers. The points are observed values taken from Fig. 1. The observed apparent angles were obtained directly from the average observer's location and the average apparent positions shown in Fig. 1. The speech levels from each of the loud speakers were calculated for each position on the pick-up stage. This was done by assuming that the waves arriving at the microphone had relative levels inversely proportional to the squares of the distances traversed. By correcting for the angle of incidence and for the known relative gains of the systems, the speech levels from the loud speakers were obtained.

A comparison of the observed and calculated results seems to indicate that the loudness difference at the two ears accounts for the greater part of the apparent angle of the reproduced sounds. If this is true,

the angular location of each position on the virtual stage results from a particular loudness difference at the two ears produced by the speech coming from the loud speakers. When three channels are used a definite

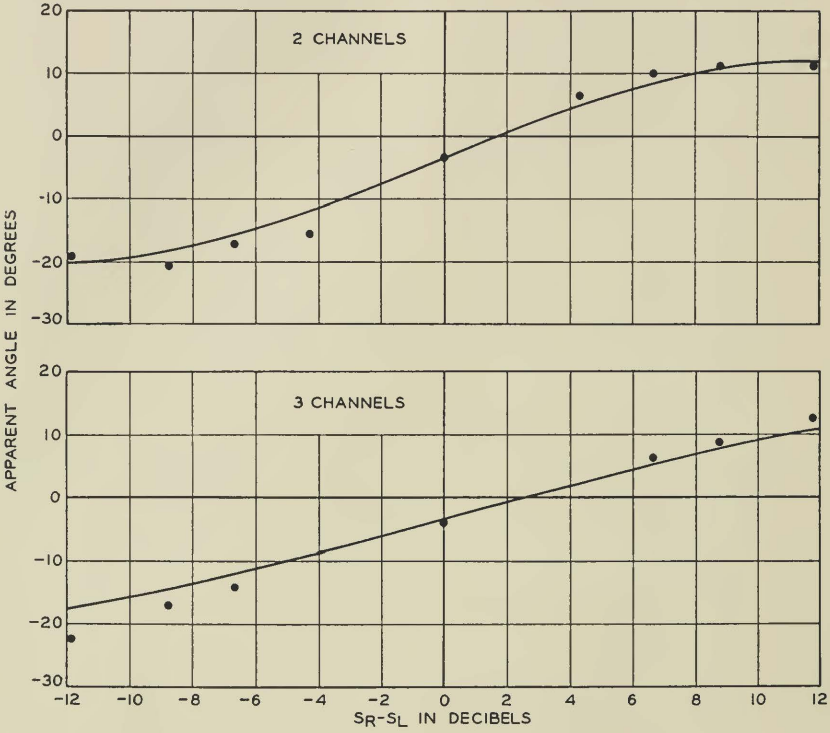


Fig. 5—Calculated and observed apparent angles for 2- and 3-channel reproduction.

set of loud speaker speech levels exists for each position on the pick-up stage. To create these same sets of loud speaker speech levels with the 3-microphone 3-loud speaker bridging arrangement already discussed, it would be necessary to change the bridging gains for each position on the pick-up stage. Hence it could not be expected that the arrangement as used (i.e., with fixed gains) would create a virtual stage identical with that created by 3-channel reproduction. However, with proper technique, bridging arrangements on a given number of channels can be made to give better reproduction than would be obtained with the channels alone.

Experimental Verification of Theory

Considerations of loudness difference indicate that all caller positions on the pick-up stage giving the same relative loud speaker outputs

should be localized at the same virtual angle. The solid lines of Fig. 6 show a stage layout used to test this hypothesis with the 2-channel system. All points on each line have a constant ratio of distances to

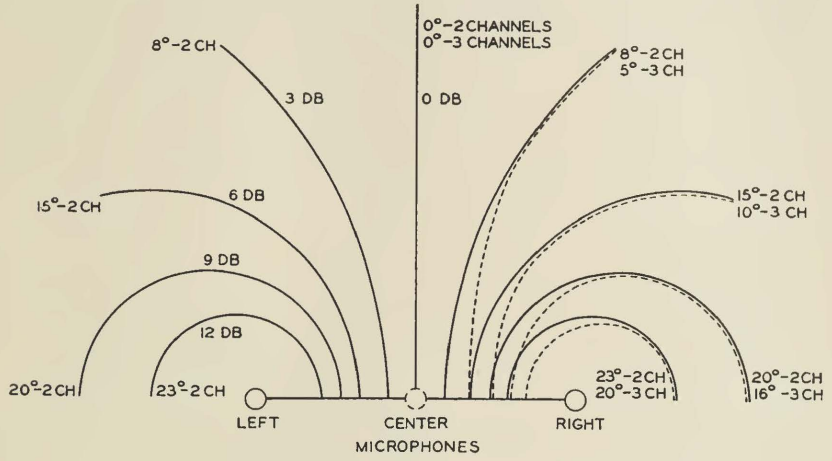


Fig. 6—Pick-up stage contour lines of constant apparent angle.

the microphones. The resulting direct sound differences in pressure expressed in decibels and the corresponding calculated apparent angles are indicated beside the curves. The apparent angles were calculated for an observing position on a line midway between the two loud speakers but at a distance from them equal to the separation between them. The microphones were turned face up at the height of the talker's lips to eliminate quality changes caused by changing incidence angle. It was found that a caller walking along one of these lines maintained a fairly constant virtual angle. For caller positions far from the microphones the observed angles were somewhat greater than those computed. For highly reverberant conditions, the tendency was toward greater calculated than observed angles. Reverberation also decreased the accuracy of localization.

A change of relative channel gain caused a change in virtual angle as would be expected from loudness difference considerations. For instance, if the caller actually walked the left 3-db line, he seemed to be on the 6-db line when the left channel gain was raised 3 db. Many of the effects of moving about the pick-up stage could be duplicated by volume control manipulation as the caller walked forward and backward on the center path. With a bridged center microphone substituted for the two side microphones similar effects were possible and, in addition, the caller by speaking close to the microphone could be brought to the front of the virtual stage.

For observing positions near the center of the auditorium the observed angles agreed reasonably well with calculations based only upon loudness differences. As the observer moved to one side, however, the virtual source shifted more rapidly toward the nearer loud speaker than was predicted by the computations. This was true of reproduction in the auditorium, both empty and with damping simulating an audience, and outdoors on the roof. Computations and experiment also show a change in apparent angle as the observer moves from front to rear, but its magnitude is smaller than the error of an individual localization observation. Consequently, observers in different parts of the auditorium localize given points on the pick-up stage at different virtual angles.

Because the levels at the three microphones are not independent, and because the desired contours depend upon the effects at the ears, a 3-channel stage is not as simple to lay out as a 2-channel stage. For a given observing position, however, a set of contour lines can be calculated. The dashed lines at the right of Fig. 6 show four contours thus calculated for the circuit condition of Fig. 1 and the observing position previously mentioned. The addition of the center channel reduces the virtual angle for any given position on the pick-up stage by reducing the resultant loudness difference at the ears. Although the 3-channel contours approach the 2-channel contours in shape at the back of the stage, a given contour results in a greater virtual angle for 2- than for 3-channel reproduction.

Similar effects were obtained experimentally. As in 2-channel reproduction, movements of the caller could be simulated by manipulation of the channel gains. From an observing standpoint the 3-channel system was found to have an important advantage over the 2-channel system in that the shift of the virtual position for side observing positions was smaller.

Effects of Quality

If the quality from the various loud speakers differs, the quality of sound is important to localization. When the 2-channel microphones were so arranged that one picked up direct sound and reverberation while the other picked up mostly reverberation, the virtual source was localized exactly in the "direct" loud speaker until the power from the "reverberant" loud speaker was from 8 to 10 db greater. In general, localization tends toward the channel giving most natural or "closeup" reproduction, and this effect can be used to aid the loudness differences in producing angular localization.

Principal Conclusions

The principal conclusions that have been drawn from these investigations may be summarized as follows:

1. Of the factors influencing angular localization, loudness difference of direct sound seems to play the most important part; for certain observing positions the effects can be predicted reasonably well from computations. When large quality differences exist between the loudspeaker outputs, the localization tends toward the more natural source. Reverberation appears to be of minor importance unless excessive.

2. Depth localization was found to vary with changes in loudness, the ratio of direct to reverberant sound, or both, and in a manner not found subject to computational treatment. The actual ratio of direct to reverberant sound, and the change in the ratio, both appeared to play a part in an observer's judgment of stage depth.

3. Observers in various parts of the auditorium localize a given source at different virtual positions, as is predicted by loudness computations. The virtual source shifts to the side of the stage as the observer moves toward the side of the auditorium. Although quantitative data have not been obtained, qualitative data on these effects indicate that the observed shift is considerably greater than that computed. Moving backward and forward in the auditorium appears to have only a small effect on the virtual position.

4. Because of these physical factors controlling auditory perspective, point-for-point correlation between pick-up stage and virtual stage positions is not obtained for 2- and 3-channel systems. However, with stage shapes based upon the ideas of Fig. 7, and with suitable use of quality and reverberation, good auditory perspective can be produced. Manipulation of circuit conditions probably can be used advantageously to heighten the illusions or to produce novel effects.

5. The 3-channel system proved definitely superior to the 2-channel by eliminating the recession of the center-stage positions and in reducing the differences in localization for various observing positions. For musical reproduction, the center channel can be used for independent control of soloist renditions. Although the bridged systems did not duplicate the performance of the physical third channel, it is believed that with suitably developed technique their use will improve 2-channel reproduction in many cases.

6. The application of acoustic perspective to orchestral reproduction in large auditoriums gives more satisfactory performance than probably would be suggested by the foregoing discussions. The instruments near the front are localized by every one near their correct positions.

In the ordinary orchestral arrangement, the rear instruments will be displaced in the reproduction depending upon the listener's position, but the important aspect is that every auditor hears differing sounds from differing places on the stage and is not particularly critical of the exact apparent positions of the sounds so long as he receives a spatial impression. Consequently 2-channel reproduction of orchestral music gives good satisfaction, and the difference between it and 3-channel reproduction for music probably is less than for speech reproduction or the reproduction of sounds from moving sources.

References

1. "Some Physical Factors Affecting the Illusion in Sound Motion Pictures," J. P. Maxfield. *Jour. Acous. Soc.*, July, 1931.
2. "Minimum Audible Sound Fields," L. J. Sivian and S. D. White. *Jour. Acous. Soc.*, April, 1933.

Loud Speakers and Microphones*

By E. C. WENTE and A. L. THURAS

In ordinary radio broadcast of symphony music, the effort is to create the effect of taking the listener to the scene of the program, whereas in reproducing such music in a large hall before a large gathering the effect required is that of transporting the distant orchestra to the listeners. Lacking the visual diversion of watching the orchestra play, such an audience centers its interest more acutely in the music itself, thus requiring a high degree of perfection in the reproducing apparatus both as to quality and as to the illusion of localization of the various instruments. Principles of design of the loud speakers and microphones used in the Philadelphia-Washington experiment are treated at length in this paper.

AS EARLY as 1881 a large scale musical performance was reproduced by telephone instruments at the Paris Electrical Exhibition. Microphones were placed on the stage of the Grand Opera and connected by wires to head receivers at the exposition. It is interesting to note that separate channels were provided for each ear so as to give to the music perceived by the listener the "character of relief and localization." With head receivers it is necessary to generate enough sound of audible intensity to fill only a volume of space enclosed between the head receiver and the ear. As no amplifiers were available, the production of enough sound to fill a large auditorium would have been entirely outside the range of possibilities. With the advent of telephone amplifiers, microphone efficiency could be sacrificed to the interest of good quality where, as in the reproduction of music, this was of primary interest. When amplifiers of greater output power capacity were developed, loud speakers were introduced to convert a large part of the electrical power into sound so that it could be heard by an audience in a large auditorium. Improvements have been made in both microphones and loud speakers, resulting in very acceptable quality of reproduction of speech and music; as is found, for instance, in the better class of motion picture theaters.

In the reproduction, in a large hall, of the music of a symphony orchestra the approach to perfection that is needed to satisfy the habitual concert audience undoubtedly is closer than that demanded for any other type of musical performance. The interest of the listener here lies solely in the music. The reproduction therefore should be

* Third paper in the Symposium on Wire Transmission of Symphonic Music and Its Reproduction in Auditory Perspective. Presented at Winter Convention of A. I. E. E., New York City, Jan. 23–26, 1934. Published in *Electrical Engineering*, January, 1934.

such as to give to a lover of symphonic music esthetic satisfaction at least as great as that which would be given by the orchestra itself playing in the same hall. This is more than a problem of instrument design, but this paper will be restricted to a discussion of the requirements that must be met by the loud speakers and microphones, and to a description of the principles of design of the instruments used in the transmission of the music of the Philadelphia Orchestra from Philadelphia to Constitution Hall in Washington. Some of the requirements are found in the results of measurements that have been made on the volume and frequency ranges of the music produced by the orchestra.

General Considerations

The acoustic powers delivered by the several instruments of a symphony orchestra, as well as by the orchestra as a whole, have been investigated by Sivian, Dunn, and White. Figure 1 was drawn on the basis of the values published by them.[1] The ordinates of the horizontal lines give the values of the peak powers within the octaves indicated by the positions of the lines. For a more exact interpretation of these values the reader is referred to the original paper, but the chart here given will serve to indicate the power that a loud speaker must be capable of delivering in the various frequency regions, if the reproduced music is to be as loud as that given by the orchestra itself. However, it was the plan in the Philadelphia-Washington experiment to reproduce the orchestra, when desired, at a level 8 or 10 db higher, so that with three channels each loud speaking system had to be able to deliver two or three times the powers indicated in Fig. 1. Sivian, Dunn, and White also found that for the whole frequency band the peak powers in some cases reached values as high as 65 watts. In order to go 8 db above this value, each channel would have to be capable of delivering in the neighborhood of 135 watts.

The chart (Fig. 1) shows that the orchestra delivers sound of comparable intensity throughout practically the whole audible range. Although it is conceivable that the ear would not be capable of detecting a change in quality if some of the higher or lower frequencies were suppressed, measurements published by W. B. Snow [2] show that for any change in quality in any of the instruments to be undetectable the frequency band should extend from about 40 to about 13,000 c.p.s. The necessary frequency ranges that must be transmitted to obviate noticeable change in quality for the different orchestral instruments are indicated in the chart of Fig. 2, which is taken from the paper by Snow.

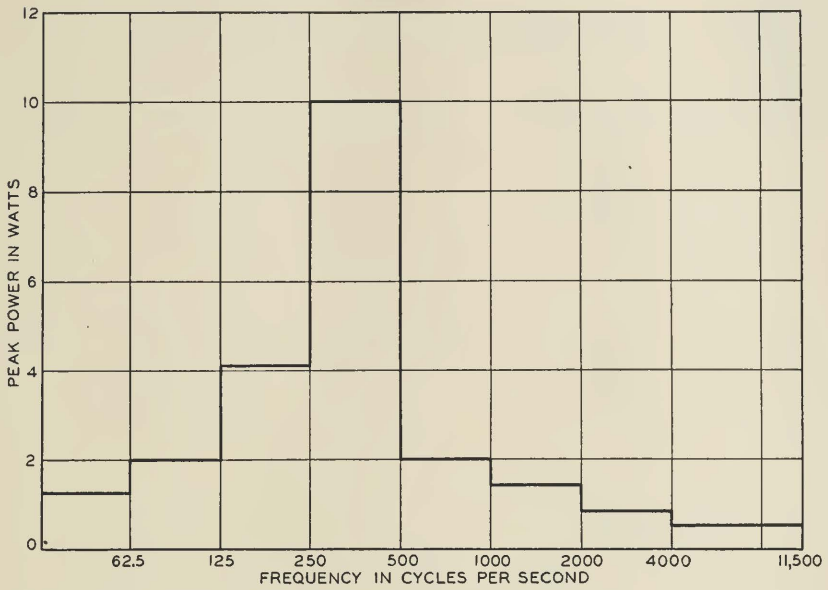


Fig. 1—Peak powers delivered by an orchestra within various frequency regions.

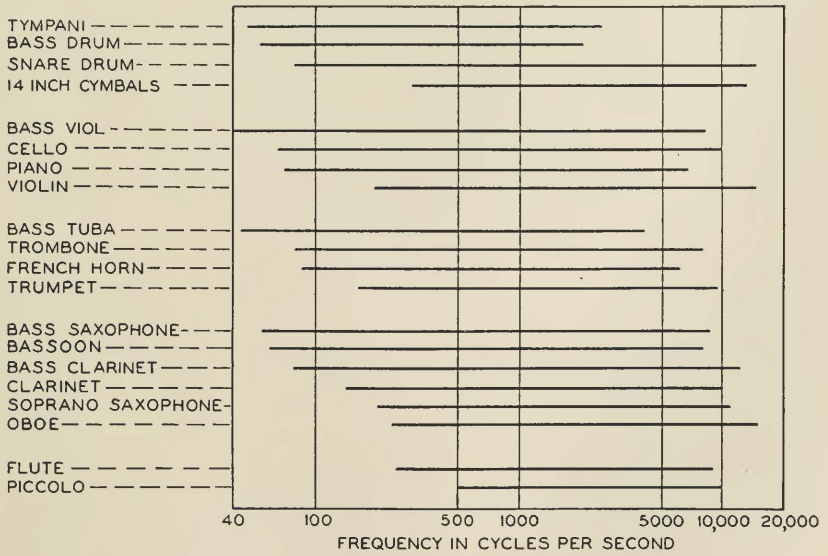


Fig. 2—Frequency transmission range required to produce no noticeable distortion for orchestral instruments.

Thus far only the sound generated by the orchestra itself has been considered. However, it is well known that the esthetic value of orchestral music in a concert hall is dependent to a very great extent upon the acoustic properties of the hall. At first thought one might be inclined to leave this out of account in considering the reproduction by a loud speaking system, as one should normally choose a hall known to have satisfactory acoustics for an actual orchestra. There would be no further problem in this if the orchestral instruments and the loud speaker radiated the sound uniformly in all directions, but some of the important instruments are quite directive; i.e., they radiate much the greater portion of their sound through a relatively small angle. As an example, a polar diagram giving the relative intensities of the sound radiated in various directions by the violin is given in Fig. 3, which is taken from a paper published by Backhaus.[3] The

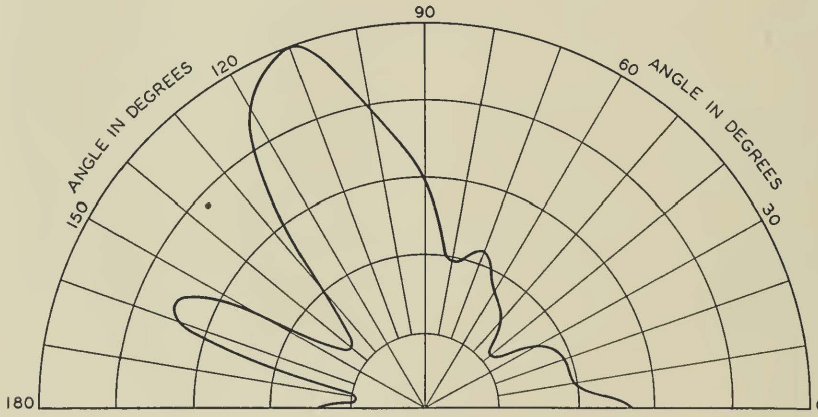


Fig. 3—Variation of intensity with direction of the sound radiated by a violin (660 c.p.s.).

directional characteristics of some of the instruments is one of the chief reasons why the music from an orchestra does not sound the same in all parts of a concert hall. The music which we hear comes to us in part directly and in part indirectly; i.e., after one or more reflections from the walls. Both contribute to the esthetic value of the music. The ratio of the direct to the indirect sound, which has been designated by Hughes [4] as the *acoustic ratio*, is to a first approximation inversely proportional to the product of the reverberation time and the angle through which the sound is radiated.[5] For a steady tone by far the greater part of the intensity at a given point in a hall remote from the source is attributable to the indirect sound. However, inasmuch as many of the tones of a musical selection are

of short duration, the direct sound is of great importance; it is this sound alone which enables us to localize the source. So far as this ratio is concerned, a decrease in the radiating angle of a loud speaker is equivalent to a reduction in the reverberation time of the hall. The effect on the music, however, is not entirely equivalent, for the rate of decay of sound in the room is unaltered by a change in directivity of the source, as this depends only on the reverberation time.

As already pointed out, some of the instruments of the orchestra are quite directive and others are nondirectional. In general, it may be said that the instruments of lower register are less directive than those of higher register. To have each instrument as reproduced by the loud speaker sound just as the instrument itself would sound in the same hall, the loud speaker would have to reproduce the music from each instrument with a directivity corresponding to that of the instrument itself. This manifestly is impossible. The best that can be hoped for is a compromise. Let the loud speaking system be designed so that it is nondirective for the lower frequencies, and at the higher frequencies it will radiate the sound through a larger angle than the most directive of the instruments and through a smaller angle than the least directive. Although this compromise means that the individual instruments will not sound exactly like the originals, it carries with it one advantage: At all the seats in the hall included in the radiating angle and at a given distance from the loud speaker the music may be heard to equal advantage, whereas with the orchestra itself the most desirable seats comprise only a certain portion of the hall. The optimum radiating angle is largely a matter of judgment; if it is too small the music will lack the spatial quality experienced at indoor concerts; if it is too large there will be a loss in definition.

There is another respect in which the directivity of the source can greatly affect the tone quality. Most loud speakers radiate tones of low frequency through a relatively large angle, but as the frequency is increased this angle becomes smaller and smaller. Under this condition the relation between the intensities of the high and low frequency tones as received directly will be different for almost all parts of the hall. Hence, even with equalization by electrical networks, the reproduction at best can be good only at a few places in the hall. Therefore, the sound radiated not only should be contained within a certain solid angle, but the radiation throughout this angle should be uniform at all frequencies.

The Loud Speaker

At present two kinds of loud speakers are in wide commercial use, the direct radiating and the horn types. Each has its merits, but the

latter was used in the Philadelphia-Washington experiment because it appears to have definite advantages where such large amounts of power are to be radiated. The horn type can be given the desired directive properties more readily, and higher values of efficiency throughout a wide frequency range are more easily realized. In consideration of the large power requirements, high efficiency is of special importance because it will keep to the lowest possible value the power capacity requirements of the amplifiers and because, with the heating proportional to one minus the efficiency, the danger of burning out the receiving units is reduced.

For efficiently radiating frequencies as low as 40 c.p.s., a horn of large dimensions is required. In order that the apparatus may not become too unwieldly the folded type of horn is preferable, but a large folded horn transmits high frequency tones very inefficiently. As actually used, therefore, the loud speaker was constructed in two units: one for the lower and the other for the higher frequencies, an electrical network being used to divide the current into two frequency bands, the point of division being about 300 c.p.s.

The Low Frequency Horn

When moderate amounts of power are transmitted through a horn the sound waves will suffer very little distortion, but when the power per unit area becomes large, second-order effects, usually neglected in considering waves of small amplitude, must be taken into account. The transmission of waves of large amplitude through an exponential horn has been investigated theoretically by M. Y. Rocard.[6] His investigation shows that if W watts are transmitted through the throat of an exponential horn a second harmonic of intensity RW will be generated, where R is given by the relation

$$R = \frac{(\gamma + 1)^2 f^2 \times 10^7 W}{2\rho c^3 f_0^2 A}, \tag{1}$$

in which f is the frequency of the fundamental, f_0 the cut-off frequency of the horn, c the velocity of sound, ρ the density of air, and A the area of the throat of the horn, all expressed in c.g.s. units. It may be noted that the intensity of the harmonic increases with the ratio of the frequency to the cut-off frequency of the horn; this is another argument against attempting to cover too wide a range of frequencies with a single horn. In Fig. 1 it is shown that in the region of 200 c.p.s. the orchestra gives peak powers of about 10 watts. If, therefore, 30 watts be set as the limit of power that the horn is to deliver at 200 c.p.s., 32 c.p.s. as the cut-off frequency of the horn, and 30 db below

the fundamental be assumed as the limit of tolerance of a second harmonic, from equation (1) a throat diameter of about 8 inches is determined.*

If the radiation resistance at the throat of a horn is not to vary appreciably with frequency, the mouth opening must be a substantial fraction of a wave-length. This condition calls for an unusually large horn if frequencies down to 40 c.p.s. and below are to be transmitted. However, the effect of variations in radiation resistance on sound output can be kept down to a relatively small value if the receiving unit is properly designed. This will be explained in the

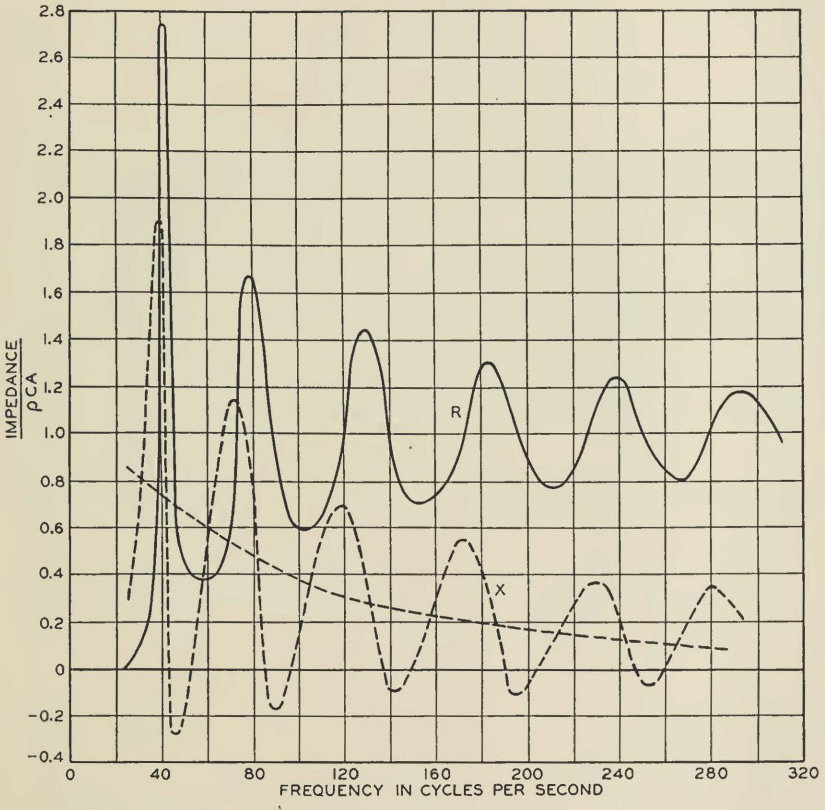


Fig. 4—Radiation resistance and reactance of low frequency horn.

next section. The low frequency horn used in these reproductions has a mouth opening of about 25 square feet. As computed from well-known formulas [7] for the exponential horn the impedance of this horn

* Since the original publication of this paper, experimental data have been obtained which indicate a second harmonic generation in horns 6 or more db below the value shown by Rocard's equation.[9]

with a throat diameter of 8 inches is shown in Fig. 4. These curves were computed under the assumption that the mouth of the horn is surrounded by a plane baffle of infinite extent, a condition closely approximated if the horn rests on a stage floor.

Low Frequency Receiving Unit

When a moving coil receiving unit, coupled to a horn, is connected to an amplifier having an output resistance equal to $n - 1$ times the damped resistance R of the driving coil, it can easily be shown that the sound power output is

$$P = \frac{\left(\frac{EBLT}{nR}\right)^2 r \times 10^{-9}}{\left[T^2 r + \frac{B^2L^2 \times 10^{-9}}{nR}\right]^2 + [x_d + T^2 x]^2} \text{ watts}, \tag{2}$$

where E is the open circuit voltage of the amplifier, L the length of wire in the receiver coil, T the ratio of the area of the diaphragm to the throat area of the horn, $r + jx$ the throat impedance of the horn, and x_d the mechanical reactance of the diaphragm and coil, the mechanical resistance of which is assumed to be negligibly small. From Fig. 4 it may be seen that the mean value of x increases as the frequency decreases to a value below 40 c.p.s., and that x is smaller than r except at the very lowest frequencies. If, therefore, the stiffness of the diaphragm be adjusted so that x_d is equal to T^2 times the mean value of x at 40 c.p.s., the second term in the denominator may be neglected without much error because it will have but little effect upon the sound output except at the higher frequencies, where the mass reactance of the coil and diaphragm may have to be taken into account.

If minimum variations in sound output are desired for variations in r,

$$\frac{B^2L^2 10^{-9}}{nRT^2} = r_0, \tag{3}$$

where r_0 is equal to the geometric mean value of r, which is approximately equal to $A\rho c$.

If α is the ratio of the resistance at any frequency to the mean value, and if the second term in the denominator is neglected, equation (2) becomes

$$P = \frac{E^2}{nR} \frac{\alpha}{(1 + \alpha)^2}. \tag{4}$$

In Fig. 4 it is shown that above 35 c.p.s. α has extreme values of 2.75 and 0.36, at which points there will be minimum values in P, but these

minimum values will not lie more than 1 db below the maximum values. Hence, if the receiver satisfies the condition of equation (3), the extreme variations in the sound output will not exceed 1 db, although the horn resistance varies by a factor of 7.5. Also it may be stated here that when the condition of equation (3) is satisfied the horn is terminated at the throat end by a resistance equal to the surge resistance of the horn. Thus equation (3) establishes a condition of minimum values in the transient oscillations of the horn.

The mean motional impedance of the loud speaker is $\frac{B^2L^2 \times 10^{-9}}{T^2 r_0}$, which, from equation (3), is equal to nR. The condition of equation (3) therefore specifies that the efficiency of the loud speaker shall be $\frac{n}{n+1}$. The maximum power that an amplifier can deliver without introducing harmonics exceeding a specified value is a function of the impedance into which it operates. Therefore, to obtain the maximum acoustic power for a specified harmonic content, the load impedance should have the value for which the product of the loud speaker efficiency and the power capacity of the amplifier has a maximum value. This optimum value of load impedance for the amplifier and loud speaker used in the Philadelphia-Washington experiments was found to be about 2.25 times the output impedance of the amplifier; the corresponding value of n then is 2.6 and the required efficiency 72 per cent. For best operating condition a definite value of receiver efficiency thus is specified.

The receiver may be made to satisfy the foregoing conditions regardless of the value of T, the ratio of diaphragm area to throat area. The area of the diaphragm has, however, a definite relation to the maximum power that the receiver can deliver at the low frequencies. The peak power delivered by the receiver is equal to $T^2 \alpha r_0 \xi^2 \omega^2 \times 10^{-7}$ peak watts where ξ is the maximum amplitude of motion of the diaphragm. Figure 1 shows that in the region lying between 40 and 60 c.p.s., peak powers reach a value of from 1 to 2 watts. However, the low frequency tones of an orchestra are undesirably weak and may advantageously be reproduced at a relatively higher level. Therefore it was decided to construct the loud speaker to be able to deliver 25 watts in this region.

As the coil moves out of its normal position in the air gap, the force factor varies. Harmonics thus will be generated, the intensities of which increase with increasing amplitude. A limit to the maximum value of the amplitude ξ thus is set by the harmonic distortion that one is willing to tolerate. In this receiver the maximum value of ξ

was taken equal to 0.060 in. Figure 4 shows that $\alpha\omega^2$ has a minimum value at about 50 cycles, where α is equal to about 0.4. These values give a ratio of 4.5 for T.

Inasmuch as $R = \frac{\sigma L^2}{v}$, where σ is the resistivity of the wire used for the coil and v the volume of the coil, from equation (3) is obtained

$$B^2v = n\sigma T^2 r_0 10^9. \tag{5}$$

The first member gives the total magnetic energy that must be set up in the region occupied by the driving coil. This value is fixed by the fact that all factors in the second member are specified. The same performance is obtained with a small coil and high flux density as with a large coil and low flux density, provided B^2v is held fixed, but the coil in any case should not be made so small that it will be incapable of radiating the heat generated within it without danger of overheating, nor so large that the mass reactance of the coil will reduce the efficiency at the higher frequencies.

This receiver unit, when constructed according to the above principles and when connected to an amplifier and a horn in the specified manner, should be capable of delivering power 3 or 4 times that delivered by the orchestra in the frequency region lying between 35 and 400 c.p.s., with an efficiency of about 70 per cent, and with a variation in sound output for a given input power to the amplifier of not more than 1 db throughout this range.

The High Frequency Horn

It is well known that a tapered horn of the ordinary type has a directivity which varies with frequency. Sound of low frequency is projected through a relatively large angle. As the frequency is increased this angle decreases progressively until, at frequencies for which the wave-length is small compared with the diameter of the mouth opening, the sound beam is confined to a very narrow angle about the axis of the horn.

If we had a spherical source of sound (i.e., a source consisting of a sphere, the surface of which has a radial vibratory motion equal in phase and amplitude at every point of the surface), sound would be radiated uniformly outward in all directions; or, if we had only a portion of a spherical surface over which the motion is radial and uniform, uniform sound radiation still would prevail throughout the solid angle subtended at the center of curvature by this portion of the sphere, provided its dimensions were large compared with the wave-length. Throughout this region the sound would appear to originate

at the center of curvature. Hence, for the ideal distribution of a spherical source within a region to be defined by a certain solid angle, it is necessary and sufficient that the radial motion be the same in amplitude and phase over the part of a spherical surface intercepted by the angle and having its center of curvature at the vertex and located at a sufficient distance from the vertex to make its dimensions large compared with the wave-length. If, further, these conditions are satisfied for this surface at all frequencies, all points lying within the solid angle will receive sound of the same wave form. A horn was designed to meet these requirements for the high frequency band.

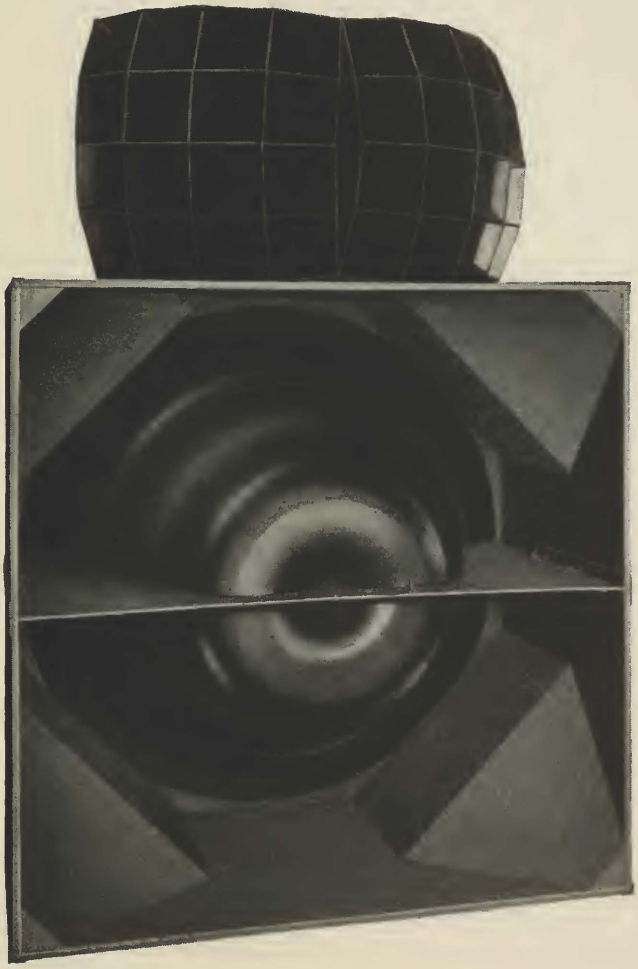


Fig. 5—Special loud speaker developed for auditory perspective experiment.

The horn, shown in the upper part of Fig. 5, comprises several separate channels, each of which has substantially an exponential taper. Toward the narrow ends these channels are brought together with their axes parallel, and are terminated into a single tapered tube which at its other end connects to the receiver unit. Sound from the latter is transmitted along the single tube as a plane wave and is divided equally among the several channels. If the channels have the same taper, the speed of propagation of sound in them is the same. The large ends are so proportioned and placed that the particle motion of the air will be in phase and equal over the mouth of the horn. This design gives a true spherical wave front at the mouth of the horn at all frequencies for which the transverse dimensions of the mouth opening are a large fraction of a wave-length.

As the frequency is increased, the ratio of wave-length to transverse width of the channels becomes less, and the sound will be confined more and more to the immediate neighborhood of the axis of each channel. The sound then will not be distributed uniformly over the mouth opening of the horn, but each channel will act as an independent horn. To have a true shperical wave front up to the highest frequencies, the horn would have to be divided into a sufficient number of channels to make the transverse dimension of each channel small compared with the wave-length up to the highest frequencies. If it is desired to transmit up to 15,000 c.p.s., it is not very practical to subdivide the horn to that extent. Both the cost of construction and the losses in the horn would be high if designed to transmit also frequencies as low as 200 c.p.s., as is the case under consideration. However, it is not important that at very high frequencies a spherical wave front be established over the whole mouth of the horn. For this frequency region it is perfectly satisfactory to have each channel act as an independent horn, provided that the construction of the horn is such that the direction of the sound waves coming from the channels is normal to the spherical wave front.

The angle through which sound is projected by this horn is about 60 degrees, both in the vertical and in the horizontal direction. For reproducing the orchestra two of these horns, each with a receiving unit, were used. They were arranged so that a horizontal angle of 120 degrees and a vertical angle of 60 degrees were covered. These angular extensions were sufficient to cover most of the seats in the hall with the loud speaker on the stage. The vertical angle determines to a large extent the ratio of the direct to the indirect sound transmitted to the audience. The vertical angle of 60 degrees was chosen purely on the basis of judgment as to what this ratio should be for the most pleasing results.

The High Frequency Receiving Unit

In the design of the low frequency receiver one of the main objectives was to reduce to a minimum the variations in sound transmission resulting from variations in the throat impedance of the horn. However, the high frequency horn readily can be made of a size such that the throat resistance has relatively small variations within the transmitting region. On the other hand, whereas the diameter of the diaphragm of the low frequency unit is only a small fraction of the wave-length, that of the high frequency unit must be several wave-lengths at the higher frequencies in order to be capable of generating the desired amount of sound. Unless special provisions are made there will be a loss in efficiency because of differences in phase of the sound passing to the horn from various parts of the diaphragm. The high frequency receiver therefore was constructed so that the sound generated by the diaphragm passes through several annular channels. There are enough of these channels to make the distance from any part of the diaphragm to the nearest channel a small fraction of a wave-length. These channels are so proportional that the sound waves coming through them have an amplitude and phase relation such that a substantially plane wave is formed at the throat of the horn.

In the appendix it is shown that, for the higher frequencies where the impedance of the horn may be taken as equal to ρc times the throat area and for the type of structure adopted, the radiation resistance is equal to

$$\rho c a T^2 \left[\frac{1}{k^2 h^2 T^2 + k^2 l^2 \cot^2 kl} \right] \tag{6}$$

and the reactance

$$-j \frac{\rho c a}{kh} T \left[1 - \frac{1}{kl \cot kl + \left(\frac{hT}{l} \right)^2 kl \tan kl} \right], \tag{7}$$

where a is the area of the throat of the horn, T the ratio of the area of the diaphragm to the throat area, $k = \omega/c$, and the other designations are those indicated in Fig. 11. At the lower frequencies the resistance is $T^2 r$ and the reactance $T^2 x$, where r and x are, respectively, the resistance and reactance of the throat of the horn.

Equation 6 shows that at a given frequency, other conditions remaining the same, the radiation resistance will have a maximum value when l is approximately equal to $\pi/2k = c/4f$. In Fig. 6 the resistances as computed from equation (6) are plotted as a function

of frequency for several values of h/w. It is seen from these curves that the resistance at the higher frequencies is determined very largely by the relation of h/w but is independent of it at the lower frequencies, where it is equal to ρcaT^2. At the lower frequencies where the mechanical impedance of the diaphragm is negligible, the efficiency, as was the case for the low frequency receiver, depends

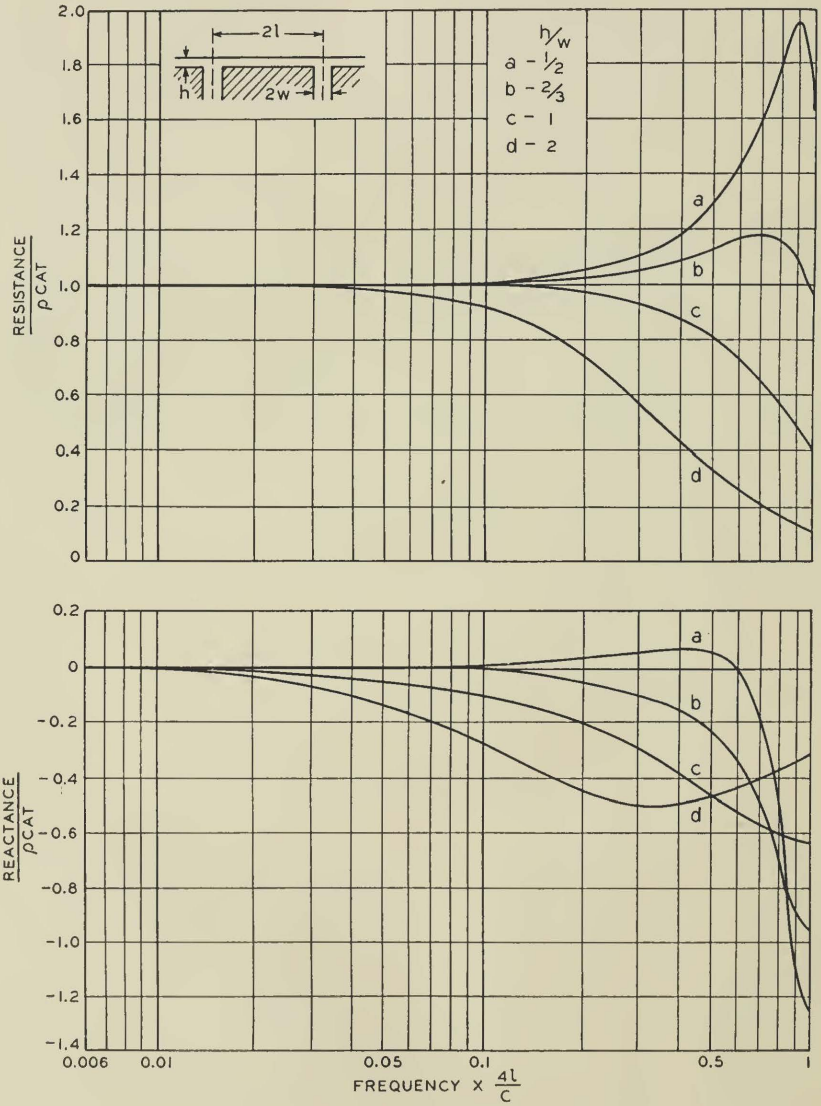


Fig. 6—Load impedance of speaker diaphragm.

upon the value of B^2v where v is the volume of the coil, but at the higher frequencies the efficiency decreases with increasing mass of the coil. It is advantageous, therefore, to keep v small and to make B as large as is practically possible. Values were selected to give the receiver an efficiency of 55 percent at the lower frequencies. For these conditions the relative sound power output was computed by equation (2) on the assumption that the receiver was connected to an amplifier having an output impedance equal to 0.45 times that of the receiver

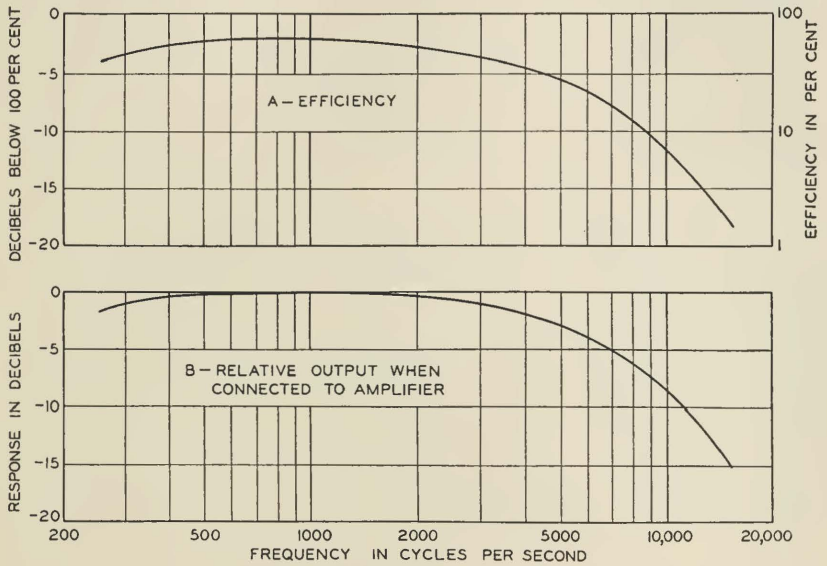


Fig. 7—Relative computed sound output of high frequency receiver.

at the lower frequencies. Figure 7 shows the values so obtained. Corresponding values obtained experimentally when the receiver was connected to the horn previously described are shown in Figs. 8 and 9, where the sizes of the rooms in which the values were obtained were, respectively, 5000 and 100,000 cubic feet. Both of these curves differ considerably from the computed curve, particularly as regards loss at high frequencies. The curve of Fig. 8 shows less, and that of Fig. 9 more, loss at high frequencies. The computed curve, however, refers to the total sound output, whereas the measured curves give average values of sound intensity in a certain part of the room, values dependent upon the acoustic characteristics of the room.

The number of high frequency receivers that must be used for each transmitting channel is governed largely by the amount of power that the system is to deliver before harmonics of an objectionable intensity

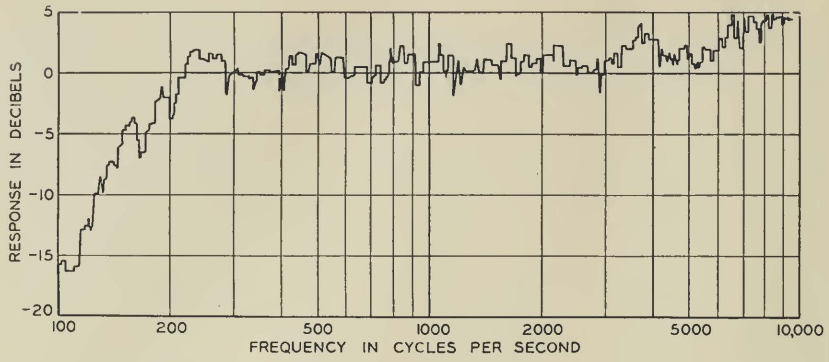


Fig. 8—Output-frequency characteristic of high frequency receiver as measured in a small room.

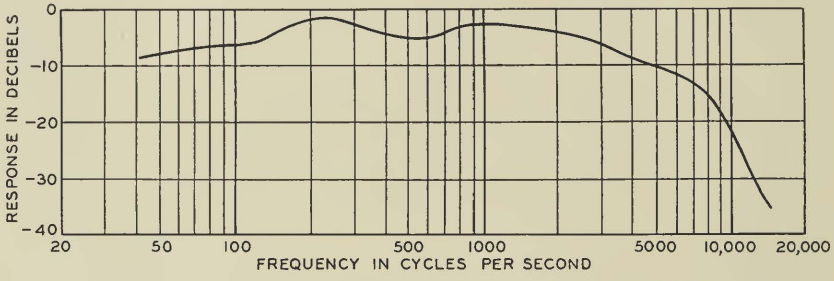


Fig. 9—Output-frequency characteristic of combined low and high frequency receivers as measured in a large room.

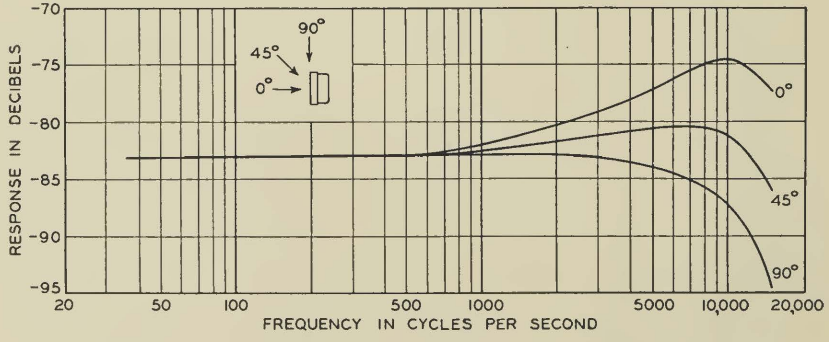


Fig. 10—Output-frequency characteristic of moving coil microphone.

are introduced. The generation of harmonics in a horn when transmitting waves of large amplitude already has been discussed. Let it suffice here to say that, for a given percentage harmonic distortion, the power that can be transmitted through the horn is proportional to the area of the throat and inversely proportional to the square of the ratio of the frequency to the cut-off frequency.

Inasmuch as the moving coil microphones used for the transmission of music in acoustic perspective have been described previously [8] they will not be discussed here at length. Their frequency response characteristic as measured in an open sound field for several different angles of incidence of the sound wave on the diaphragm are shown in Fig. 10 where it is seen that the response at the higher frequencies becomes less as the angle of incidence is increased. In general, this is not a desirable property, but with the instruments as used in this experiment the sound observed as coming from each loud speaker is mainly that which is picked up directly in front of each microphone; sound waves incident at a large angle do not contribute much.

At certain times the sound delivered by the orchestra is of very low intensity. Therefore it is important that the microphones have a sensitivity as great as possible, so that the resistance and amplifier noises may readily be kept down to a relatively low value. At 1,000 c.p.s. these microphones, without an amplifier, will deliver to a transmission line 0.05 microwatt when actuated by a sound wave having an intensity of 1 microwatt per square centimeter. This sensitivity is believed to be greater than that of microphones of other types having comparable frequency response characteristics, with the possible exception of the carbon microphone.

APPENDIX

Load Impedance of a Diaphragm Near a Parallel Wall with Slot Openings

First assume a diaphragm and a parallel wall of infinite extent separated by a distance h, and that the wall is slotted by a series of equally spaced openings as shown in Fig. 11. From symmetry it is known that when the diaphragm vibrates there will be no flow perpendicular to the plane of the paper or across the planes indicated by the dotted lines. Therefore only one portion of unit width, such as *abcdef* need be considered. Let the x and y reference axes be located as shown. If the general field equation

$$\frac{\partial^2 \varphi}{\partial x^2} + \frac{\partial^2 \varphi}{\partial y^2} + k^2 \varphi = 0 \tag{8}$$

is applied when the diaphragm has a normal velocity equal $\xi e^{i\omega t}$ the following boundary conditions are obtained:

When
$$x = 0, \quad \partial\varphi/\partial x = -\xi,$$
$$x = h, \quad \partial\varphi/\partial x = 0,$$
$$y = 0, \quad \partial\varphi/\partial y = 0,$$

and when $y = l$, the pressure is equal to the product of acoustic impedance and volume velocity or

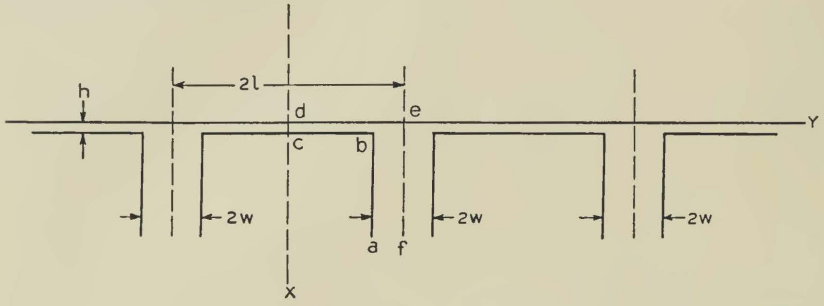


Fig. 11—Schematic diagram of diaphragm and parallel slotted wall of infinite length.

$$\frac{\rho}{h}\int_0^h \left(\frac{\partial\varphi}{dt}\right)_{y=l} dx = \frac{c\rho}{w}\int_0^h \left(-\frac{\partial\varphi}{\partial y}\right)_{y=l} dx$$

where φ is the velocity potential, $k = \omega/c$, and c is the velocity of sound.

The appropriate solution of equation (8) then is

$$\varphi = \frac{\xi}{k}\left[\frac{\cos ky}{kh\left(\cos kl + i\frac{h}{w}\sin kl\right)} - \frac{\cos k(x-h)}{\sin kh}\right].$$

The average reacting force per unit area of the diaphragm is

$$\frac{ik\rho c}{l}\int_0^l (\varphi)_{x=0}\, dy$$

Thus, for the impedance per unit area, which is equal to the force divided by the velocity, is obtained

$$\frac{\rho c l}{w}\left\{\left[\frac{\sin^2 kl}{k^2l^2}\,\frac{1}{\cos^2 kl + \left(\frac{h}{w}\right)^2 \sin^2 kl}\right] - j\frac{w}{h}\left[\frac{\frac{kh\cos kh}{\sin kh}\,kl - \frac{\sin kl\cos kl}{\cos^2 kl + \left(\frac{h}{w}\right)^2\sin^2 kl}}{k^2l^2}\right]\right\} \equiv r' + jx'.$$

In all practical types of loud speakers $kh \cos kh/\sin kh$ would be very nearly equal to 1; then

$$r' = \frac{\rho c l}{w}\left[\frac{1}{k^2 l^2\left(\left(\frac{h}{w}\right)^2 + \cot^2 kl\right)},\right]$$

$$x' = -\frac{\rho c l}{h}\left[\frac{kl - \dfrac{1}{\cot kl + \left(\dfrac{h}{w}\right)^2 \tan kl}}{k^2 l^2}\right].$$

If the total area of the diaphragm is A and that of the corresponding channels a, then $A/a = l/w$, approximately, and the total impedance becomes

$$r = \frac{\rho c A^2}{a} \cdot \frac{1}{\left(\frac{kh}{a}\right)^2 A^2 + k^2 l^2 \cot^2 kl},$$

$$x = -j\frac{\rho c A}{kh}\left[1 - \frac{1}{kl \cot kl + \left(\frac{h}{l}\frac{A}{a}\right)^2 kl \tan kl}\right].$$

References

1. "Absolute Amplitudes and Spectra of Certain Musical Instruments and Orchestra," L. J. Sivian, H. K. Dunn, and S. D. White. *Jour. Acous. Soc. Am.*, v. 2, Jan. 1931, p. 330.
2. "Audible Frequency Ranges of Music, Speech, and Noise," W. B. Snow. *Jour. Acous. Soc. Am.*, v. 3, July 1931, p. 155.
3. Backhaus, *Zeits. f. Tech. Physik*, v. 9, 491, 1928.
4. "Engineering Acoustics," L. E. C. Hughes, p. 47. Benn, London.
5. W. J. Albersheim and J. P. Maxfield, similar relations were presented in a paper before the Acoustical Society in May 1932.
6. "Sur la Propagation des Ondes Sonores d'Amplitude Finie," M. Y. Rocard. *Comptes Rendus*, Jan. 16, 1933, p. 161.
7. "Theory of Vibrating Systems and Sound," Crandall. P. 163 ff. D. Van Nostrand, New York.
8. "Moving Coil Telephone Receivers and Microphones," E. C. Wente and A. L. Thuras. *Jour. Acous. Soc. Am.*, v. 3, 1931, p. 44.
9. "Extraneous Frequencies Generated in Air Carrying Intense Sound Waves," A. L. Thuras, R. T. Jenkins and H. T. O'Neil. To be presented at mtg. of Acous. Soc. Am., April 30–May 1, 1934.

Amplifiers*

By E. O. SCRIVEN

Appreciable care is required in the design of a system which must amplify with great fidelity practically the whole range of audible frequencies and be capable of delivering a high level while at the same time providing a wide volume range. Some of the problems involved are discussed, particularly as applying to the equipment used in the reproduction in Washington, D. C., of the Philadelphia Symphony Orchestra playing in Philadelphia.

VACUUM tube amplifiers have been closely identified with the extension of the channels of communication since, with completion of the initial transcontinental telephone line 20 years ago, they first enabled New York to converse with San Francisco. There are now thousands of audio frequency amplifiers in telephone circuits and in sound picture theaters, public address systems, and other similar services as well as in the millions of radio receiving sets.

Along with the extension of the field of usefulness of audio amplifiers there has been continuing progress toward more faithful reproduction, better transmitters, better receivers, and better amplifiers. Those first telephone repeaters, although quite adequate for their immediate purpose, transmitted a frequency band only a few octaves wide. Very few radio sets even now cover a range above 3,000 c.p.s. without distortion, and the most up-to-date sound picture installation rarely can be depended upon for accurate reproduction of frequencies above 7000 or 8000 c.p.s. The requirements as to frequency range and freedom from distortion for any particular service are, in the last analysis, determined by public demand.

However, when one undertakes to reproduce an orchestra like the Philadelphia Symphony and to reproduce it in such a manner as to satisfy the critical ear of the director, or that of the devotee of symphonic concerts, one has to provide something out of the ordinary in audio amplifiers.

In his paper, which forms a part of this symposium, Dr. Fletcher has pointed out that only the elimination of those frequencies below 40 c.p.s. and those above 15,000 c.p.s. produces no detectable difference in the reproduction of symphonic music. This, then, is the

* Fourth paper in the Symposium on Wire Transmission of Symphonic Music and Its Reproduction in Auditory Perspective. Presented at Winter Convention of A. I. E. E., New York City, Jan. 23–26, 1934. Published in *Electrical Engineering*, January, 1934.

frequency spectrum that the amplifier must be designed to handle. Also, it is important that there shall be uniform amplification of all parts of the frequency range and that no extraneous frequencies shall be introduced.

Of importance commensurate with the distortionless amplification of the complete frequency range of the orchestra is the provision of an equivalent volume of sound. The amplifier must be capable of supplying to the loud speakers without distortion an amount of energy that will produce a sound volume at least equivalent to that produced by the orchestra (the Philadelphia-Washington installation was designed to produce about 10 times this amount). And equally important, the amplifier must be so free from internal disturbances and from self-induced electrical fluctuations that the softest music, the weakest input to the microphone, can be reproduced without appreciable background noise. According to Fletcher the ratio of the heaviest playing of a large orchestra such as the Philadelphia Symphony Orchestra to the softest music such as that of a violin is about 10,000,000 to 1, or 70 db. Thus it is required that any noise be at least 75 db below the loudest tones; that is, there must be at least a 75-db volume range.

The sources of noise may be divided into 2 groups. In the first group are included the 60-cycle alternating current power supply, vibration or jar of mechanically unstable vacuum tubes, contact and thermoelectric potentials, and similar disturbances, which may be reduced to practically any degree depending upon the lengths to which one is willing to go to reduce them. In the second group are those electronic irregularities intimately associated with the operation of the vacuum tube and which depend somewhat upon the design, manufacture, and method of operation of the vacuum tube; and which, when sufficiently amplified and fed into a loud speaker, may be heard as noise. In general, the maximum volume range of an amplifier is reached when all other disturbances are reduced to the level of this tube noise.

It is evident, then, that under ordinary circumstances the limiting volume range of an amplifier is a function of the amount of amplification following the first tube. In other words, the magnitude of the signal voltage with respect to the noise voltage in the plate circuit of the first tube in a multistage amplifier determines the limiting volume range obtainable with that amplifier.

It will appear that in a sound reproduction system a highly efficient microphone simplifies the amplifier volume range requirements, and that loud speakers of high efficiency reduce the volume required from the amplifier.

Perhaps it is in order to inquire as to what makes an amplifier free from frequency distortion over a wide range. The answer might well be: attention to impedance relations. A compact, efficient amplifier requires several pieces of reactive apparatus such as transformers, retardation coils, and capacitors. One must remember that an inductance of one henry is equivalent to an impedance of 250 ohms at 40 c.p.s. but that it is nearly 100,000 ohms at 15,000 c.p.s.; that the grid circuit of the vacuum tube is not actually an open circuit even though the grid is maintained negative with respect to the cathode, but has a reactance which becomes important at high frequencies or with large ratio input transformers. Many years of development in this field have advanced the art to the point where transformers transmitting extremely wide bands now can be designed. The commercial production of such designs requires rigid inspection including shop transmission measurements under the actual conditions of use. The transformer must be designed for the particular type of vacuum tube with which it is to be used. First, however, the tube must be designed to permit its use under the proposed conditions and then it must be manufactured to close limits, every tube of a type like every other tube of that type.

This is, then, the general requirement for a wide frequency range amplifier: (1) attention to impedance relations; (2) meticulous design of each component for the particular job it has to do, and rigid inspection to insure that it does that job.

One might suppose that when the tube designer and the coil designer each had done his part the job was done. Such is not the case. The various pieces of apparatus have to be gathered together into a unit (often a current supply set for supplying anode, cathode, and grid potential is assembled with the amplifier) and out of this electrical and physical association is apt to arise "feed-back" and "noise."

When there is coupling between two parts of the amplification circuit which are at different potential or different phase there is feed-back. Feed-back sometimes is employed designedly to modify an amplifier characteristic, but, feed-back which may arise as a result of a particular arrangement of apparatus or wiring ordinarily will cause more or less severe frequency distortion. It may be induced due to stray electromagnetic or electrostatic fields, which must be eliminated by rearrangement of apparatus or by shielding; or it may be caused by common circuit impedance, requiring circuit modifications. In general, a low gain amplifier or one with limited frequency range presents no feed-back problems, but a study of a high-gain wide-range equipment usually is necessary in order to determine the best

arrangement. Often modifications of tentative circuit or apparatus must be made to obtain satisfactory operation.

The provision of a volume range of some 75 db on an energy basis became largely a matter of the suppression of a.-c. hum. The low inherent electronic noise effect of the Western Electric No. 262A vacuum tube and the relatively high level from the microphones kept electronic tube noise well in the background. Careful and in some cases rather elaborate shielding of audio transformers and leads and the segregation of the 60-cycle power equipment coupled with the use of vacuum tubes having indirectly heated cathodes and specially designed to have small stray fields prevented a.-c. hum trouble in the early stages. However, the Western Electric No. 242A vacuum tubes used in the push-pull final stage have filamentary cathodes, and when such tubes have raw a.-c. filament supply, a very appreciable 120-cycle component appears in the space current. Although theoretically in a perfectly balanced push-pull amplifier this component would be eliminated, in practice an exact balance cannot be obtained. As a final step in noise elimination, advantage was taken of the fact that each channel employed two amplifiers in parallel. Under such conditions and with proper phasing of the power supply to the two amplifiers the net a.-c. noise output of the two amplifiers in parallel will be less than that of either one alone.

Having reduced feed-back and noise to tolerable values, it remains to determine the operating conditions for maximuum output. The vacuum tube is not strictly a linear device, but, when properly used, the total harmonic content can be held to a low figure. For a high quality system the total harmonics produced in the system should not exceed one per cent of the fundamental. This requires that impedance and potential relations in the vacuum tubes should be adjusted to give approximately linear operation; and also that the design of audio transformers, particularly those carrying considerable levels, must be scrutinized carefully to insure that they operate over an essentially linear portion of the magnetization curve of the core material.

An instrument really essential to the design of high quality amplifiers is a high sensitivity harmonic analyzer that is capable of quickly and accurately resolving a complex wave into its simple components. By this means the effect of variations in circuit relations can be evaluated and the optimum condition for maximum distortionless power output determined.

It may be desirable at this point to examine the make-up of the audio amplification system used in the Philadelphia-Washington experiments. It should be noted that the arrangement of equipment

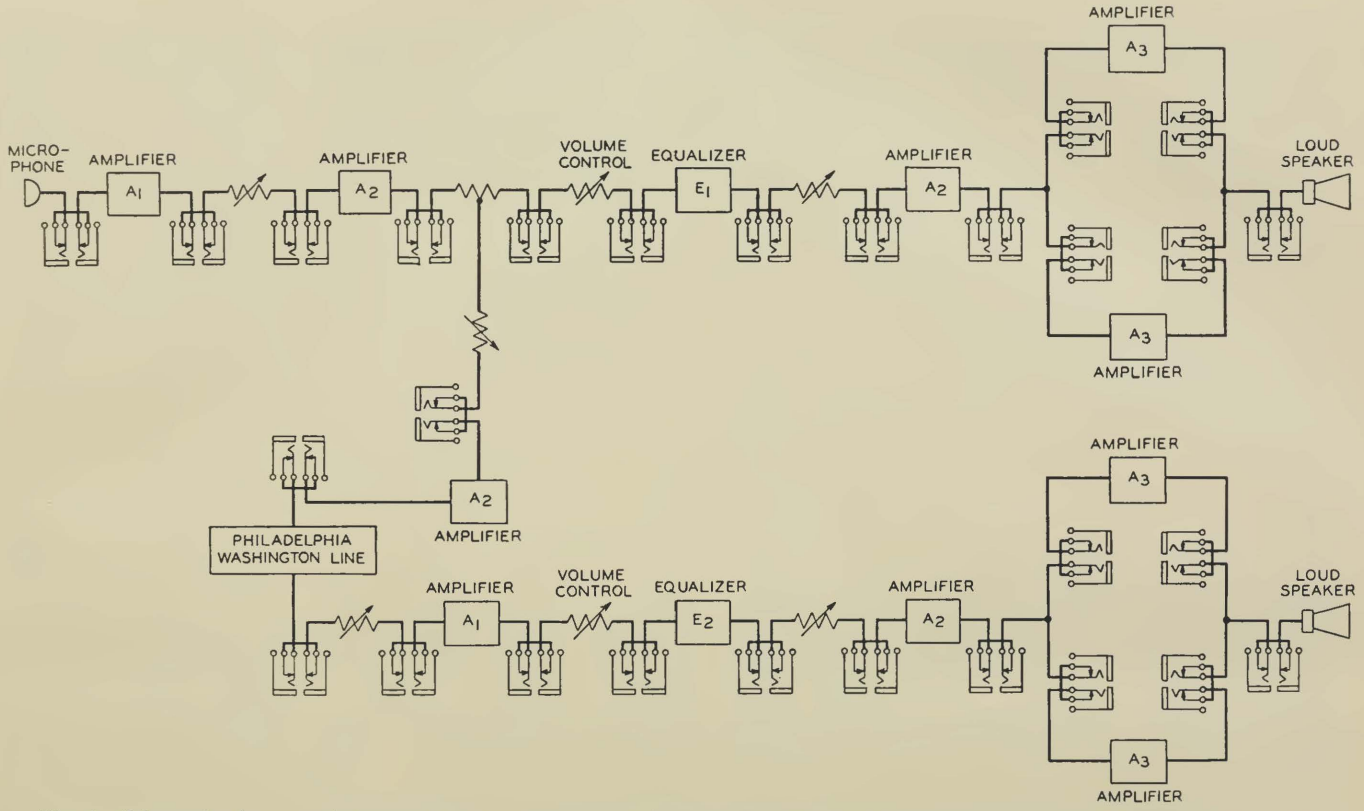


Fig. 1—Schematic diagram of the amplification system used in conveying Philadelphia symphonic music to Constitution Hall in Washington, D. C., and there reproducing it in auditory perspective.

provided for simultaneous reproduction at both Philadelphia and Washington. There were three complete and essentially equivalent channels of equipment actually in use and a fourth complete channel held in reserve as a spare.

Several stages of so-called voltage amplification were required preliminary to the final or power stage. There is, of course, no essential difference between a voltage amplifier and a power amplifier, the term "voltage amplifier" being applied to those preliminary stages of an amplification system the function of which is so to amplify the output of the pick-up device as to supply adequate driving voltage to the grids of the power stage. Theoretically, inasmuch as no energy is absorbed in the ideal grid circuit, this voltage increase might be supplied entirely by a high ratio input transformer. However, there are practical difficulties to the design of such a single stage amplifier and therefore multistage vacuum tube amplification is employed.

As a matter of convenience the voltage amplification for this system was obtained through the use of several separate amplifier units in tandem. This arrangement not only enabled the ready replacement of any unit of the system in case of failure, but it also facilitated the insertion of a pad, control potentiometer, or other network at any desired point. Several of these devices were required, and of course each introduced a loss. Thus the gross amplification of the system used for reproduction at Philadelphia was approximately 160 db and for Washington 240 db, although the actual difference in level between microphone output and loud speaker input was but from 80 to 90 db.

The general scheme of the amplification system is shown in Fig. 1. A_1 is a single-stage, single-tube Western Electric No. 80A amplifier slightly modified to meet the particular conditions of use; it has a gain of 30 db, and employs a Western Electric No. 262A vacuum tube. This tube has an equipotential cathode, the heater being operated on 10-volt 60-cycle alternating current and the anode being supplied from rectified alternating current. A_2 is a 2-stage amplifier having a single Western Electric No. 262A vacuum tube in the first stage and push-pull Western Electric No. 272A tubes in the second stage. It has a gain of 50 db. The cathodes of the tubes are energized with low-voltage 60-cycle alternating current and the anodes with rectified alternating current. A_3, the final or power amplifier, is a single stage amplifier employing two Western Electric No. 242A vacuum tubes in parallel on each side of a push-pull circuit, thus having four tubes per amplifier. Two of the A_3 amplifiers were used in parallel on each channel, and were capable of supplying 60 watts each, or a total of 120 watts, to the loud speakers. These are r.m.s. values. The instantaneous peaks of power of course could equal twice this value, or 720

watts, for the three channels. E_1 and E_2 are equalizers to compensate for any amplitude distortion that would cause a listener to obtain a different tone effect from the loud speakers than he would from the



Fig. 2—Amplifying equipment used at Philadelphia. The taller racks are 8 ft. high and contain A_1 and A_2 amplifiers, volume indicators, and various controls.

actual orchestra. These equalizers are loss networks and principally equalize for the acoustic characteristic of the loud speakers in the particular hall, but they are placed in a low energy part of the amplification circuit so as not to waste the energy of the final power stage.

In view of the inclusion of the equalizers in the amplification system, and particularly because of the fact that the amplification of the A_3 amplifier deliberately was made to increase with frequency in order to compensate in part for acoustic losses in the overall system, the actual amplification-frequency curves of the amplifiers are of little importance. The equalizers of the system are discussed in the paper by Bedell and Kerney.

Transmission Lines*

By H. A. AFFEL, R. W. CHESNUT and R. H. MILLS

Describing methods whereby high quality sound reproduction in auditory perspective can be accomplished over long distances, this discussion centers largely upon a description of the exact technique employed in providing communication transmission circuits for the Philadelphia-Washington demonstration. Problems that might be involved in carrying out such transmission on a more widespread scale also are touched upon.

MICROPHONES have been described that will pick up without noticeable distortion all the sounds given forth by a symphony orchestra. Loud speakers and amplifiers also have been described that will accurately reproduce this highest quality music in its full range of tone quality and volume. Therefore, the situation obviously requires connecting transmission paths so perfect in their characteristics that reproduction 100 or 200 miles away may not suffer in comparison with reproduction which may be only 100 or 200 feet from the source of music.

There are several respects in which a long line circuit possibly may distort the speech or music passed over it, unless considerable effort is expended to overcome these tendencies. For example, there may be frequency-amplitude distortion; i.e., all the notes and overtones may not be transmitted with the proper relative volumes. Similarly there may be phase or delay distortion, the different frequencies may not arrive at the receiving end of the line circuit in the same time relationships in which they originated. A line circuit is subject also to possible inductive disturbances from other communication circuits ("crosstalk"), or from power or miscellaneous circuits which cause "noise" at the receiving terminal. If the circuit contains amplifiers, transformers, and inductances having magnetic cores, it is subject to possible nonlinearity effects; i.e., the current at the receiving end of the line may not follow exactly the amplitude variations of the current applied to the transmitting end or, what is more important, spurious intermodulation frequencies may be generated within the transmission circuit and mar the purity of the musical tones. The problem of reproduction in auditory perspective, using two or three paralleling

* Fifth paper in the Symposium on Wire Transmission of Symphonic Music and Its Reproduction in Auditory Perspective. Presented at Winter Convention of A. I. E. E., New York City, Jan. 23–26, 1934. Published in *Electrical Engineering*, January, 1934.

channels, also adds the requirement that these channels must be substantially identical in their transmission characteristics.

With the exception of the last, all these aspects of the problem are, of course, not peculiar to symphony music transmission. They exist as part of the problem of satisfactorily transmitting any telephone message. However, the requirements of this new high quality transmission have set a new high standard of refinement, even as compared with that required for ordinary radio chain broadcasting. For example, ordinary telephone message transmission commonly is carried out by circuits having a frequency range not exceeding 200 to 3000 cycles per second. Much present-day radio broadcasting involves a transmission band only from about 100 to 5000 c.p.s. This new high quality transmission, however, requires a range from approximately 40 to 15,000 c.p.s. Further, with reference to the required freedom from interference, ordinary radio broadcasting seldom exceeds a volume range greater than 30 decibels. The new high-quality system, however, requires a volume range of at least 65 db, which is more than 3,000,000 to 1 expressed as a power ratio.

In considering the specific problem of transmitting from Philadelphia to Washington for the demonstration given on April 27, 1933, several alternative methods of providing the required transmission paths presented themselves. The arrangement chosen consisted in bridging the distance between the two cities by means of carrier channels over cable conductors. From the telephone toll office in Philadelphia to the toll office in Washington, three carrier transmission paths were provided in which the music frequencies were stepped up from their normal position in the audible range to considerably higher frequencies. The frequency range from 40 to 15,000 c.p.s. picked up by the microphones was transmitted over line circuits in a range from 25,000 to 40,000 c.p.s. After being thus stepped up in frequency, the high frequency currents were applied to three non-loaded pairs in an all-underground cable which was equipped with repeaters at approximately 25-mile intervals. At Washington, step-down or demodulation apparatus restored the frequencies to their normal position in the spectrum.

For transmission between the auditorium in Philadelphia and the toll office there, a distance of approximately three miles, and for transmission in Washington between the telephone toll office and the auditorium, about half this distance, normal frequency transmission over small-gauge pairs in ordinary exchange cables was employed.

The use of the carrier method for the long distance transmission has several advantages. In general, it permits multiplex operation; i.e., more than one message or program on the same pair of wires. As a

matter of expediency in this particular case this feature of operation was not used, and three separate pairs were employed, one for each channel. In the future the same technique undoubtedly would permit two or possibly more of these extra-broad-band transmission paths to be obtained on the same pair of conductors. The most important reason for choosing the carrier method rather than transmission in the natural audio-frequency range in this particular case was that, because all other transmission circuits in the same cable were at a considerably lower frequency and because the lead sheath of the cable acts efficiently at the high frequencies to shield the pairs from induced disturbances from the outside, it offered a special freedom from crosstalk and noise.

With these arrangements, which will be described in somewhat greater detail in what follows, requirements of transmission were met very satisfactorily and the reproduction of the symphony music in Washington with the orchestra playing in Philadelphia suffered not the least in comparison with the reproduction of the same program in an auditorium in Philadelphia located but a few feet from the hall in which the orchestra played. It is believed that, if necessary, by the use of the same principles, line circuits may be set up and comparable quality reproduction given throughout the country. However, as will be evident from part of the discussion which follows, in some respects the problem of meeting the requirements in transmission between Philadelphia and Washington was not as difficult as might be encountered in other localities. Hence even more complex arrangements might be necessary if it were desired to establish such transmission circuits to other points, and particularly for greater distances.

Line Circuits

There are several all-underground cables between Philadelphia and Washington. As described in a paper [1] by Clark and Green given before the A. I. E. E. in 1930, recently laid cables contain several 16-gauge conductors distributed throughout the cross section of the cable for possible use as program circuits in chain broadcasting. These pairs, however, ordinarily are loaded and equipped with repeaters at approximately 50-mile intervals so that they transmit a frequency range up to about 8000 c.p.s.

In one of the cables several pairs of this type had not yet been loaded, and these pairs were used for this newer transmission. Because of the higher frequencies employed and the greater attenuation encountered, it was necessary to install repeaters at more frequent intervals. As may be noted in Fig. 1, the normal cable layout between Philadelphia and Washington includes two intermediate repeater

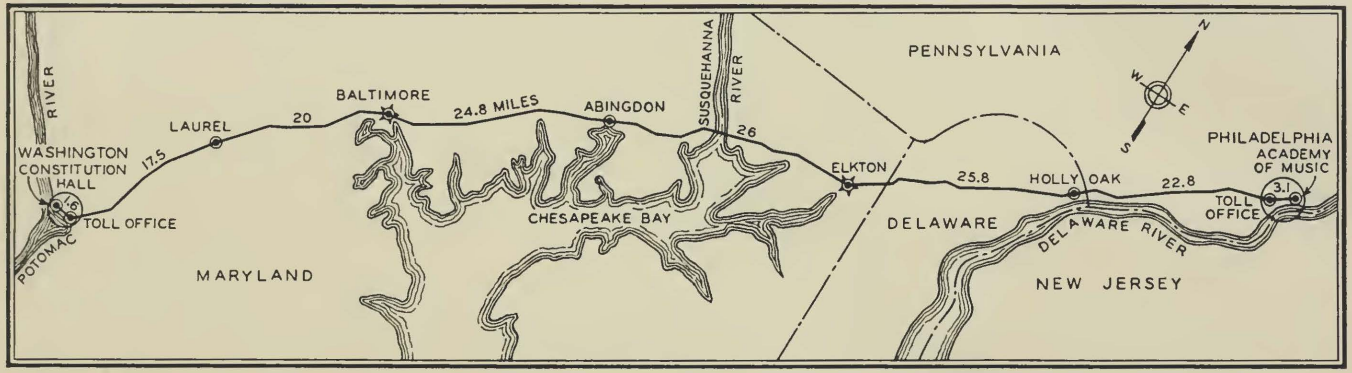


Fig. 1—Geographical layout of 3-circuit communication line used to carry Philadelphia symphony music to Washington, D. C., for reproduction in auditory perspective.

stations, one at Elkton and one at Baltimore. Additional repeater stations were established accordingly at in-between points—Holly Oak, Abingdon, and Laurel. One of these repeater points, Holly Oak, was established in a local telephone office. No such convenient housing existed at the other two points, and it was necessary to establish new repeater stations. These were small metal structures large enough to house only the repeaters, their power supply, and testing equipment. This apparatus was arranged to be normally non-attended, various switching actions being remotely controlled from the nearest regular repeater station.

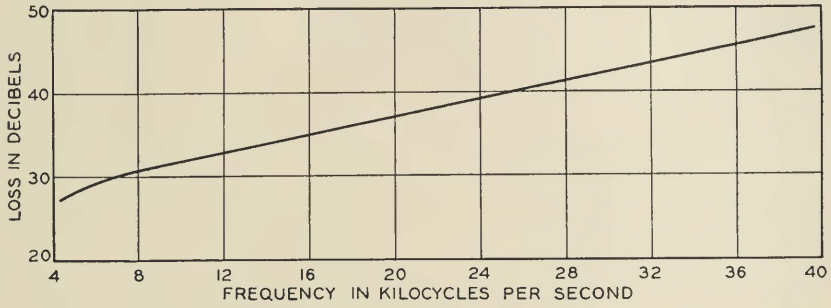


Fig. 2—Line attenuation characteristic of typical repeater section.

The line attenuation between repeater points is shown in Fig. 2. It may be noted that the attenuation is approximately 50 db for the highest carrier frequency involved. A diagram showing the variations in power level as the carrier waves traverse the complete circuit is shown in Fig. 3. Because of the variation in attenuation over the

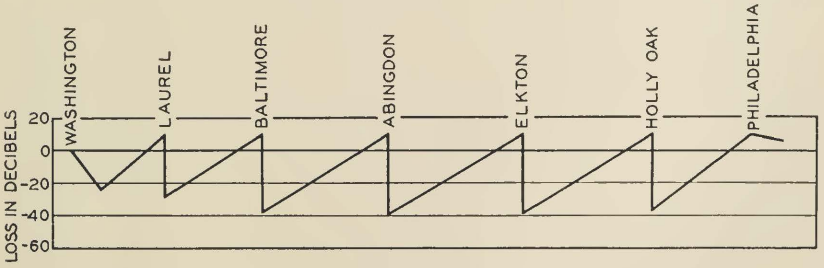


Fig. 3—Transmission level diagram.

frequency range employed it was necessary, of course, to use equalizers at the input of each repeater; i.e., networks having an attenuation variation with frequency approximately the inverse of that of the line circuit.

Noise

In setting up these circuits various tests, including measurements of noise currents picked up by the conductors to be employed, were made prior to the actual installation of the apparatus. It was discovered that on the cable circuits north of Baltimore these pairs were picking up sufficient noise even at the higher frequencies to constitute a possible limitation in the volume range that might be delivered. This noise was generated chiefly as a by-product of relay and other similar operations within the Baltimore office and was propagated over the longitudinal circuits of various pairs in the cable from which, by induction, it entered the special selected pairs. As a remedial measure, longitudinally acting choke coils applied to all but the specially selected pairs in the cable greatly reduced the noise. Shielding and physical separation were employed in the Baltimore office to prevent induction between the repeaters and the connection to the main cable. If it is desired to use existing cables for carrier transmission, particularly for such high grade transmission circuits, it seems likely that filtering arrangements of this kind, or other precautions, generally will be required.

Carrier Apparatus

The carrier system employed may be characterized briefly as single-sideband carrier-suppressed, with perfectly synchronized carrier frequencies of 40,000 c.p.s. Most present-day commercial telephone carrier systems are of the single-sideband carrier-suppressed type. Suppressing one sideband saves frequency space and suppressing the carrier reduces the load on the line amplifiers or repeaters. Ordinarily the exact synchronization of the carrier frequencies at the sending and receiving ends is not required for message telephone service.

Obtaining a single sideband after modulation commonly is carried out by providing band filters which transmit the desired sideband and suppress the unwanted sideband. For the requirements of message telephone transmission this does not impose severe requirements in the design of filters because audio frequencies less than about 100–200 c.p.s. ordinarily are not transmitted, in which case, if the filter in suppressing the unwanted sideband tends to cut off the lower frequencies of the desired sideband, it is not important.

For the requirements of this new high quality system, however, where it was desired to transmit all frequencies to at least as low as 40 c.p.s., the problem was considerably more difficult. Two alternatives presented themselves in the design of the required filters. The first consisted in attempting to provide the required selectivity in the filters themselves, perhaps supplementing the actions of induc-

tance coils and condensers (which normally make up such a filter structure) by quartz crystals to provide the sharp selectivity required on the sides of the band. The other alternative consisted in providing a filter of moderate selectivity so that in the neighborhood of the carrier frequency the unwanted sideband is not completely suppressed, and in arranging that the resultant reproduced music at the receiving terminal is obtained by the proper coordination of the desired and the vestige of the unwanted sideband. The "vestigial" sideband method was decided upon. Although this does not require filters having particularly sharp selectivity on the sides of the band, it does, however, impose more severe requirements upon the control of the phase characteristics of the filters in the neighborhood of the carrier frequency. It makes it necessary also to have the carrier frequencies at the sending and receiving ends not only synchronized, but phase controlled as described later.

For the modulating elements in the system at both the sending and receiving terminals, copper oxide rectifying disks were chosen. These elements can be made very simple. In stability, with respect to transmission loss and the ability to suppress the unwanted carrier frequency by balanced circuits, this arrangement is superior to the usual vacuum tube circuits.

In Fig. 4 is shown schematically the arrangements of the carrier circuit at the transmitting and receiving ends. At the transmitting terminal the circuit from the microphones is led first through low- and high-pass filters to limit the bands to the desired width; i.e., 40 c.p.s. to 15,000 c.p.s. The 40-cycle limiting filter was included because tests had demonstrated that lower frequencies are not required for the satisfactory transmission of music of symphony character, and because it was feared that occasional high energy pulses of subaudible frequency might cause overloading. When these 40-cycle filters are omitted, as was done in tests, the carrier channels are capable of transmitting frequencies down to and including zero frequency, a characteristic which could not possibly be obtained in a single sideband system by other than such a vestigial sideband technique.

As may be noted further in Fig. 4, carrier current is supplied to the rectifying disks of the modulator along with the incoming music frequencies. The balancing connection of the four rectifying disks making up the modulator is arranged to suppress the carrier frequency, the final degree of suppression being adjusted by means of the variable condenser and resistance shown, which were included to make up for slight dissimilarities in the characteristics of the individual copper disks. A very high degree of carrier suppression can be achieved by

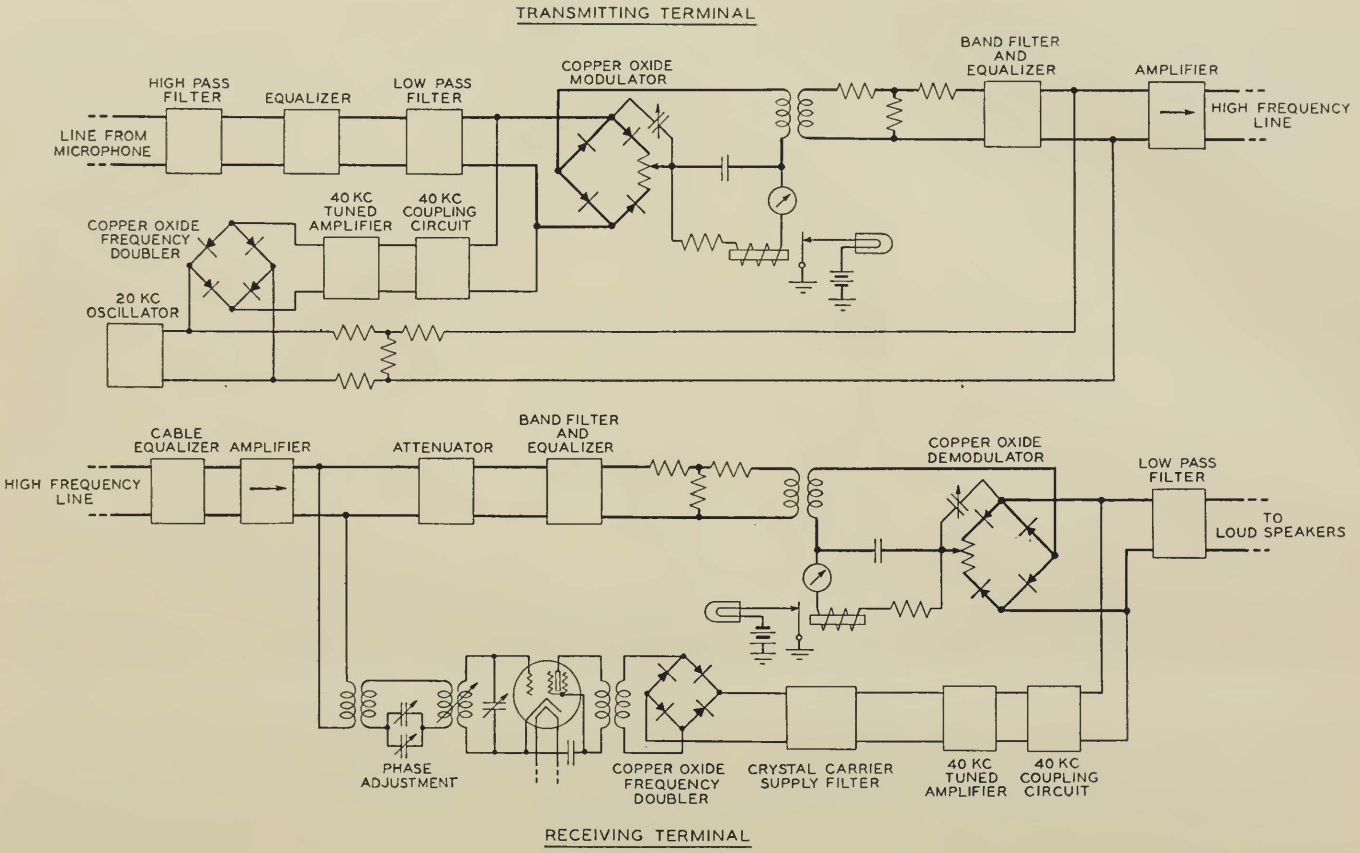


Fig. 4—Schematic diagram of carrier terminal circuits.

this means. No difficulty was experienced in maintaining a ratio of at least 60 db between the carrier voltage applied to the unit and the residual carrier current not completely balanced out. Over short periods an even higher degree of balance can be readily obtained.

There is a certain amount of electrical noise generated in the rectifying disks over and above that caused by thermal agitation [2, 3] effects. The amount of this noise compared with the maximum permissible modulation output determines the volume range possibilities of a modulator of this type. Measurements indicated that this range was approximately 90 db, which obviously was more than sufficient to meet the requirements desired.

The circuit includes a relay and a meter through both of which flows the d.-c. component produced by the rectification of the carrier frequency. These supplementary units give a check on the magnitude of the carrier supply and afford an alarm in case of failure. From the modulator unit the circuit is connected to the band filter which transmits only the lower sideband lying between approximately 25,000 and 40,000 c.p.s. and the vestige of the upper sideband. From the band filter the currents are led to an amplifier and thence to the line circuit leading to the farther terminal.

It may be noted that at the transmitting terminal the 40,000-cycle carrier current is derived from a 20,000-cycle oscillator by passing its output through a series of copper oxide rectifiers connected to form a frequency doubler. Part of the originally generated 20,000 cycles also is connected to the input of the transmitting amplifier and sent over the line to be used in producing the 40,000-cycle carrier supply for demodulation.

At the receiving terminal a similar modulation or demodulation process occurs through the use of copper oxide disk circuits. A relay and meter also are included in the circuit to check the carrier supply, in this case providing also a check or pilot of the transmission over the long line circuit. The 20,000-cycle synchronizing current is selected at the receiving terminal, amplified and applied to a frequency doubler, and thence applied to the demodulator circuit. The input of this carrier supply circuit includes also a phase adjusting variable condenser arrangement so that the phase of the carrier supplied to the demodulator may be adjusted properly in relation to that of the carrier supplied to the modulator at the sending end. An interesting feature of the receiving terminal carrier supply is the quartz crystal filter employed to select the 40,000-cycle carrier after frequency doubling. The transmission characteristic of this extremely selective filter is shown in Fig. 5.

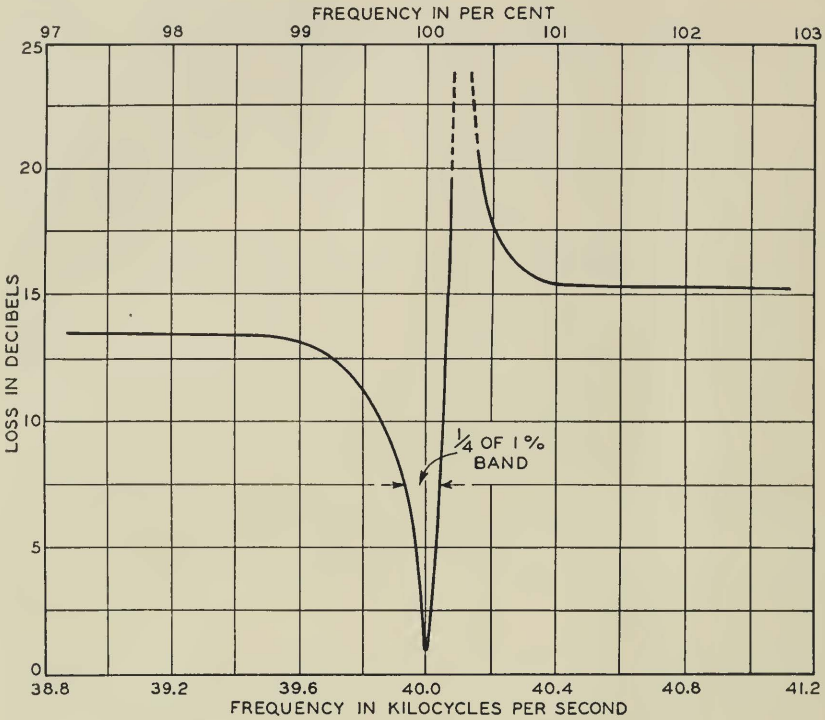


Fig. 5—Transmission characteristic of carrier supply crystal filter.

Filters

The transmission characteristics of the carrier channels are determined largely by the filters and associated equalizers. The filters principally affecting transmission are the band filters. Identical units are employed at the sending and receiving ends. The transmission and phase shift characteristics of one of these units are shown in Fig. 6. These band filters are equalized to produce the desired squared band characteristic.

The characteristics in the frequency region near the carrier (i.e., at 40,000 cycles) are shown on a large scale. This region is of particular interest because it is here that the degree of success in the application of the vestigial sideband method, for the purpose of insuring the satisfactory transmission of the low music frequencies, is determined. If for a given frequency interval above the carrier the phase change is arranged to be equal and opposite to that of the same frequency interval below the carrier, then in the action of demodulation the demodulated current produced by the action of one sideband adds

itself arithmetically to that produced by the other sideband. It will be noted that this desirable phase characteristic has been achieved closely in the characteristics shown. If, in addition, the attenuation

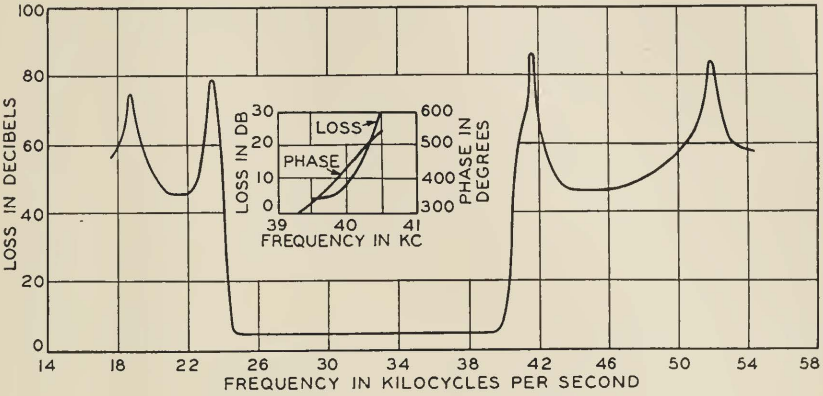


Fig. 6—Transmission and phase characteristics of band filter.

loss in the filter is adjusted so that the sum of the regular and vestigial sideband amplitudes corresponding to the low music frequencies is substantially constant and equal to the amplitude of the frequencies at midband, the desired flat transmission characteristic is assured.

As was noted previously, this action can be carried out only if the phase angle of the receiving carrier is properly related to that of the sideband frequencies and, in turn, to the carrier applied to the modulator at the transmitting terminal. The curves shown in Fig. 7

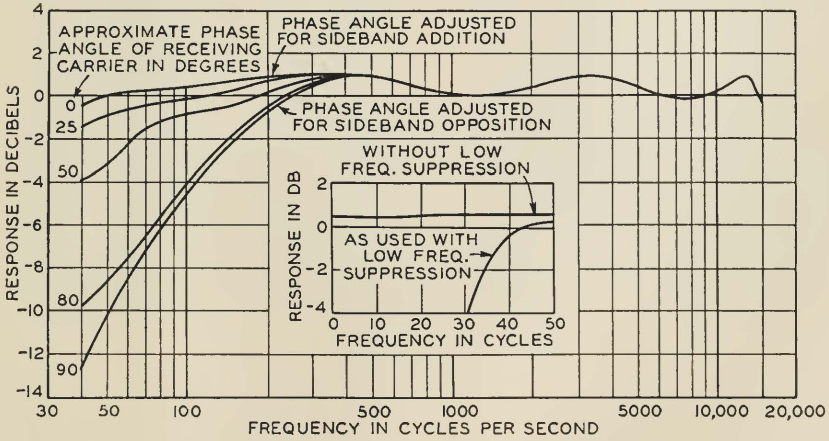


Fig. 7—Over-all transmission characteristics as a function of the phase relation of the receiving carrier.

illustrate the influence that the phase adjustment of the carrier frequency has on the transmission of the lower frequencies in a system of this kind.

The upper curve shows the transmission frequency characteristic of one of the carrier channels measured from terminal to terminal between distortionless lines, when the phase angle of the receiving carrier is adjusted for its optimum value. Under these conditions the vestigial sideband and normal sideband supplement each other in their effects to produce substantially flat transmission. (The insert indicates the sustained transmission toward zero frequency when the 40-cycle highpass filter is omitted from the circuit.) It may be noted also that with this proper phase adjustment the full band transmission characteristic provided is substantially flat within a fraction of a decibel from 40 c.p.s. to 15,000 c.p.s. The lower curves indicate successively what happens if the phase angle of the receiving carrier is adjusted different amounts from the optimum adjustment. It may be noted that for a 90-degree departure the transmission of a 40-cycle tone over the carrier channel would suffer more than 12 db in comparison with a 1000-cycle tone.

Repeaters

As noted previously, the line circuit between Philadelphia and Washington included five intermediate repeater points. A schematic drawing of the apparatus installed at each point is shown in Fig. 8.

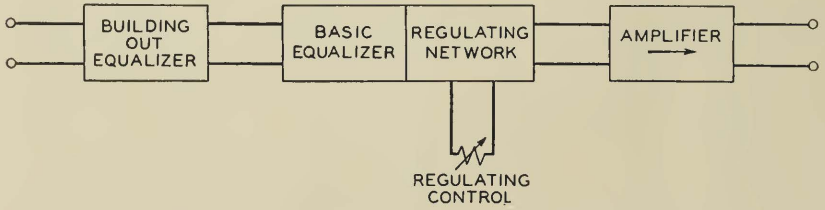


Fig. 8—Schematic diagram of repeater station apparatus.

The amplifiers at these points, as well as those used at the transmitting and receiving terminal, consisted of a new form of amplifier employing the principle of negative feed-back. The principal virtues of amplifiers of this type are their remarkable stability with battery and tube variations and great freedom from nonlinearity or modulation effects. Each amplifier is supplemented at its input by an equalizer designed to have its attenuation approximately complementary in loss to that of the line circuit in a single section. The amplifiers actually employed for the purpose were taken from a trial of a cable

carrier system described in a recent A. I. E. E. paper by A. B. Clark and B. W. Kendall.[4]

The losses in the cable circuits do not, of course, remain absolutely constant with time, and slow variations due to change of temperature are compensated for by occasional adjustments of the variable equalizer arrangements provided. These adjustments were required only infrequently; approximately at weekly intervals because in an underground cable the temperature experiences only slow, seasonal variations.

As noted, new repeater stations were established at two points. The housing arrangements for one of these points, Abingdon, is shown in Fig. 9. The equipment at this repeater point also included relays remotely controlled from the nearest attended repeater station to permit the repeaters to be turned on and off at will and the power supply, which consisted of storage batteries, to be switched from the regular to the reserve battery or either battery put on charge if required.

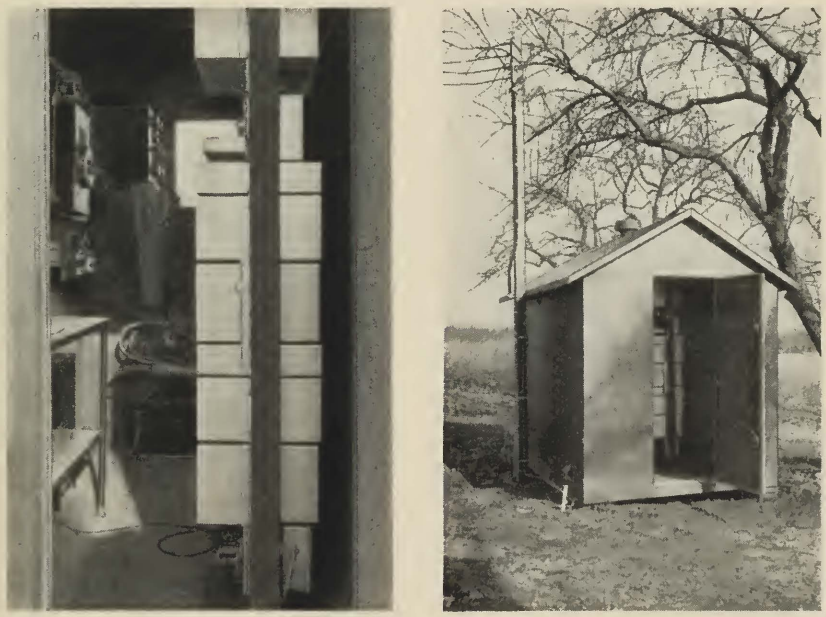


Fig. 9—Interior and exterior of special intermediate repeater station at Abingdon, Md.

Over-All Performance

While the system was set up specifically to provide transmission for the demonstration into Washington on April 27, 1933, it was operated over a period of several weeks and complete tests and measurements were carried out for the purpose of gathering information on cable carrier systems. The complete layout of apparatus and lines provided between Philadelphia and Washington is shown in Fig. 10.

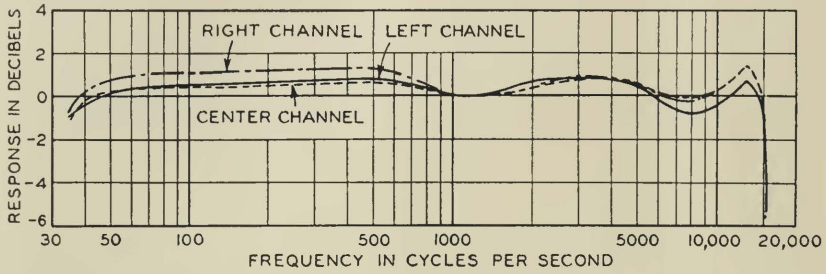
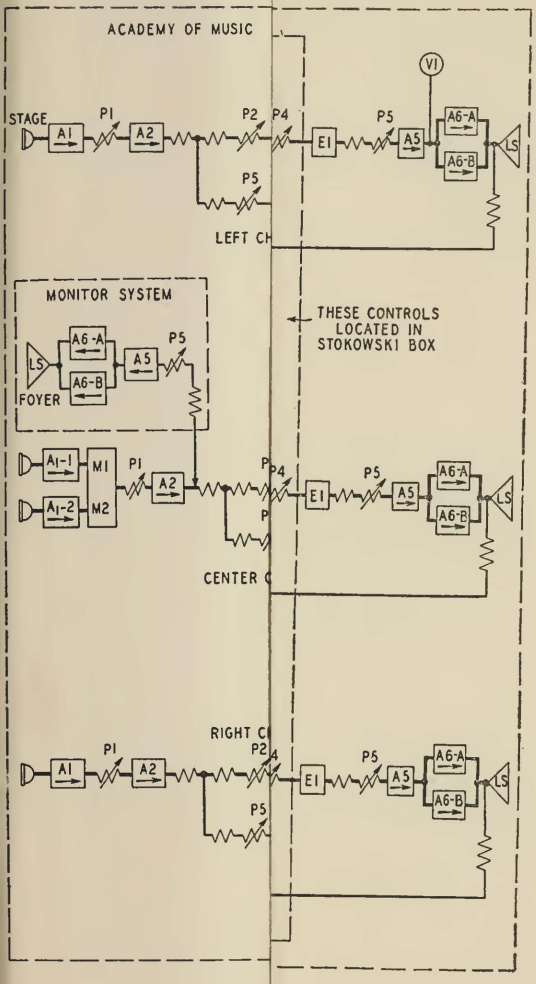


Fig. 11 (right)—Frequency characteristics over carrier channels used between Philadelphia and Washington, D. C.

The over-all frequency transmission characteristics of the three channels that were set up are shown in Fig. 11. These curves differ from those shown in Fig. 7, and include the complete high frequency line circuit with its 150 miles of cable, repeaters, equalizers, and other equipment. It may be seen that between the desired frequency limits the circuit is substantially flat in transmission performance to within $\pm$ 1 db. Various noise measurements made on the over-all circuit indicated that the circuits fully met the requirements that had been set up, and that the line and apparatus noise was inaudible in the auditorium at Washington even during the weakest music passages. The circuit also was found to be free from nonlinear distortion to a satisfactory degree. Harmonic components generated when single-frequency tones were applied to the channels at high volumes were found with one unimportant exception to be more than 40 db below the fundamental.

As a means of obtaining a further increase in volume range, which was not actually required for this demonstration, tests were made with a so-called predistortion-restoring technique. In this the higher frequency components of the music were transmitted over the carrier channels at a volume much higher than normal in relation to the volume of the lower frequencies. By this means any noise entering the carrier channels at frequencies equivalent to the higher music frequencies is greatly minimized in effect.


ACADEMY OF MUSIC
VI
STAGE
A1
P1
A2
P2
P4
E1
P5
A5
A6-A
A6-B
LS
P5
LEFT CH
MONITOR SYSTEM
A6-A
P5
LS
A5
A6-B
FOYER
THESE CONTROLS
LOCATED IN
STOKOWSKI BOX
A1-1
M1
P1
A2
A1-2
M2
P4
E1
P5
A5
A6-A
A6-B
LS
CENTER C
RIGHT C
A1
P1
A2
P2
P4
E1
P5
A5
A6-A
A6-B
LS
P5

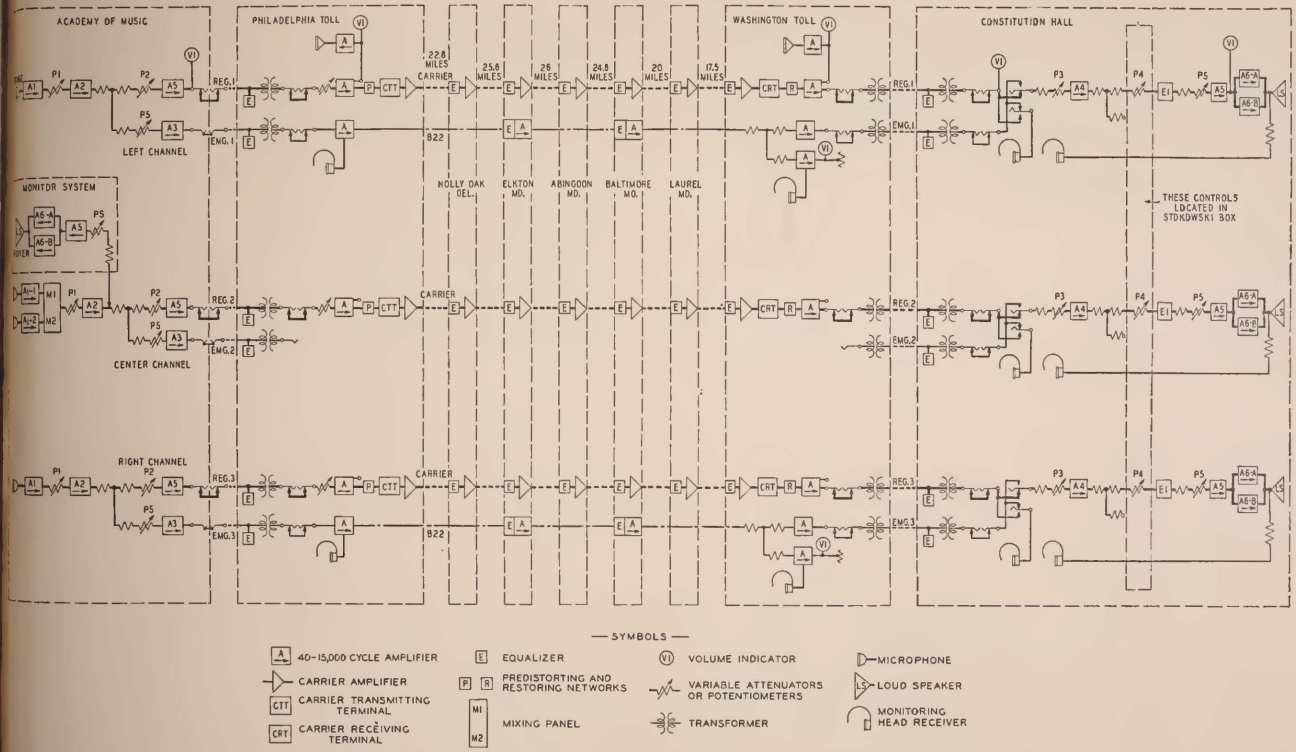


Fig. 10—Schematic diagram of circuit layout for 15-kc channel used for symphonic program demonstration.

This predistortion is accomplished by including in the circuit at the input to the modulator a network having relatively high loss for the lower frequencies and tapering to low loss for the higher frequencies. Its maximum loss is compensated for by adding in the circuit an equivalent amount of additional amplification. The characteristics of such a network are illustrated in Fig. 12. To restore the normal volume relationships between the different tones and overtones a

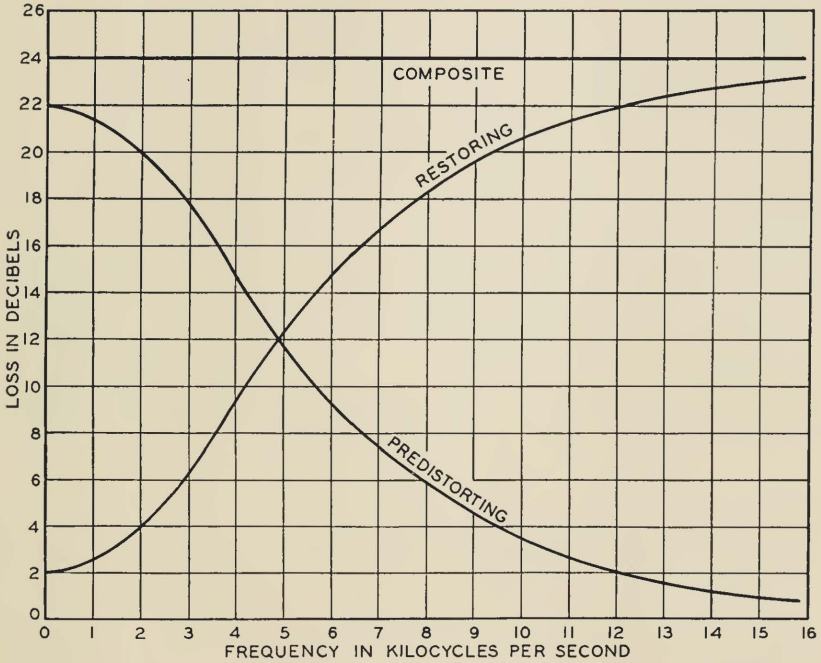


Fig. 12—Attenuation characteristics of "predistorting and restoring" networks.

restoring network having complementary transmission frequency characteristics is, of course, included at the output of the receiving circuit. It was found with this predistortion-restoring technique that a volume range increase of something like 10 db could be obtained over the circuits described.

There is available also another method which might have been employed for obtaining a further increase in volume range. This method, the so-called volume compression-expansion system, very likely will be necessary if in the future it is desired to obtain such high quality circuits on long routes where the carrier frequency range is being used also for regular telephone message transmission or for other purposes, and where the problem of freedom from noise and crosstalk

This predistortion is accomplished by including in the circuit at the input to the modulator a network having relatively high loss for the lower frequencies and tapering to low loss for the higher frequencies. Its maximum loss is compensated for by adding in the circuit an equivalent amount of additional amplification. The characteristics of such a network are illustrated in Fig. 12. To restore the normal volume relationships between the different tones and overtones a

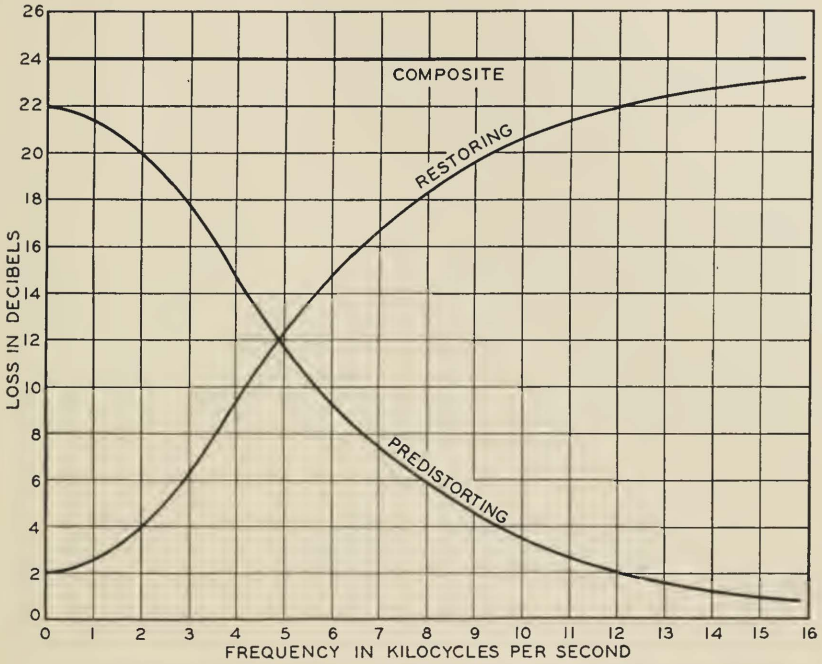


Fig. 12—Attenuation characteristics of "predistorting and restoring" networks.

restoring network having complementary transmission frequency characteristics is, of course, included at the output of the receiving circuit. It was found with this predistortion-restoring technique that a volume range increase of something like 10 db could be obtained over the circuits described.

There is available also another method which might have been employed for obtaining a further increase in volume range. This method, the so-called volume compression-expansion system, very likely will be necessary if in the future it is desired to obtain such high quality circuits on long routes where the carrier frequency range is being used also for regular telephone message transmission or for other purposes, and where the problem of freedom from noise and crosstalk

no doubt will be more serious than experienced in the Philadelphia-Washington demonstration. Such a volume compression-expansion system requires additional apparatus at the sending and receiving terminals of the line circuit. At the sending end this apparatus is used to raise in volume the weak passages of the music or other program for transmission over the line circuits in order that the proper ratio between the desired program and unwanted noises may be retained. At the receiving terminal coordinating apparatus reexpands the compressed volume range to the volume range originally applied to the transmitting terminal.

In the demonstration, to provide supplementary control features required by Dr. Stokowski at Washington for communicating with the orchestra at Philadelphia, additional wire circuits were established between these points. Order wire circuits also were provided for communication between the terminals and repeater points to make possible the location troubles if any should arise. Rather elaborate switching means were included at the terminals to permit switching the carrier channels to different microphones and to different amplifier equipment at the loud speaker end. To take care of the contingency of a cable pair failure, spare pairs of wires were made available to be switched in at short notice. Fortunately, none of the reserve facilities actually were required for the demonstration.

References

1. "Long Distance Cable Circuits for Program Transmission," A. B. Clark and C. W. Green. *A. I. E. E. Trans.*, v. 49, 1930, p. 1514–23.
2. "Thermal Agitation of Electricity in Conductors," J. B. Johnson. *Phys. Rev.*, v. 32, 1928, p. 97.
3. "Thermal Agitation of Electric Charge in Conductors," H. Nyquist. *Phys. Rev.*, v. 32, 1928, p. 110.
4. "Carrier in Cable," A. B. Clark and B. W. Kendall. *Elec. Engg.*, July 1933, p. 477–81.

System Adaptation*

By E. H. BEDELL and IDEN KERNEY

A communication system for the pick-up and reproduction in auditory perspective of symphonic music must be designed properly with respect to the acoustics of the pick-up auditorium and the concert hall involved. The reverberation times and sound distribution in the two auditoriums, the location of the microphones and loud speakers, and the response-frequency calibration of the system and its equalization are considered. These and other important factors entering into the problem are treated in this paper.

WHEN the effect of music or the intelligibility of speech is spoiled by bad acoustics in an auditorium, the audience is well aware that acoustics do play a most important part in the appreciation of the program. One may not be conscious of this fact when the acoustical conditions are good, but a simple illustration will show that the effect still is present. Thus, of the sound energy reaching a member of the audience as much as 90 per cent may have been reflected one or more times from the various surfaces of the room, and only 10 per cent received directly from the source of the sound.

In listening to reproduced sound in an auditorium or concert hall, the effect of the room acoustics is perhaps even more important, for in this case the audience does not see any one on the stage and must rely entirely upon the auditory effect to create the illusion of the presence there of an individual or a group. Imperfections in the reproduced sound that are caused by defects in the acoustics of the auditorium may destroy the illusion and be ascribed improperly to the reproducing system itself.

In some types of reproduced sound, radio broadcast for example, where the reproduction normally takes place in a small room, the attempt is made to create the illusion that the listener is present at the source.[1, 2] In the case considered here, however, where symphonic music is reproduced in a large auditorium, the ideal is to create the illusion that the orchestra is present in the auditorium with the audience. Since the orchestra is playing in one large room and the music is heard in another, the acoustical conditions prevailing in both must be considered.

* Sixth and final paper in the Symposium on Wire Transmission of Symphonic Music and Its Reproduction in Auditory Perspective. Presented at Winter Convention of A. I. E. E., New York City, Jan. 23–26, 1934. Published in *Electrical Engineering*, January, 1934.

Pick-Up Conditions

The source room is the auditorium of the American Academy of Music in Philadelphia. This room has a volume of approximately 700,000 cubic feet, and a seating capacity of 3000. Measured reverberation time curves for this auditorium, and preferred values [3, 4] for a room of this volume, are given in Fig. 1. It may be seen that with a

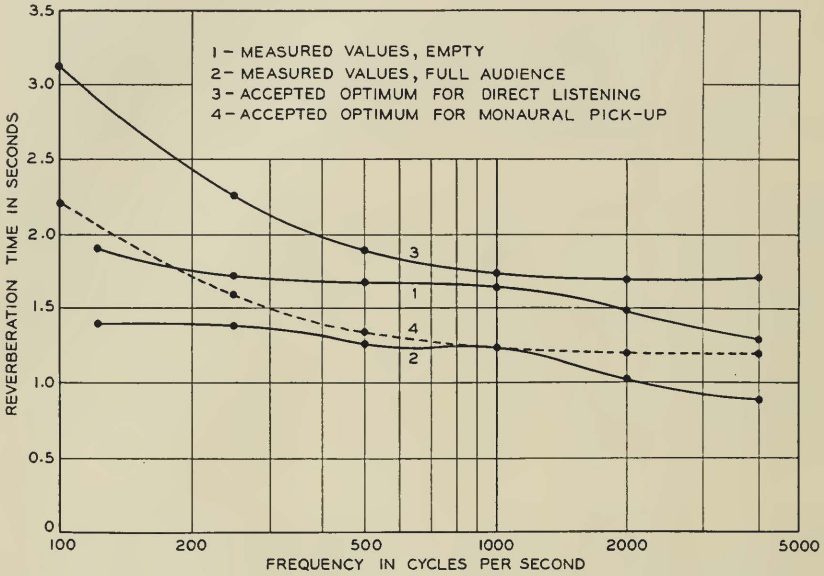


Fig. 1—Reverberation characteristics of Academy of Music, Philadelphia, Pa.

full audience this room might be considered somewhat dead, but would be considered generally satisfactory for pick-up either with or without an audience. A floor plan of the Academy auditorium and stage, showing the location of the three microphones used, is given in Fig. 2. The microphone positions were selected after judgment tests using several locations and are much nearer the orchestra than they would be for single channel pick-up.[2] The use of the microphones near the orchestra results in picking up a high ratio of direct to reverberant sound and thus reduces the effect of reverberation in the source room upon the reproduced music. A high ratio of direct sound is desirable in the present case also because of the use of three channels. The perspective effect obtained with three channels depends to a considerable extent upon the relative loudness at the three microphones, and since the change in loudness with increasing distance from the source is marked for the direct sound only, and not for the reverberant, there would be

a definite loss in perspective effect if the microphones were placed at a greater distance from the orchestra. This effect is discussed more fully in another paper of this symposium.

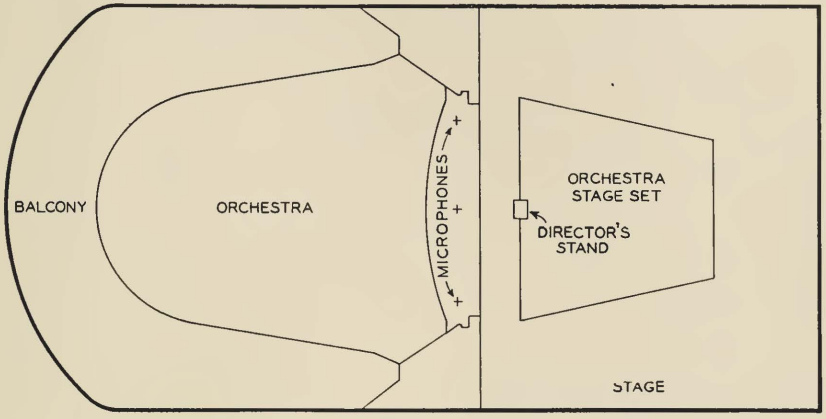


Fig. 2—Floor plan of Academy of Music, showing location of microphones.

With the microphones located close to the orchestra their response-frequency characteristics will be essentially those given by the normal field calibration, since relatively little energy is received from the sides and back. For a distant microphone position it would be necessary to use the random incidence response characteristic, which differs from the normal because of the variation in directional selectivity of the microphones as the frequency varies. This difference in response characteristic depends upon the size of the microphone and may amount to as much as 10 db at 10,000 c.p.s. It may be pointed out here that this difference in response is one factor frequently overlooked in the placement of microphones.

In addition to the three microphones regularly used, a fourth was provided to pick up the voice when a soloist accompanied the orchestra. In this case only the two side channels were used for the orchestra, the voice being transmitted and reproduced over the center channel. The solo microphone was so shielded by a directional baffle that it responded mainly to energy received from a rather small, solid angle. This arrangement permitted independent volume and quality control for the vocal and orchestral music.

The Concert Hall

The music was reproduced before the audience in Constitution Hall in Washington, D. C. This hall has a volume of nearly 1,000,000

cubic feet, and a seating capacity of about 4000. A floor plan of the auditorium showing the location of the loud speakers and of the control equipment is given in Fig. 3. The loud speakers are placed so that each of the three sets radiates into a solid angle including as nearly

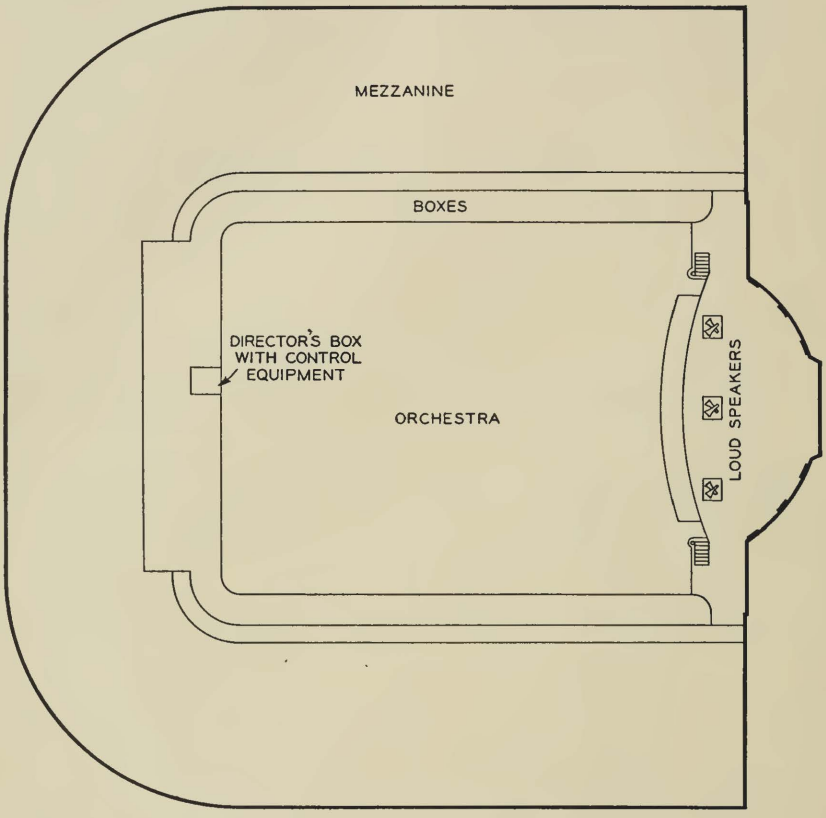


Fig. 3—Floor plan of Constitution Hall, Washington, D. C., showing locations of loud speakers.

as possible all the seats of the auditorium. Figure 4 shows the reverberation-frequency characteristics of Constitution Hall. The values given by the curve for the empty hall were measured through the use of the three regular loud speakers and several microphone positions in the room. The values for the hall with an audience present were calculated from known absorption data for an audience, and the optimum values are taken from accepted data for an auditorium of the volume of this one.[3] The reverberation times were considered satisfactory and no attempt was made to change them for this demonstration. The

reverberation time measurements for both Constitution Hall and the Academy of Music were made with the high speed level recorder.[5] This instrument measures and plots on a moving paper chart a curve

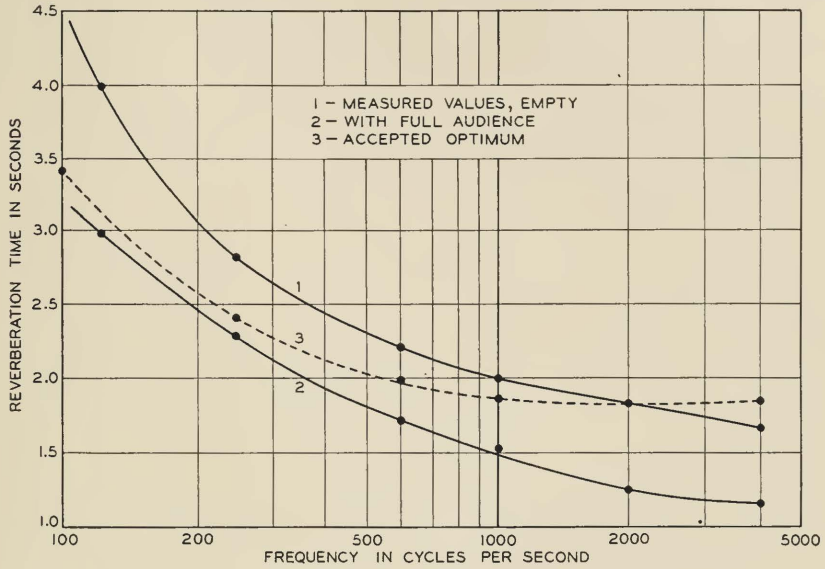


Fig. 4—Reverberation characteristics of Constitution Hall.

the ordinate of which is proportional to the logarithm of the electrical input furnished to it. When used in connection with a microphone for reverberation time measurements, curves are obtained showing the intensity of sound at the microphone during the period of sound decay. The rates of decay, and hence the reverberation times, are obtainable immediately from the slopes of these recorded curves and the speed of the paper chart.

Calibration of the System

In calibrating the system, a heterodyne oscillator connected to the loud speakers through the amplifiers was used. The oscillator was equipped with a motor drive to change the frequency, and as the frequency was varied through the range from 35 to 15,000 c.p.s. the sound was picked up with a microphone connected to the level recorder. Continuous curves of microphone response as a function of frequency thus were obtained for several positions in the auditorium, and for each channel independently. These response curves provided a check on a uniform coverage of the audience by each loud speaker, and also provided data for the design of the equalizing networks required to

give an over-all flat response-frequency characteristic. If the system, including the air path from the loud speakers to one position in the auditorium, is made flat, it will not, in general, be flat for other positions or for other paths in the room. This variation in characteristic is due partly to the variation in the ratio of direct to reverberant sound, and partly to the fact that the sounds of higher frequency are absorbed more rapidly by the air during transmission.[6, 7] This latter effect is of considerable importance; it depends upon the humidity and temperature of the air, and may cause a loss of more than 10 db in the high frequencies at the more distant positions in a large auditorium. Some compromise in the amount of equalization employed therefore is necessary. Probably the most straightforward procedure would be to design the networks according to the response curves obtained with the microphones near the loud speakers. This would insure that for both the response measurements and the pick-up the microphone characteristics would be the same, and any deviation from a uniform response in the microphones would be corrected for in this way, along with variations in the loud speaker output. This procedure was modified somewhat for the case under discussion, however, because by far the greater portion of the audience was at a distance from the stage such that they received a relatively large ratio of reverberant sound, and it was believed that a better effect would be achieved by equalizing the system characteristic in accordance with response measurements taken at some distance from the loud speakers.

Control Equipment

In addition to the equalizing circuits used to obtain a uniform response characteristic, two sets of quality control networks which could be switched in or out of the three channels simultaneously were employed. One set modified the low frequencies as shown at *A*, *B*, and *C* of Fig. 5, while the other gave high frequency characteristics as shown at *D*, *E*, *F*, and *G*. These latter networks permitted the director to take advantage of the fact that the electrical transmission and reproduction of music permits the introduction of control of volume and quality which can be superimposed on the orchestral variations. Quality of sound can be divorced from loudness to a greater degree than is possible in the actual playing of instruments, and the quality can be varied while the loudness range is increased or decreased. Electrical transmission therefore not only enlarges the audience of the orchestra, but also enlarges the *capacity* of the orchestra for creating musical effects.

The quality control networks and their associated switches were

mounted in a cabinet (Fig. 6) at the right side of the director's position. Continuously variable volume controls for the three channels were mounted on a common shaft and housed in the center cabinet of Fig. 6.

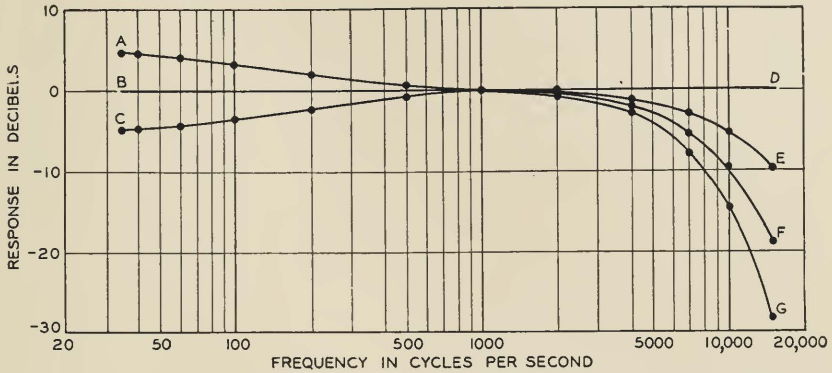


Fig. 5—Transmission characteristics of quality control networks used in the Philadelphia-Washington experiment.

A separate control for the center channel was provided when that was used for the soloist. In addition to the high quality channels certain auxiliary circuits were supplied to aid the smoothness of performance. Supplementing the order wire connecting all technical operators, a monitor circuit was provided in the reverse direction. The microphone was located on the cabinet before the director, and loud speakers were connected in the control rooms and on the stage with the orchestra, enabling the control operator to hear what went on in the auditorium and allowing the director to speak to the orchestra. Two useful signal circuits were employed; one giving the orchestra a "play" or "listen" signal, and at the same time connecting either the auditorium or the orchestra's loud speakers, respectively; the other being a

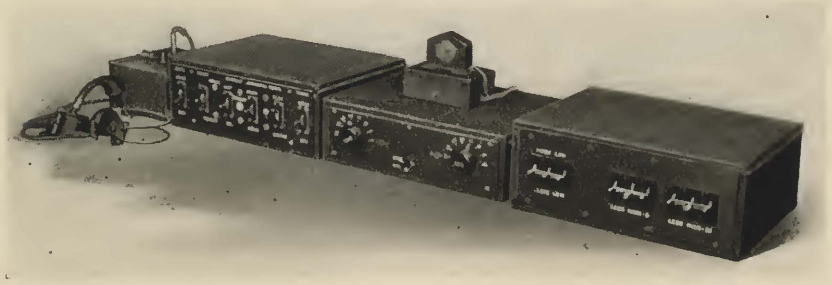


Fig. 6—Cabinets housing quality control networks and providing communication facilities for operation.

"tempo" signal to the assistant director leading the orchestra that could be operated during the rendition of the music. The switches for the auxiliary circuits and the order wire subset are shown at the control operator's position at the left in Fig. 6.

That a reproducing system may have quite different characteristics in different auditoriums is well illustrated in the case of the two halls considered here. From Fig. 3 it may be seen that in Constitution Hall the stage is built into the auditorium itself, and that there is no back stage space. The Academy of Music, however, has a large volume back stage. When the orchestra plays in the Academy the reflecting shell shown in Fig. 2 is used to concentrate the radiated sound energy toward the audience. When the reproducing system was set up in the Academy the shell could not be used because of the stage and lighting effects desired, and a large part of the energy radiated by the loud speakers at the low frequencies was lost back stage. The loss of low frequency energy is attributable partly to the fact that the loud speakers cannot well be made as directional for the very low frequencies as for the higher. The loss amounts to about 10 db at 35 c.p.s., and becomes inappreciable at 300 c.p.s. or more, as measured in comparable locations in the two auditoriums. This difference in characteristics emphasizes the fact that for perfect reproduction the acoustics of the auditorium must be considered as a part of the system, and that in general the equalizing networks must have different characteristics for different auditoriums.

References

1. "Acoustics of Broadcasting and Recording Studios," G. T. Stanton and F. C. Schmid. *Jour. Acous. Soc. Am.*, v. 4, No. 1, part 1, July 1932, p. 44.
2. "Acoustic Pick-Up for Philadelphia Orchestra Broadcasts," J. P. Maxfield. *Jour. Acous. Soc. Am.*, v. 4, No. 2, Oct. 1932, p. 122.
3. "Optimum Reverberation Time for Auditoriums," W. C. MacNair. *Jour. Acous. Soc. Am.*, v. 1, No. 2, part 1, Jan. 1930, p. 242.
4. "Acoustic Control of Recording for Talking Motion Pictures," J. P. Maxfield. *Jour. S. M. P. E.*, v. 14, No. 1, Jan. 1930, p. 85.
5. "A High Speed Level Recorder for Acoustic Measurements," E. C. Wente, E. H. Bedell, K. D. Swartzel, Jr. Unpublished paper presented before the Acous. Soc. Am., May 1, 1933.
6. "The Effect of Humidity Upon the Absorption of Sound in a Room, and a Determination of the Coefficients of Absorption of Sound in Air, V. O. Knudson. *Jour. Acous. Soc. Am.*, v. 3, No. 1, July 1931, p. 126.
7. "Absorption of Sound in Air, in Oxygen, and in Nitrogen—Effects of Humidity and Temperature," V. O. Knudson. *Jour. Acous. Soc. Am.*, v. 5, No. 2, Oct. 1933, p. 112.

Abstracts of Technical Articles from Bell System Sources

Effects of Rectifiers on System Wave Shape.[1] P. W. Blye and H. E. Kent. Operation of mercury arc rectifiers generally results in increased harmonic currents in the rectifier supply circuits and may result in increased harmonic voltages. While these harmonics usually are not serious from the standpoint of the power system, they may result in interference to communication circuits exposed to the power circuits. This paper presents a method of computing these harmonic voltages and currents, and discusses methods of coordinating telephone systems and a-c. power systems supplying rectifiers.

Joint Use of Poles with 6,900-Volt Lines.[2] W. R. Bullard and D. H. Keyes. A plan has been developed for joint occupancy of poles by power and telephone circuits in the Staten Island, N. Y. area, involving 6,900-volt distribution. The aim of this plan is to secure to the public and to the power and telephone companies over-all safety, convenience, and economy. Results of this cooperative study of joint use are presented in this paper.

Sound Film Printing—II.[3] J. Crabtree. The production of sound-film prints from variable density negatives by the Model D Bell & Howell printer has been studied from the point of view of high-frequency response and uniformity of product. The account of this study, begun in Part I, is continued here, with particular reference to the degree of influence of slippage on the high-frequency response, occasioned particularly by non-conformity of the perforation pitch of the negative and positive films. It is found that to improve printing conditions in practice, it is first necessary to achieve consistency in the pitch of the processed negative and positive materials and to make the pitch of the processed negative 0.0004 inch less than that of the positive raw stock.

The Determination of the Direction of Arrival of Short Radio Waves.[4] H. T. Friss, C. B. Feldman, and W. M. Sharpless. In this paper are described methods and technique of measuring the direction with

[1] *Elec. Engg.*, January, 1934.
[2] *Elec. Engg.*, December, 1933.
[3] *Jour. S. M. P. E.*, February, 1934.
[4] *Proc. I. R. E.*, January, 1934.

which short waves arrive at a receiving site. Data on transatlantic stations are presented to illustrate the use of the methods. The methods described include those in which the phase difference between two points constitutes the criterion of direction, and those in which the difference in output of two antennas having contrasting directional patterns determines the direction. The methods are discussed first as applied to the measurement of a single plane wave. Application to the general case in which several fading waves of different directions occur then follows and the difficulties attending this case are discussed.

Measurements made with equipment responsive to either the horizontal or the vertical component of electric field are found to agree.

The transmission of short pulses instead of a steady carrier wave is discussed as a means of resolving the composite wave into components separated in time. More detailed and significant information can be obtained by this resolving method. The use of pulses indicates that (1) the direction of arrival of the components does not change rapidly, and (2) the components of greater delay arrive at the higher angle above the horizontal. The components are confined mainly to the plane of the great circle path containing the transmitting and receiving stations.

A method is described in which the angular spread occupied by the several component waves may be measured without the use of pulses.

Application of highly directional receiving antennas to the problem of improving the quality of radiotelephone circuits is discussed.

Electron Diffraction and the Imperfection of Crystal Surfaces.[5] L. H. GERMER. Bragg reflections are obtained by scattering fast electrons (0.05A) from the etched surfaces of metallic single crystals. The surfaces studied are a (100) face of an iron crystal, (111) face of a nickel crystal and (110) face of a tungsten crystal. In each case the reflections occur accurately at the calculated Bragg positions with no displacement due to refraction. A given reflection is found, however, even when the glancing angle of the primary beam differs considerably from the calculated Bragg value—by over 1.0° in some cases—so that several Bragg orders occur simultaneously. The accuracy with which this glancing angle must be adjusted is a measure of the degree of imperfection of the crystal. From the electron experiments, estimates are made of the widths at half maximum of electron rocking curves. These widths are 0.8° for the iron crystal, 1.5° for the nickel crystal and somewhat over 1.0° for the tungsten crystal. X-ray rocking curves for these same crystals are much narrower, although the observed

[5] *Phys. Rev.*, December 15, 1933.

widths vary considerably with the treatment of the surfaces. It is concluded that the values obtained from the electron measurements apply to projecting surface metal only, and that *the degree of misalignment is much greater at the surface than deep down within the crystal.* Furthermore, even the x-rays [Mo $K\alpha$ radiation $-$ 0.71A] are not sufficiently penetrating to yield values certainly characteristic of these metal crystals.

Mutual Impedance of Grounded Wires Lying on the Surface of the Earth when the Conductivity Varies Exponentially with Depth.[6] MARION C. GRAY. This paper presents a formula for the mutual impedance of any insulated wires of negligible diameter lying on the surface of the earth and grounded at their end-points, on the assumption that the conductivity of the earth varies exponentially with depth. Various special cases are briefly discussed.

Signals and Speech in Electrical Communication.[7] JOHN MILLS. This book is written by a member of the technical staff of Bell Telephone Laboratories who is well-known for his text on "Radio Communication" (1917) and the more popular presentations of "Within the Atom" (1921) and "Letters of a Radio-Engineer to His Son" (1922). In this book he presents for the general reader a synthesis of the electrical arts of communication in terms of their general fundamental principles. In separate chapters, which are discrete essays in popular and semi-technical language, the fundamental principles of dial operation, transmitters and receivers, loading coils, repeaters, multi-channel or carrier systems, and transoceanic radio-telephony are graphically expounded. The entertaining treatment of engineering achievements in allied fields of the sound picture, broadcasting, television, stereophonic reproduction and the teletypewriter, will intrigue the layman and assist him in acquiring a general understanding of these highly technical developments.

Some Earth Potential Measurements Being Made in Connection with the International Polar Year.[8] G. C. SOUTHWORTH. For several years the Bell System has been studying the relation between radio transmission and earth potential disturbances. A paper dealing with this subject was published in 1931. Prompted by the needs of the International Polar Year, together with the prospect that further work would throw additional light on the nature of radio transmission, the work was extended somewhat in 1932.

[6] *Physics*, January, 1934.

[7] Published by Harcourt Brace and Company, New York, N. Y., 1934.

[8] *Proc. I. R. E.*, December, 1933.

It is expected that useful correlation will be found between the normal earth potential effects which occur day after day during undisturbed periods and the corresponding diurnal and seasonal variation of radio transmission. It seems entirely probable, for instance, that earth potentials are but the terrestrial manifestations of certain changes taking place in the Kennelly-Heaviside layer which may not be found by other methods.

This paper is intended to serve mainly as a progress report outlining briefly the methods and scope of the work and showing the type of data being obtained. It leaves to a later date most of their correlation and their interpretation. The data here presented are in a conventional form used by other investigators for many years. Their value lies mainly in their extent and in the rather wide range of circumstances under which they were obtained.

Investigation of Rail Impedances.[9] HOWARD M. TRUEBLOOD and GEORGE WASCHECK. Measurements of impedance made on five sizes of rails and on two types of bonds are reported in this paper; the investigation covered a range of current per rail of 20 to 900 amperes, and frequencies of 15 to 60 cycles per second. Results are given in a form convenient for engineering use, and include information for applying corrections for bond impedance and for temperature.

[9] *Elec. Engg.*, December, 1933.

Contributors to this Issue

H. A. Affel, S.B. in Electrical Engineering, Massachusetts Institute of Technology, 1914; Research Assistant in Electrical Engineering, 1914–16. Engineering Department and the Department of Development and Research, American Telephone and Telegraph Company, 1916–34. Bell Telephone Laboratories, 1934–. Mr. Affel has been engaged chiefly in development work connected with carrier telephone and telegraph systems.

E. H. Bedell, B.S., Drury College, 1924; University of Missouri, 1924–25. Bell Telephone Laboratories, Acoustical Research Department, 1925–. Mr. Bedell's work has had to do mainly with studies of sound absorption and transmission and allied subjects in the field of architectural acoustics.

Arthur G. Chapman, E.E., University of Minnesota, 1911. General Electric Company, 1911–13. American Telephone and Telegraph Company, Engineering Department, 1913–19, and Department of Development and Research, 1919–34. Bell Telephone Laboratories, 1934–. Mr. Chapman is in charge of a group engaged in developing methods for reducing crosstalk between communication circuits, both open wire and cable, and evaluating effects of crosstalk on telephone and other services.

R. W. Chesnut, A.B., Harvard University, 1917; War Department of French Government, 1917; U. S. Army, 1917–19. Engineering Department, Western Electric Company, 1920–25; Bell Telephone Laboratories, 1925–. Mr. Chesnut has been engaged in the development of carrier telephone and long-wave radio systems.

Harvey Fletcher, B.Sc., Brigham Young University, 1907; Ph.D., University of Chicago, 1911; Instructor of Physics, Brigham Young University, 1907-08, and University of Chicago, 1909–10; Professor, Brigham Young University, 1911–16. Engineering Department, Western Electric Company, 1916–25; Bell Telephone Laboratories, 1925–. As Acoustical Research Director, Dr. Fletcher is in charge of investigations in the fields of speech and audition.

Frederick S. Goucher, A.B., Acadia University, 1909; A.B., Yale University, 1911; M.A., Yale, 1912; Ph.D., Columbia University, 1917. Western Electric Company, 1917–18. University College, London, 1919. Research Laboratories, General Electric Company, Limited, North Wembley, England, 1919–26. Bell Telephone Laboratories, 1926–. Dr. Goucher has been engaged in a study of the carbon microphone.

IDEN KERNEY, B.S. in Communication Engineering, Harvard Engineering School, 1923. Development and Research Department, American Telephone and Telegraph Company, 1923–34. Bell Telephone Laboratories, 1934–.

R. H. MILLS, S.B. in Electrical Engineering, Massachusetts Institute of Technology, 1916. Western Union Telegraph Company, 1916–18. Western Electric Company, Transmission Development Branch, 1918–25. Bell Telephone Laboratories, Apparatus Development Department, 1925–. Mr. Mills is responsible for the development of carrier frequency filters for use in commercial communication systems.

E. O. SCRIVEN, B.S., Beloit College, 1906; Instructor, Fort Worth University, 1906–08; S.M., Massachusetts Institute of Technology, 1911. Engineering Department, Western Electric Company, 1911–25. Bell Telephone Laboratories, 1925–. Mr. Scriven is in charge of the electrical design of special products apparatus other than radio.

W. B. SNOW, A.B., Stanford University, 1923; E.E., 1925. Engineering Department, Western Electric Company, 1923–24. Acoustical research, Bell Telephone Laboratories, 1925–. Mr. Snow has been engaged in articulation testing studies and investigations of speech and music quality.

J. C. STEINBERG, B.Sc., M.Sc., Coe College, 1916, 1917. U. S. Air Service, 1917–19. Ph.D., Iowa University, 1922. Engineering Department, Western Electric Company, 1922–25; Bell Telephone Laboratories, 1925–. Dr. Steinberg's work since coming with the Bell System has related largely to speech and hearing.

A. L. THURAS, B.S., University of Minnesota, 1912; E.E., 1913. Laboratory assistant with U. S. Bureau of Standards, 1913–16. Graduate student in physics, Harvard, 1916–17. Bell Telephone Laboratories, 1920–. At the Laboratories, Mr. Thuras has worked on the study and development of electro-acoustic devices and instruments.

E. C. WENTE, A.B., University of Michigan, 1911; S.B. in Electrical Engineering, Massachusetts Institute of Technology, 1914; Ph.D., Yale University, 1918. Engineering Department, Western Electric Company, 1914–16 and 1918–24; Bell Telephone Laboratories, 1924–. As Acoustical Research Engineer, Dr. Wente has worked principally on general acoustic problems and on the development of special types of acoustic devices.

VOLUME XIII JULY, 1934 NUMBER 3

THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

The Compandor—An Aid Against Static in Radio Telephony—*R. C. Mathes and S. B. Wright* . . 315

The Effect of Background Noise in Shared Channel Broadcasting—*C. B. Aiken* 333

Wide-Band Open-Wire Program System—*H. S. Hamilton* 351

Line Filter for Program System—*A. W. Clement* . . 382

Contemporary Advances in Physics, XXVIII—The Nucleus, Third Part—*Karl K. Darrow* 391

Electrical Wave Filters Employing Quartz Crystals as Elements—*W. P. Mason* 405

Some Improvements in Quartz Crystal Circuit Elements —*F. R. Lack, G. W. Willard and I. E. Fair* 453

A Theory of Scanning and Its Relation to the Characteristics of the Transmitted Signal in Telephotography and Television—*Pierre Mertz and Frank Gray* 464

Abstracts of Technical Papers 516

Contributors to this Issue 520

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

50c per Copy *$1.50 per Year*

THE BELL SYSTEM TECHNICAL JOURNAL

Published quarterly by the
American Telephone and Telegraph Company
195 Broadway, New York, N. Y.

EDITORIAL BOARD

Bancroft Gherardi — H. P. Charlesworth — F. B. Jewett
L. F. Morehouse — E. H. Colpitts — O. B. Blackwell
D. Levinger — O. E. Buckley — H. S. Osborne
Philander Norton, *Editor* — J. O. Perrine, *Associate Editor*

SUBSCRIPTIONS

Subscriptions are accepted at $1.50 per year. Single copies are fifty cents each. The foreign postage is 35 cents per year or 9 cents per copy.

Copyright, 1934

PRINTED IN U. S. A.

The Bell System Technical Journal

July, 1934

The Compandor—An Aid Against Static in Radio Telephony *

By R. C. MATHES and S. B. WRIGHT

One of the important conditions which must be met by any speech transmission system is that it should transmit properly a sufficient range of speech intensities. In long-wave radio telephony, even after the speech waves are raised to the maximum intensity before transmission, there remain energy variations such that weak syllables and important parts of strong syllables may be submerged under heavy static. The compandor is an automatic device which compresses the range of useful signal energy variations at the transmitting end and expands the range to normal at the receiving end, thus improving the speech-to-noise ratio.

This paper deals with some of the fundamental characteristics of speech waves and explains how the task of changing them for transmission over the circuit and restoring them at the receiving end is accomplished. It is also shown that raising the strength of the weaker parts of speech gives these results: 1, the successful transmission of messages for a large percentage of the time previously uncommercial; 2, a reduction of the noise impairment of transmission for moderate and heavy static during time classed commercial; and 3, the ability to deliver higher received volumes due to the improved operation of the voice controlled switching circuits. In addition to these advantages, the compandor makes it possible to economize on radio transmitter power in times of light static.

Introduction

WHEN the original New York–London long-wave radiotelephone circuit was designed, it was recognized that radio noise would often limit transmission, especially for the weaker voice waves. Accordingly provision was made for manually adjusting the magnitude of the speech waves entering the radio transmitters to such a value as to load these transmitters to capacity.[1] While this treatment was very effective in improving the average speech-to-noise ratio and in preventing the strong peaks of speech from overloading the transmitter, it was, of course, unsuitable for following the rapidly varying amplitudes of the various speech sounds.

The total range of significant intensities applied to the circuit is in the order of 70 db, an energy ratio of 10 million to one. The manual adjustments referred to above were succesful in reducing this range to about 30 db. To further reduce this residual range an interesting

* Presented at Summer Convention of A. I. E. E., June, 1934. Published in *Electrical Engineering*, June, 1934.

device called the compandor has been developed. This device which works automatically makes a further reduction of one-half in the residual db range so that the range transmitted over the circuit is then only 15 db, an energy ratio of about 32 to one.

Speech Energy

Quantitative designation of speech intensity and hence of a range of intensities is rendered difficult by the rapidly varying amplitude characteristics of the various speech sounds. Devices called volume indicators are used fairly extensively to indicate the so-called "electrical volume" * of speech waves. A volume indicator is essentially a rectifier combined with a damped d-c. indicating meter on which are read in a specified manner the standard ballistic throws due to partly averaged syllables at a particular speech intensity. These devices are so designed and adjusted that they are insensitive to extremely high peak voltages of short duration, but their maximum deflection is approximately proportional to the mean power in the syllable. It has been found that, if commercial telephone instruments are used, the ear does not detect amplifier overloading of the extremely high peaks of short duration. Consequently, the volume indicator is a useful device for indicating the noticeable repeater overloading effect of a voice wave. These devices do not tell us much about the effect of the weaker voltages in overriding interference or operating voice-operated devices but they give a fairly satisfactory indication of loudness and possibilities of interference into other circuits.

The sound energy that the telephone transmits consists of complicated waves made up of tones of different pitch and amplitude. The local lines and trunks connecting the telephone to the subscribers' toll switchboard have little effect in changing the fundamental characteristics of these waves but, on account of various amounts of dissipation, the waves received at the toll switchboard are always weaker than those transmitted by the telephone. Furthermore, the strength of signals varies with the method of using the telephone, loudness of talking, battery supply, and transmitter efficiency. The subscriber may be talking over a long distance circuit from a distant city, in which case the loss of the toll line further attenuates the received waves. Figure 1 † shows that the range of outgoing speech volumes as measured by a volume indicator at the transatlantic switchboard at New York is nearly 40 db for terminal calls. When via calls and variation in volume

* The term volume will be used through the rest of this paper to designate this quantity and not as synonymous with loudness.[2]

† This curve is plotted on so-called *probability paper*, in which the scale is such that data distributed in accordance with the *normal law* will produce a straight line.

of the individual talker are taken into account, it is even greater than 40 db.

Volume Range of a Telephone Circuit

There are two limits on the range of volumes which a system can transmit. The upper limit of volume is set by the point at which

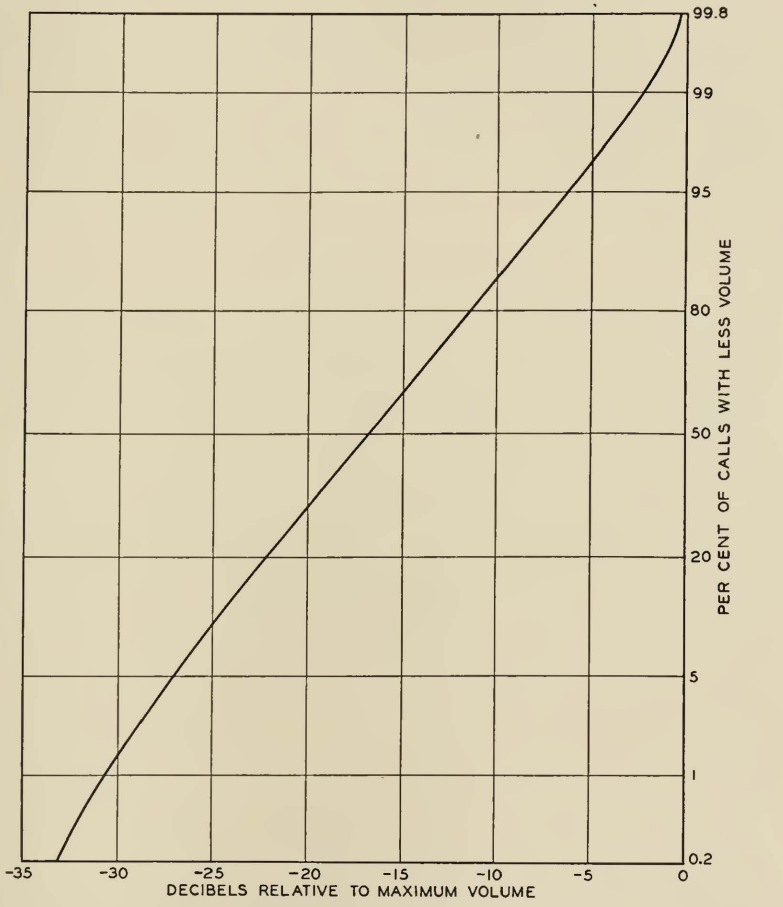


Fig. 1—Volumes of 950 local subscribers at New York transatlantic switchboard, January–April, 1931.

overloading appreciably impairs the signal quality or endangers the life of the equipment. It is an economic limit set by the cost of building equipment of greater load capacity. The lower limit of volume is set by the combination of the amount of attenuation and the amount of interference in the system such that the signal should not be appre-

ciably masked by noise. This also is ordinarily an economic problem depending on the cost of lowering the attenuation or of guarding against external interference. In some cases, however, this limitation is a physical one. A striking case is that of radio transmission in which we have no means of controlling the attenuation of the electromagnetic waves in transit to the receiving station. They may arrive at levels below those of thermal [3, 4] noise in the antenna and other receiving apparatus. Thus, even in the absence of static there is a definite useful lower limit to the received and hence the transmitted volume. In such cases the problems raised by the spread in signal intensities become a matter of particular importance. Radio telephony was therefore one of the fields of use particularly in view for the development of the device to be described.

Effect of Volume Control

Until recently the only method in use for reducing the range of signal intensities on radio circuits was a special operating method for constant volume transmission. At each terminal the technical operator, with the aid of a volume indicator, adjusted the speech volume going to the radio transmitter to that maximum value consistent with the transmitter load capacity.

Referring to Fig. 2, we have a diagram showing the normal relation of input to output intensities of a zero loss transducer as given by the diagonal line. Points $A_{\text{max.}}$ and $A_{\text{min.}}$ on this line indicate the extreme values of signal intensities for sustained loud vowels covering a volume range of 40 db. The effect of the volume adjustments made by the technical operator is to bring all the applied volumes to a single value indicated by point B in Fig. 2. The value of B could be any convenient intensity. Here it is set at a value determined by transmission conditions in the line between the technical operator's position and the radio transmitter.

As the technical operator has reduced the strongest volumes 5 db and increased the weakest volumes 35 db, the result of this volume control is to increase the volume range which the circuit can handle by 40 db. It is possible to make this adjustment for two-way transmission in the case of radio circuits without danger of singing because of the use of voice-controlled switching arrangements [5] which permit transmission in only one direction at a time. By this method of operation volumes initially strong or weak are delivered to the distant receiving point with equal margins relative to interference and the transmission capacity of the whole system is thereby improved.

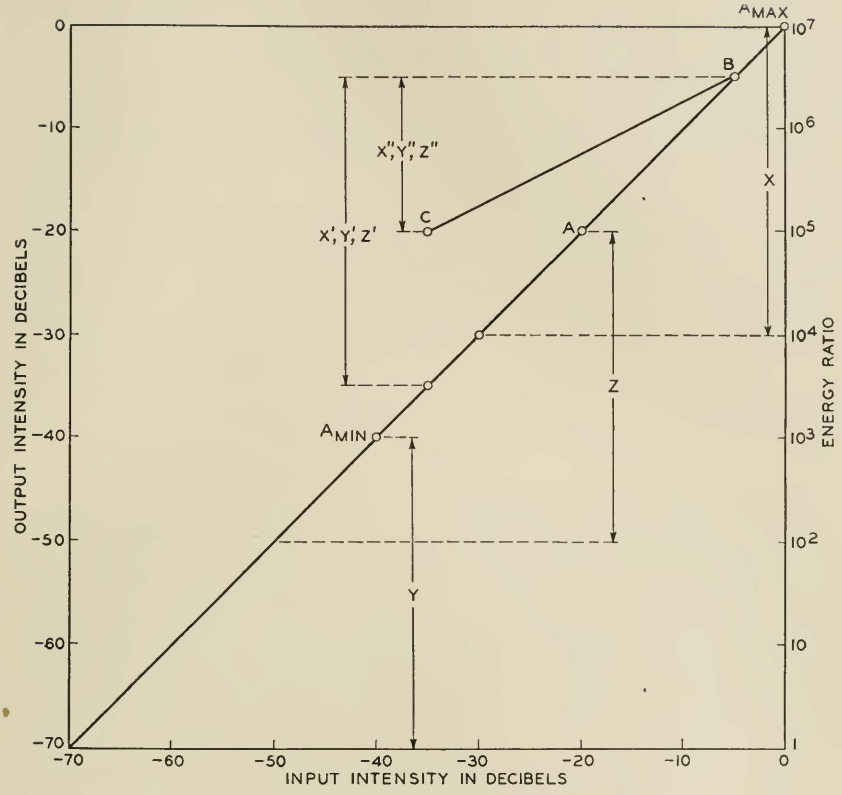


Fig. 2—Range control.

Intensity Range at Constant Volume

However, even with speech adjusted to constant volume at the transmitting point there are large variations in signal intensity from syllable to syllable and within each syllable. For example, the energy of some consonants as compared with the stronger vowels is down about 30 db. The importance of the weaker sounds is brought out by the fact that in the case of commercial telephone sets a steady noise 30 db below the energy in the strongest parts of the speech syllables produces an appreciable impairment in transmission efficiency. It is accordingly desirable to maintain transmission conditions such that generally more than this range is kept free from the masking effect of noise. This range of intensities within the syllable is also of importance in the operation of the voice-controlled switches used in the radio system. The sensitivity spread between a voice operated relay which

just operates on the crests of loud syllables and one which operates sufficiently well not to clip speech is also about 30 db.

Considering on Fig. 2 that the coordinates are in terms of the average r.m.s. value over a period of time small compared with the time of a syllable, there is a spread of at least 30 db in signal intensity extending down from the maximum for each talker. Thus for the weakest talker this spread is indicated by the bracket Y and for the strongest, by X. Any other talker, as Z, falls somewhere in between. After manual control of volume this spread of intensities is represented by the bracket X', Y', Z' for all talkers. This residual spread makes desirable a means for further compressing the range of intensities in the speech signals so that the weaker parts of sound are transmitted at a higher level without at the same time raising the peak values of speech and so overloading the transmitter.

Types of Compression Systems

This problem can be approached in several ways. One, for instance, is from the frequency distortion standpoint. As many of the weaker consonants have their chief energy contribution in the upper part of the speech band, a simple equalizer which relatively increased the energy of the higher frequency consonants before transmission and another which restored the frequency energy relations after transmission should be found of value. Tests have confirmed this expectation to some degree. Unfortunately, the best type of equalizer depends upon the type of subscriber station transmitter, so that in general only a compromise improvement can be obtained.

Another general method of approach is that of amplitude distortion in which the weaker portions of the syllable are automatically increased in intensity in some inverse proportion to their original strength. The manual control of volume described above may be considered the genesis of this method. Early suggestions [6] included the use of an auxiliary channel such as a telegraph channel for duplicating the control operations in the reverse sense at the receiving end, thus restoring the original energy distribution. Another early suggestion along this line was made by George Crisson of the American Telephone and Telegraph Company.[7] If a voltage be applied to a circuit consisting of a two-element vacuum tube (with a parabolic characteristic) in series with a large resistance, the instantaneous voltages across the tube are approximately the square root of corresponding voltages applied. Thus a voltage originally 1/100 of the peak voltage can be transmitted at an intensity of 1/10 of the peak or ten times its original intensity. If the instantaneous energy is expressed on the logarithmic

or db scale, the energy range is then cut in half. Such a device may be called an instantaneous compressor. At the distant end a circuit which is simply the inverse of that at the transmitting end is used. The output voltage is taken off of a low resistance in series with a parabolic element, thus restoring the signal substantially to its original form. This circuit may be called an instantaneous expandor. This scheme was successfully tested in the laboratory but unfortunately possesses a very serious limitation for practical application in the telephone plant. This is due to the fact that, to properly maintain the characteristics of the compressed signals, a transmission band width without appreciable amplitude or phase distortion of about twice the normal proved necessary.

The Compandor

The principle of the present device is the use of a rate of amplitude control for the compressing and expanding devices intermediate between manual and instantaneous control which may be considered approximately as a control varying as a function of the signal envelope.[8, 9] Such a modulation of the original signal in terms of itself does not appreciably widen the frequency band width of the modified signal as compared with the original signal. The transmitting device is called the compressor; the receiving device, the expandor; and the complete system, the compandor.

The functional behavior of a typical compressor may be considered with reference to the simplified schematic circuit No. 1 of Fig. 3.

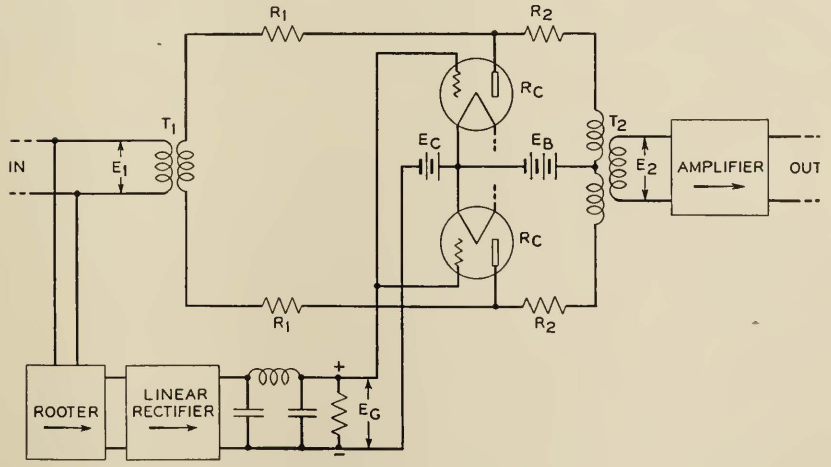


Fig. 3—Compressor circuit No. 1.

This circuit is of the forward-acting type; that is, the control energy is taken from the line ahead of the point of variable loss. The variable loss consists of a high impedance pad connected in the circuit through two high ratio transformers T_1 and T_2. The high resistances R_1 and R_2 are shunted by a pair of control tubes connected in push-pull. The push-pull arrangement is desirable for two reasons. It reduces the even order non-linear distortion effects caused by the shunt path on the transmitted speech and it balances out the control impulse and unfiltered rectified speech energy from the control path which might otherwise add distortion to the speech. The impedances of these tubes are controlled by the control voltage E_G, which is roughly proportional to the envelope of speech energy and which is derived from the line through a non-linear or "rooter" * circuit, a linear rectifier and a low-pass filter which may have a cutoff frequency in the range 20 cycles to 100 cycles. In the following analysis it is assumed that the delay due to this filtering is negligible:

Let E_1 = r.m.s. speech voltage at input

and E_2 = r.m.s. speech voltage at output in same impedance

R_C = a-c. impedance of control tubes.

Now if R_C is kept small compared to the pad impedance, we have approximately

$$E_2 = k_1 E_1 R_C. \tag{1}$$

Let E_G be the control voltage applied to the grids of the control tubes. With the plate voltage E_B just neutralized by the steady biasing grid voltage E_C, then only E_G may be considered as determining the space current and we may assume ideally that the space current

$$I_B = k_2 E_G^{s}.$$

Then

$$R_C = \frac{dE_B}{dI_B} = \mu \frac{dE_G}{dI_B} = \frac{1}{k_3 E_G^{s-1}}, \tag{2}$$

wheres s is determined by tube design and the ks are constants for constant μ tubes. For variable μ tubes equation (2) can be used to set requirements on the tube design.

From (1) and (2)

$$E_2 = \frac{k_1 E_1}{K_3 E_G^{s-1}}. \tag{3}$$

Now let the rooter be a non-linear circuit such that the instantaneous voltage is the tth root of E_1. After rectification and filtering we

* So called because the output is a root of the input; see equation (4).

shall have approximately

$$E_G = k_4 E_1^{1/t}. \tag{4}$$

From (4) and (3) we have

$$E_2 = \frac{KE_1}{E_1^{(s-1)/t}} = KE_1^{(t-s+1)/t}. \tag{5}$$

If $t = s = n$

$$E_2 = KE_1^{1/n}. \tag{6}$$

Now if the input voltage be increased by a factor x, the input increment in db will be 20 log x. The new output will be

$$E_2' = K(xE_1)^{1/n}.$$

The increment in output in db will be

$$20 \log \frac{E_2'}{E_2} = 20 \log x^{1/n} = \frac{20}{n} \log x.$$

The ratio of the output increment to the input increment in db is $1/n$ and the device is said to have a compression ratio of $1/n$. In other words, the per cent change in relative speech voltages in passing through the compressor is the same at all points in the intensity range. In the general form of this circuit, t and s need not be equal to secure a particular value of $1/n$.

In Fig. 4, Compressor Circuit No. 2 is shown, a backward-acting type of circuit. In this circuit the control tubes can be used to per-

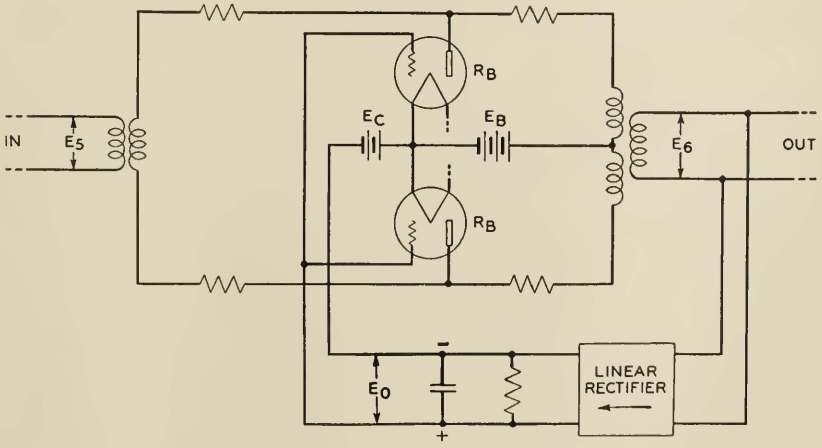


Fig. 4—Compressor circuit No. 2.

form the function of the rooter in circuit No. 1 when $s = t = n$. We may write for this circuit

$$E_6 = k_1 E_5 R_B,$$
$$E_0 = k_2 E_6,$$
$$R_B = \frac{1}{k_3 E_0^{n-1}} = \frac{1}{k_4 E_6^{n-1}},$$
$$E_6 = \frac{k_1 E_5}{k_4 E_6^{n-1}} = K E_5^{1/n}, \tag{7}$$

which is the same as equation (6) for circuit No. 1.

In Fig. 5 is shown the Expandor Circuit. If the resistances r are

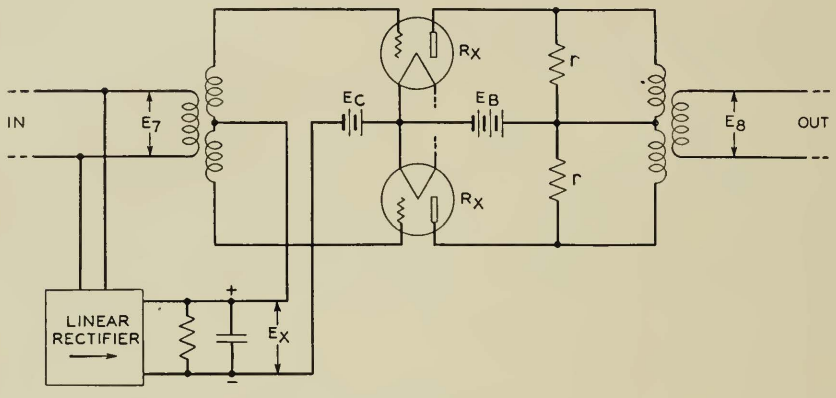


Fig. 5—Expandor circuit.

kept small compared with those of the control tubes, we may write

$$E_8 = \frac{k_1 E_7}{R_x},$$
$$E_x = k_2 E_7,$$
$$R_x = \frac{1}{k_3 E_x^{n-1}} = \frac{1}{k_4 E_7^{n-1}},$$
$$E_8 = K E_7 E_7^{n-1} = K E_7^n. \tag{8}$$

This relation is just the inverse of that given in equations (6) and (7). The increment ratio in db of output to input is n and the expansion ratio may be said to be n. When a compressor and expandor having the same value of n in their indices are put in tandem, the final output and input intensity ranges are the same. However, between the compressor and expandor the range of signal intensities, whose

rate of change is not faster than the usual syllabic envelope, is $1/n$ in terms of db. In terms of voltage ratios the intermediate signal intensities are proportional to the square root of their original values if n equals 2, the cube root if n equals 3, etc.

The ideal relations postulated above cannot all be met in the physical design of the circuits. The indices s and t must be the dynamic characteristics of the tube and circuit and can be held to constant value only over limited ranges of operation. Equation 2 is only approximately true as some space current is permitted to flow when no speech is passing; otherwise, impractical values of control impedances would be involved. However, they do serve to illustrate the functional operation and can be approximated sufficiently well in commercial equipment for useful amounts of compression and expansion. Figure 6 shows experimental steady-state input versus output characteristics for devices built to have a compression ratio of 1/2 and an expansion ratio of 2.

The compressor is seen to operate substantially linearly over a 45 db range of inputs and the expandor over a 22.5 db range. This is about as much range as can be secured conveniently from a single stage of vacuum tubes. As such ranges would be entirely insufficient to handle the seventy odd db range at speech intensities, it is necessary to control volumes to a given point before sending through these devices, rather than compress or expand first and then control. The range is adequate, however, to take care of the range of signal intensities for commercial speech at constant volume.

Effect of Compandor

The compressor curve of Fig. 6 indicates that, when the input is 15 db above 1 milliwatt, the compressor gives no gain or loss. If the levels are adjusted so that this point corresponds to the intensity at point B on Fig. 2, then the line BC indicates the controlled intensities corresponding to the assumed 30 db spread of speech controlled to constant volume. The new range of intensities as indicated by the bracket X'' Y'' Z'' is now finally reduced to about 15 db. Tests show that a volume indicator on the output of the compressor reads from 1 to 2 db higher than on uncompressed speech at its input. Compressed speech sounds slightly unnatural but the effects of compression upon articulation in the absence of noise are negligible.

In considering the action of the expandor it is important to note that all of the improvement in signal-to-noise ratio is put in by the compressor. Considering any narrow interval of speech the insertion of the expandor does not change the signal-to-noise ratio. The de-

sirability of using it depends on other reasons. First, it restores the naturalness of the speech sounds. Second, the apparent magnitude of the noise is greatly reduced since noise comes in at full strength only when speech is loudest and is reduced by the loss introduced by the expandor at times when the energy is low between syllables. When no speech is being transmitted, noises up to a certain limit, which

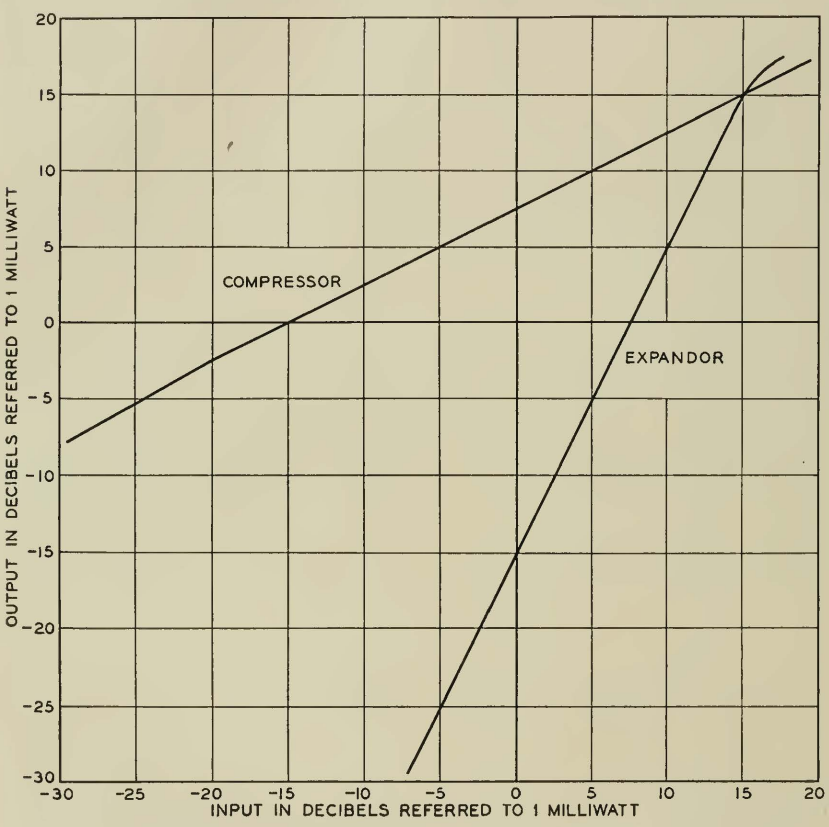


Fig. 6—Experimental input *vs.* output characteristics (1000 cycles steady state).

corresponds to the maximum energy in received speech, are reduced in varying amounts from about 20 db to zero depending on their value.

When speech is present the effect of the expandor is determined by the sum of the instantaneous speech and noise voltages, so that the effect on the noise, whether it is large or small, is determined largely by the existing speech intensity. For a circuit having somewhere near the limit of static, the use of the compandor allows on the average 5 db more noise than when it is not used. When the noise is less than

this limit, somewhat greater improvements are obtained from the compandor, ranging up to at least 10 db.

The particular values of compression and expansion ratio were chosen initially for the relative ease in the design of the system with commercially available vacuum tubes whose characteristics closely approximated a parabola. Tests of the equipment have shown that this degree is sufficient for present telephone circuit intensity range requirements. Increasing the amount of compression is limited by increase in quality distortion and by increased variation in the intensity of radio noise as heard by the listener. A noise which is constant at the input to the expandor varies on the output as the speech intensity changes. Also variations in attenuation equivalent between the compandor terminals are multiplied by the expandor. Herein lies a reason for having a constant compression and expansion ratio over the working range. If it were different at different intensities, attenuation changes would distort the reproduced speech as well as appearing as a somewhat increased change in intensity. This change in intensity is n times the attenuation change in front of the expandor in db.

The degree of compression may obviously be controlled in a variety of ways: such as, using different values for the indices s and t, applying control voltages upon more than one variable stage in tandem, the use of variable μ vacuum tubes, etc. The circuits as shown use variable shunt control for the compressor and variable series control for the expandor. Either or both may be changed to the other by inverting the polarity of the control potential and properly designing the rectifier characteristics of the control circuits.

There are two major sources of possible speech distortion which must be considered in the design and use of these devices in addition to those ordinarily present. The first is due to the non-linear characteristics of the vacuum tubes used for controlling. The even order distortion terms are largely balanced out by using two tubes in a push-pull arrangement. The remaining distortion is minimized by having speech pass through the control tubes at a sufficiently low level. In the operating ranges for the device shown on Fig. 6, the harmonics of a single-frequency tone are 30 db or more below the fundamental.

The second major source of distortion is the time lag in the control circuits due to the presence of the filters after the linear rectifier. However, with a complete compandor circuit using the compressor circuit No. 1, it was found on careful laboratory tests with expert listeners that it was almost impossible to distinguish whether the device

was in or out of circuit. Furthermore, distortion of this type is largely eliminated when compressor circuit No. 2 is used. In that case it will be noted that, if the two terminals are connected by a substantially distortionless transmission system, the identical control circuits of the two devices receive identical operating voltages. As the gain changes put in are reciprocal and occur now with equal time lag, the deviations from ideal compression are virtually counterbalanced by the inverse deviations from ideal expandor action. In Fig. 7 are shown

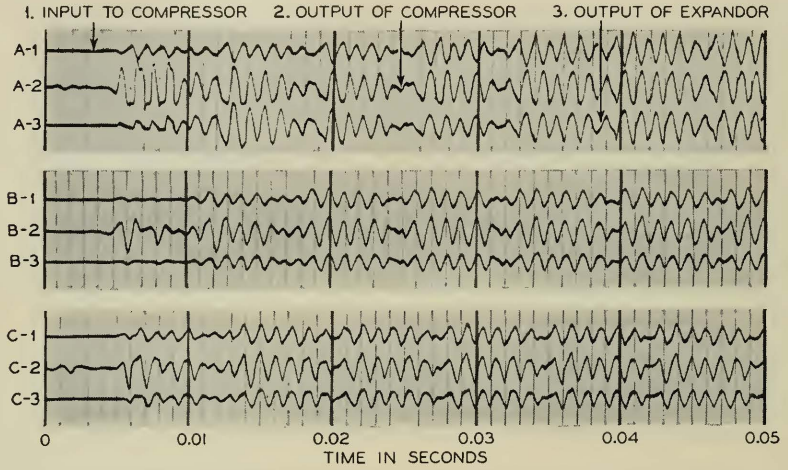


Fig. 7—Operation of compandor on beginning of word "bark." *A*. Compressor circuit No. 1. *B*. Compressor circuit No. 2 with low-pass filter in control circuit. *C*. Compressor circuit No. 2 without filter.

oscillograms taken of the first part of the word "bark." Each record shows the intensity changes before the compressor, between the compressor and expandor and on the output of the expandor.

Application to Transatlantic Circuit

A compandor system has been in service on the New York–London long-wave radiotelephone circuit since about July 1, 1932. At first compressor circuit No. 1 was used, and later a change was made to compressor circuit No. 2. Figure 8 is a photograph of the experimental installation at New York. It occupies about five feet of standard relay rack space. The blank panel shown in the photograph indicates the saving of apparatus resulting from the change to compressor circuit No. 2. Figure 9 is a schematic diagram showing the method of inserting the compressor and expandor in the radio telephone terminals at each end of the circuit. Since the two ends are similar, only one

end is shown. The compandor circuits are indicated in their relation to the subscriber, the toll switchboard, the vodas and privacy apparatus, and the radio transmitter and receiver.[5] A meter located at the point designated *A* would indicate the full range of applied volumes, at *B*, the controlled volumes and at *C*, the compressed speech signals.



Fig. 8—Experimental installation of compandor at New York.

When the United States subscriber talks, electrical waves set up by his voice pass over a wire line to the toll switchboard. They then divide in a hybrid set; part of the energy is dissipated in the output of a receiving repeater and part is amplified by a transmitting repeater whose gain is controlled by noting the reading of a volume indicator at

B and adjusting a potentiometer ahead of the transmitting repeater. The waves then act on the vodas which consists of amplifier-detector, delay circuit and relays for switching the transmission paths in such a manner as to prevent echoes, singing and other effects. When in the transmitting condition, the vodas is arranged to have zero loss so that the waves impressed on the compressor are practically the same as at *B*. The waves put out from the compressor are then sent through the privacy apparatus, the output of which is then sent over a wire line to the radio transmitter. The radiated waves are picked up by the distant radio receiver, amplified and transformed into voice-frequency energy which passes over a wire line to the terminal at the distant end.

The path of received waves in either terminal may be traced in the lower branch of the circuit shown on Fig. 9. After being made intel-

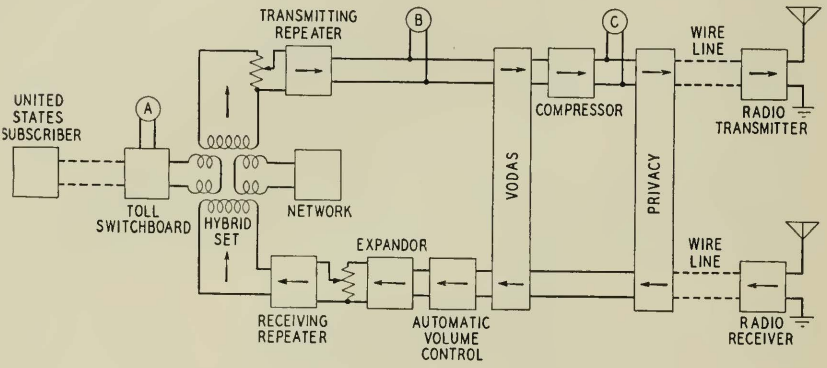


Fig. 9—Compandor applied to one end of a radio telephone circuit.

ligible by passing through the receiving privacy device, the compressed incoming waves are sent through the vodas into an automatic volume control and then into the expandor. The expanded waves are sent through a receiving repeater from whose output the amplified waves pass into the hybrid set, part being dissipated in the network and the other part going through the toll switchboard to the subscriber. Due to imperfect balance between the subscriber's line and the network, a portion of the received energy is transmitted across the hybrid set and amplified by the transmitting repeater. This echo might operate the transmitting vodas under certain conditions. For this reason a potentiometer is inserted in the receiving branch of the circuit so as to reduce the echo, and consequently the received volume, so that false operation of the transmitting vodas is prevented.

Results of Compandor Operation

The effectiveness of the compandor in service depends not only on its ability to reduce noise but also on its relation to the other characteristics of the circuit. Tests in the laboratory and on the long-wave transatlantic circuit have indicated that the presence of the compandor does not affect the quality appreciably, provided compressor circuit No. 2 is employed and provided the compression in the circuit itself is not serious. Delay distortion can be tolerated up to about the same amount as when no compandor is used. Frequency changing for privacy purposes is not materially affected by the compandor.

The expandor increases the transmission variations in the circuit exactly as it increases the voltage range of the waves applied to it. It is therefore necessary to guard against excessive variations in the overall circuit including the wire line extensions as well as the radio links. At the New York terminal there has been installed an automatic volume control operated from received speech signals which performs this function.

The received volume is limited by incoming waves which do not operate the receiving side of the vodas but which return as echoes from the land line to cause false operation of the transmitting side. The compressor increases these weak waves so that they are better able to operate the receiving side of the vodas, and the expandor effectively increases the stronger waves relative to the weak. This results in more received volume being delivered to the two-wire terminal than when the compandor is not used. The overall improvement in volume delivered to the subscriber varies with the noise, being greatest when the noise is low.

Summary

The allowable increase of about 5 db in noise before reaching the commercial limit increases the time when the circuit can be used for service. The increased circuit time is greatest in the seasons of the year when it is needed the most.

For conditions of moderate disturbances now classed as commercial, a reduction of the noise transmission impairment to very low values is accomplished by the compandor.

The improvement in the vodas operation results in delivering substantially higher volumes to the subscribers.

The beneficial effect of the compandor might alternately be applied to a reduction of transmitter power.

References

1. "The New York–London Telephone Circuit," S. B. Wright and H. C. Silent, *Bell System Technical Journal*, Vol. VI, pp. 736–749, October, 1927.
2. "Speech Power and Its Measurement," L. J. Sivian, *Bell System Technical Journal*, Vol. VIII, pp. 646–661, October, 1929.
3. "Thermal Agitation of Electricity in Conductors," J. B. Johnson, *Physical Review*, Vol. 32, pp. 97–109, July, 1928.
4. "Thermal Agitation of Electric Charge in Conductors," H. Nyquist; presented before the *American Physical Society*, February, 1927, and published in *Physical Review*, Vol. 32, pp. 110–113, July, 1928.
5. "Two-Way Radio Telephone Circuits," S. B. Wright and D. Mitchell, *Bell System Technical Journal*, Vol. XI, pp. 368–382, July, 1932, and *Proceedings of The Institute of Radio Engineers*, Vol. 20, pp. 1117–1130, July, 1932.
6. U. S. Patent 1,565,548, December 15, 1925, issued to A. B. Clark.
7. U. S. Patent 1,737,830, December 3, 1929, issued to George Crisson.
8. U. S. Patent 1,738,000, December 3, 1929, issued to E. I. Green.
9. U. S. Patent 1,757,729, May 6, 1930, issued to R. C. Mathes.

The Effect of Background Noise in Shared Channel Broadcasting

By C. B. AIKEN

The interference which occurs in shared channel broadcasting consists of several components of different types. Of these the program interference is usually the most important in the absence of a noise background, while if a strong noise background is present another component, which may be called flutter interference, predominates.

A simple theory of the flutter effect is developed and it is shown that its importance is dependent upon the type of detector employed. If manual gain control is used, flutter may be greatly reduced by the use of a linear rectifier. However, if automatic gain control is used this superiority of the linear detector cannot be realized and flutter is bound to be troublesome.

The results of experimental studies of the various types of interference are given and a comparison is made of the relative importance of flutter and program interference. The effects of the type of detector used and of the width of the received frequency band are observed. It is evident from these studies that improvements in the size of the lower grade service areas of shared channel stations might be obtained by close synchronization of the carrier frequencies, even though different programs are transmitted.

THE regulation requiring that carrier frequencies be maintained to within fifty cycles of their assigned values has resulted in the practical disappearance from shared broadcast channels of the heterodyne whistle, that most pernicious of all types of radio interference. Consequently, it is now unnecessary to have so large a ratio of the field strength of the desired signal to that of the undesired as was the case before the banishment of the high pitched squeal. Nevertheless, the field strength ratio which is necessary to permit of satisfactory reception on shared channels is still much higher than we should like it to be, and interference still abounds.

A very common type of interference is that which manifests itself as a fluttering or heaving sound, often very unpleasant in character. This phenomenon is caused by the periodic rise and fall of the background noise (static, R. F. tube and circuit noise, etc.) as the weak interfering carrier wave swings alternately in and out of phase with the carrier from the stronger station. In the complete absence of a noise background, program interference, or "displaced sideband interference" [1] as it may be called, is more troublesome than are flutter effects. Consequently, it is in regions other than the high grade service areas of shared channel stations that flutter effects are most annoying. In such regions they occur most prominently when the

[1] "The Detection of Two Modulated Waves Which Differ Slightly in Carrier Frequency," *Proc. I. R. E.*, January, 1931, and *Bell. Sys. Tech Jour.*, January, 1931.

frequency difference of the desired and interfering carriers is only a few cycles per second. As this difference is increased the flutter is transformed into a more sustained sound, rather harsh in character, and as it is still further increased a low growl appears which becomes more objectionable as it rises in frequency. The pitch of this growl cannot exceed 100 cycles unless one or both stations are violating the 50 cycle regulation. With the increasing use of very precise frequency control, heterodyne frequencies of a few cycles have become very common, and so, therefore, have flutter effects.

It has been pointed out in an earlier paper [2] that the magnitude of the flutter effect will depend upon the type of rectifier employed in the receiving set, and that it will be very much more objectionable when a square law detector is used than when a linear detector is employed. This is to be expected, since in the former case the audio-frequency output of the receiver will be proportional to the amplitude of the incoming carrier, while in the latter case the output will be essentially independent of the carrier amplitude, provided over-modulation does not occur. However, these statements refer to the case in which automatic gain control is not used. When the receiver is equipped with automatic control, as in most better grade modern receivers, the superiority of the linear detector is nullified and a serious flutter may occur.

In addition to displaced sideband interference and flutter, trouble may arise from distortion of the desired program by the action of the interfering carrier. One or both of the first two types of interference are likely to occur at lower field strength ratios than is the last, but at higher levels of the undesired carrier all three types are of importance and combine to degrade the quality of reception. In this paper, studies of all these types will be reported. Audible beat interference will not be discussed since it has been considered in other papers and, as just mentioned, is much less important than it used to be.

Theoretical Estimation of Flutter Effects

As has already been stated, the flutter effect is due to the rise and fall of the level of the noise background with variation in the effective amplitude of the impressed carrier. In order to study this effect, let us suppose that there are impressed upon the detector a component of radio frequency noise which may be represented by $N \cos (\omega + n)t$, and a desired carrier $E \cos \omega t$. $n/2\pi$ is assumed to be an audio-frequency.

If a square law, or quadratic, detector is employed, the audio-

[2] "Theory of the Detection of Two Modulated Waves by a Linear Rectifier." *Proc. I. R. E.*, Vol. 21, pp. 601–629, April, 1933.

frequency output will be proportional to the audio-frequency component of

$$[E \cos \omega t + N \cos (\omega + n)t]^2,$$

which is

$$EN \cos nt. \tag{1}$$

Now suppose that there is impressed, in addition to the desired carrier and noise component, a weak carrier $e \cos (\omega + u)t$. The sum of the strong and weak carriers may be conveniently regarded as a single wave of amplitude

$$(E + e \cos ut).$$

This may be substituted for the amplitude E in (1), giving for the noise output

$$EN(1 + K \cos ut) \cos nt \tag{2}$$

in which

$$K = e/E. \tag{3}$$

The noise which is heard will consist of a steady portion, the amplitude of which is proportional to EN, and another portion of variable amplitude which is proportional to $ENK \cos ut$.

The factors that determine the importance of the flutter are many and complex, but it seems likely that the most important of them is the ratio of the variable component of the noise output to the steady component. As long as the noise is loud enough to be obvious, this ratio should be a fairly good measure of the perceptibility of the flutter, and we shall venture to regard it as such. The experimental data to be reported later will bear out this assumption.

From (2) it is evident that the ratio mentioned is merely K, the ratio of the amplitude of the interfering carrier to that of the desired carrier. We shall call this ratio the "flutter factor" for the quadratic detector and designate it by F_Q.

$$F_Q = K = e/E. \tag{4}$$

It is interesting that F_Q is independent of the amplitude N of the high frequency noise.

It is possible to derive a similar factor, giving the ratio of the variable to the steady components of noise, for the linear detector. From equations (70a) and (71) of the paper [2] already mentioned it follows that the flutter factor for the linear detector, at low modulations of the desired wave, is

$$F_L = \frac{Ne}{4E^2} = \frac{kK}{4}, \tag{5}$$

in which $k = N/E$.

F_L is seen to be dependent upon the strength of the high frequency noise as well as upon that of the interfering carrier. It is also to be noted that the flutter will be more serious with the quadratic than with the linear detector by a factor $4/k = 4E/N$, which is usually large.

This derivation of F_Q and F_L on the basis of a single frequency noise component serves to indicate important differences between the two types of detector and to show how the flutter changes with the noise level and with the ratio of the incoming carrier amplitudes. In any practical case the noise field would consist of numerous frequency components, but it is reasonable to expect that the proportionalities expressed in (4) and (5) would still hold. However, the absolute values of N and K at which the flutter becomes detectable must be determined experimentally and may be expected to depend upon the width of the received frequency band.

In the foregoing derivations it has been assumed that there is no automatic volune control in the receiving set. A brief examination of the effect of such a device will now be made.

Action of an Automatic Volume Control

The comparative freedom from flutter effects which has been noted in the case of the linear detector may be regarded as due to the fact that the audio-frequency output of such a detector is independent of carrier amplitude over a wide range. If automatic volume control is used in the receiving set, the amplitude of the carrier wave will be maintained practically constant at the input terminals of the detector. If the effective carrier amplitude impressed upon the antenna undergoes a periodic fluctuation, due to very low frequency heterodyning between the two stations, the gain of the radiofrequency amplifier will undergo cyclic variations, so as to keep the carrier constant at the detector. Obviously this will cause a fluctuation in the amplitude of the side-bands, be they due to noise or program.

From this it is evident that, on the one hand, flutter effects in the presence of a noise background will usually be of minor importance if a good linear rectifier is employed in conjunction with a manual volume control; while, on the other hand, these effects may become extremely objectionable if automatic volume control is used. Because of the prevalent use of AVC in modern radio receivers the low flutter characteristics of the linear detector cannot be generally employed to reduce flutter interference on shared channels.

In the case of the square law detector, the output is proportional to the product of the amplitudes of the carrier and side frequencies. At first glance it might seem that the use of automatic volume control

should reduce the flutter effects, since it would iron out the variations in carrier amplitude impressed upon the detector. However, it is evident that this stabilization of the carrier will be exactly offset by the variation imposed upon the sideband amplitudes, and that consequently the flutter effects should be as evident when a normally functioning automatic volume control is used as they are in the case of manual control.

A perfectly functioning automatic volume control should make flutter effects approximately independent of the type of detector employed when the beat frequency is of the order of 2 or 3 cycles. However, at some of the higher frequencies, of the order of 20 to 40 cycles, the control will function with reduced efficiency, and at still higher frequencies will not function at all. Consequently, in this intermediate range the gain control may have some special effect and may make the flutter either worse or better than it would be with the same type of detector and manual control.

Experimental Studies

Equipment Used in the Study of the Effects of a Noise Background

A laboratory investigation was made of the interference between two waves of slightly different carrier frequency. A block schematic of the equipment used is shown in Fig. 1.

A modulated signal could be received from Station WABC, or, by throwing the switch *S*, it was possible to obtain an unmodulated carrier from a Western Electric No. 700A Oscillator, which is of very great frequency stability.[3] Whichever signal was used was fed through an impedance matching transformer to a radio frequency attenuator. The output of this attenuator was fed into the grid of one tube of a mixing amplifier. As indicated in the drawing, this amplifier consists merely of two shield grid tubes having a broadly tuned common plate circuit load.

The other tube of the mixing amplifier was energized, through a second radio frequency attenuator, by an unmodulated carrier derived from a crystal controlled laboratory oscillator of the same type as that which served as an alternative to WABC. This oscillator was part of a Western Electric No. 1A Frequency Monitoring Unit.[4] The monitor includes arrangements for measuring frequency differences between the oscillator included within it and an external source. In this case the external source was WABC, or the alternative carrier. The energy

[3] O. M. Hovgaard, "A New Oscillator for Broadcast Frequencies," *Bell Laboratories Record*, 10, 106–110, December, 1931.

[4] R. E. Coram, "A Frequency Monitoring Unit for Broadcast Stations," *Bell Laboratories Record*, 11, 113–116, December, 1932.

required by the frequency measuring device was supplied through a tuned buffer amplifier.

The voltage developed across the tuned circuit of the mixing amplifier was measured by a conventional form of vacuum tube voltmeter. By setting one attenuator at a very high loss, the magnitude of the signal supplied through the other could be measured, and the process then reversed. If the two signals were adjusted so as to give equal amplitudes across the tuned load, then any desired carrier ratio could be obtained by adding a known loss in one attenuator.

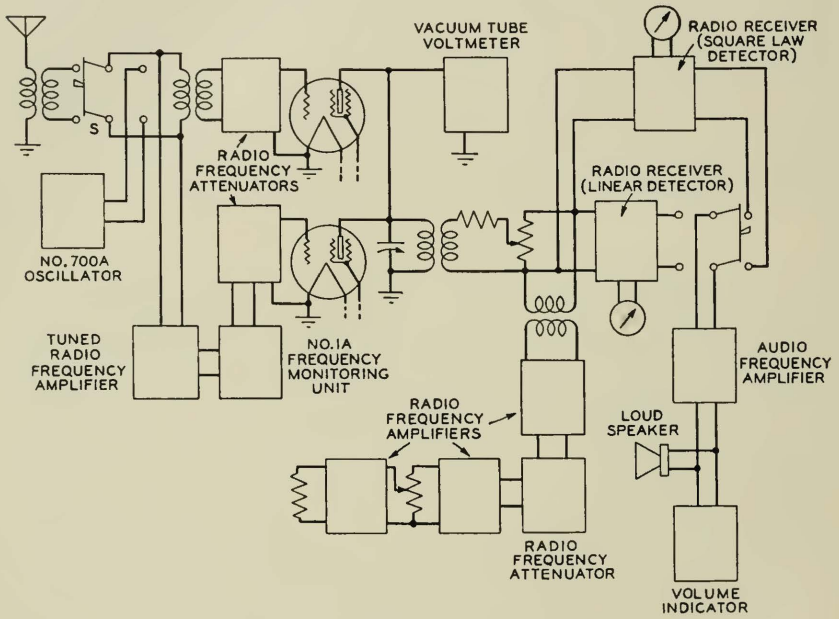


Fig. 1—Schematic circuits of experimental setup.

The mixing amplifier fed a shielded transmission line which included an adjustable pad. The line supplied energy to either of two radio receivers, one of which contained a square law and the other a linear detector. The output of the receiver was monitored on a loud speaker and also on a volume indicator. Meters were provided for indicating the change in direct current flow in the detector circuit of both receivers.

In order to study the effects of a noise background, a noise source of constant and controllable level was required. Furthermore, it was desirable that the noise be of a type frequently encountered in practice. The thermal noise generated in a high gain amplifier seemed to be suitable. Consequently, there were connected in cascade two amplifiers having a gain of approximately 44 db each, over the entire broad-

cast band. The output of the second of the units was fed through a radio frequency attenuator to the grid of a single stage amplifier, the output circuit of which contained a step-down transformer bridged across the transmission line feeding the radio receivers. With zero loss in the attenuator the noise energy fed to the line was ample for the purposes of the present study.

An additional description of some of the pieces of equipment used in the foregoing set-up may be of interest.

Source of Constant Unmodulated Carrier Frequency

The oscillator contained in the No. 1A Frequency Monitoring Unit is of unusual frequency stability. The piezo-electric crystal is mounted in a specially designed thermal insulating chamber which reduces the temperature fluctuations to an extremely small fraction of a degree. Voltage regulating equipment is included in the unit, giving further assistance in stabilizing the frequency. Detailed descriptions of the oscillator [3] and of the frequency monitor [4] have been published.

A similar oscillator is used as a control unit at Station WABC. Hence, it was expected that a very constant beat frequency could be obtained between that station and the local oscillator. The frequency of the latter was adjustable over a narrow range by means of a vernier condenser in the crystal circuit. Figure 2 shows a number of plots of

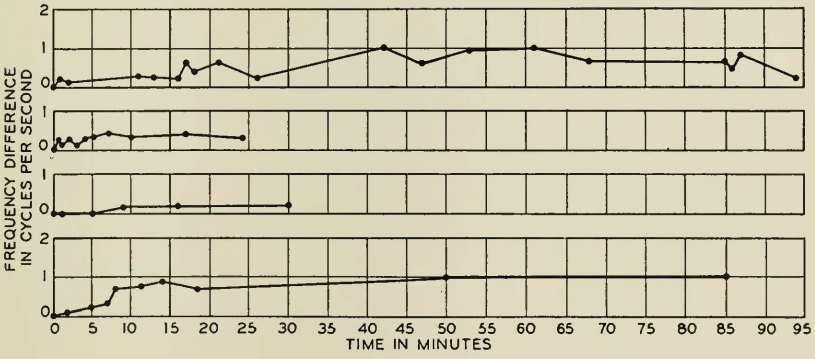


Fig. 2—Beat frequency between WABC and Western Electric No. 1A Frequency Monitoring Unit.

the beat frequency against time. These curves indicate an extremely slow drift, and experience has shown that the beat frequency would hold to within 0.4 cycle over a period of at least five minutes, and usually considerably longer. This high stability greatly facilitated work which required a very small difference in frequency of the two carriers.

Radio Receivers

Both receivers were high fidelity (7000 cycles) units of the tuned radio frequency type. One of these was modified so that either manual or automatic volume control could be used, and the level impressed upon the detector was reduced so that it would function as a strictly square law device. The cathode resistor which normally furnishes a grid bias for the detector tube was replaced by a battery. This was necessary in order to prevent straightening out of the characteristic by degeneration at very low frequencies. In Fig. 3 is shown a plot of

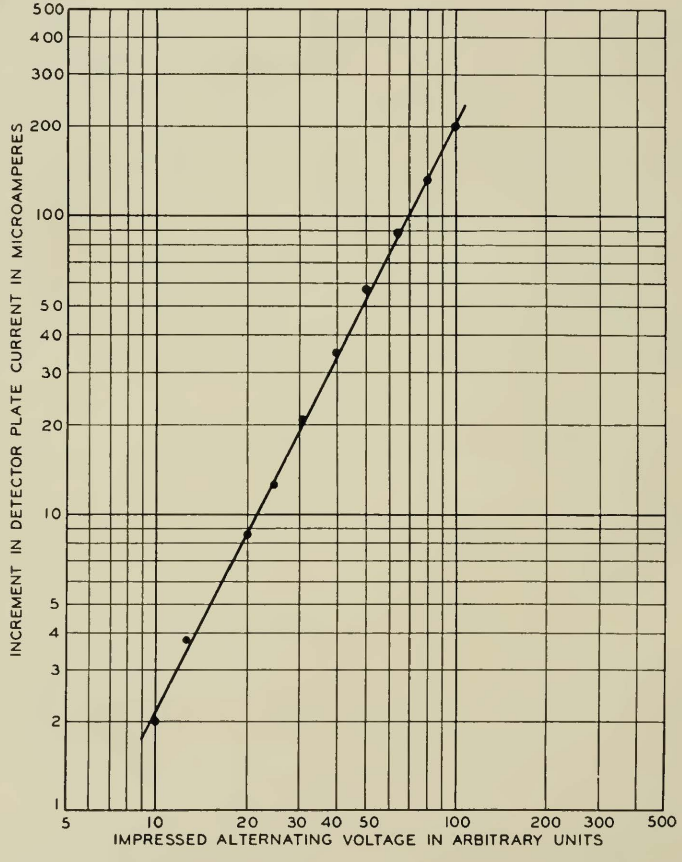


Fig. 3—Characteristic of radio frequency amplifier and square law detector.

the change in detector space current as a function of the impressed voltage. It will be observed that for increments of less than 200 μA the characteristic has a slope of two to one. All observations were

made at signal levels which were low enough to stay well within this range.

The other set was provided with a diode rectifier which functioned as a linear detector. In order to improve the linearity of the characteristic, an initial bias was used and was adjusted to obtain the best characteristic as indicated by the following test:

If a large unmodulated carrier is impressed on a linear rectifier, together with a much smaller unmodulated carrier, the beat frequency output should be independent of the amplitude of the larger carrier over a wide range. This phenomenon was observed experimentally and the initial bias was altered until the range, over which the large carrier could be adjusted without changing the output, was a maximum. In Fig. 4, the horizontal curve shows the magnitude of the

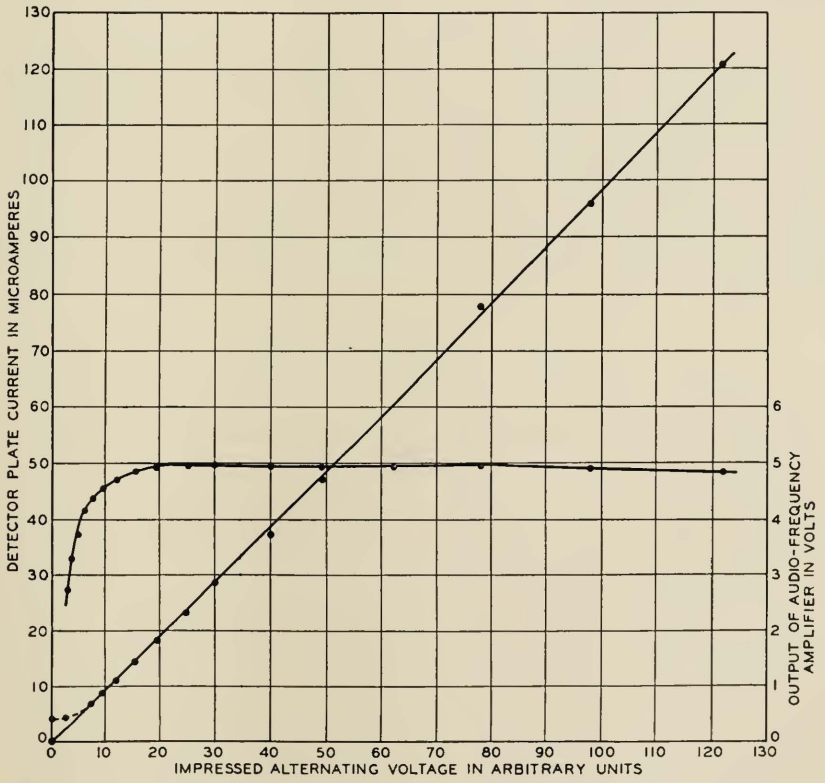


Fig. 4—Characteristics of radio frequency amplifier and linear detector.

audio-frequency output, while the sloping curve shows the direct current flowing in the detector circuit. The dashed curve is due to the

presence of the weak signal, while the solid one represents the effect of the large signal alone. The curves of this figure were taken with a bias of + 0.5 volt, which was found not to be critical.

The results of the experimental observations made with this detector were entirely in accord with theory, as will be discussed later, while similar observations made with a zero bias gave results which differed considerably from those predicted by the theory of the linear rectifier. Lack of the small bias caused a considerable departure from linearity, as was plainly evidenced by the fact that when it was absent the audio-frequency output due to the two carriers was by no means independent of the magnitude of the larger.

The tuned circuit in the mixing amplifier was so broad as to have an entirely negligible effect on the fidelity of the radio receivers.

Listening Conditions

In studying the effects of noise background some observations were made in the open laboratory, and a greater number in a partially deadened room 10 feet x 10 feet x 10 feet. The sound-proofing of this room was sufficient to keep out street noises and other extraneous disturbances of moderate intensity.

In determining the dependence of a given effect upon the magnitude of the carrier ratio, there was recorded that value of the ratio at which the effect was just perceptible.

Results of Experimental Work

A number of observations have been made with the intention of obtaining practical data on the characteristics of reception in the presence of a noise background, and with the purpose of checking the theoretical predictions already given. It has been pointed out that the flutter effects depend upon the type of detector which is employed and upon the ratio of the two carriers. If a square law detector is used the effect should be very nearly independent of the magnitude of the noise level, so long as it is within reasonable limits and does not either overload any of the equipment (including the ear of the listener) or fall so low as to be hardly noticeable. On the other hand, if a linear detector is employed, flutter effects should increase with the noise level. In either case the modulations of the two stations play no important part in determining the flutter effects except in so far as high modulations may temporarily mask them.

As a result of these considerations it was decided to employ unmodulated carriers for the greater part of the work. In order that a suitable level might be chosen, the strong carrier was first adjusted to

give the proper change in detector current. It was then modulated 30 per cent with a pure tone, and the gain of the audio-frequency output amplifier was adjusted until a fairly loud, but entirely comfortable, level was delivered to an observer placed about six feet in front of the loud speaker. The output level of the audio-frequency amplifier was read on a meter so that its gain might be checked later on.

The Linear Rectifier

The detector of a radio receiver was adjusted to have a linear rectifier characteristic in the manner just described and manual gain control was employed. In the first set of runs the carrier ratio was determined at which the flutter effect at low frequencies, or the carrier beat-note at higher frequencies, became just noticeable, the frequency being the variable. In Fig. 5 is shown a curve representing a number of

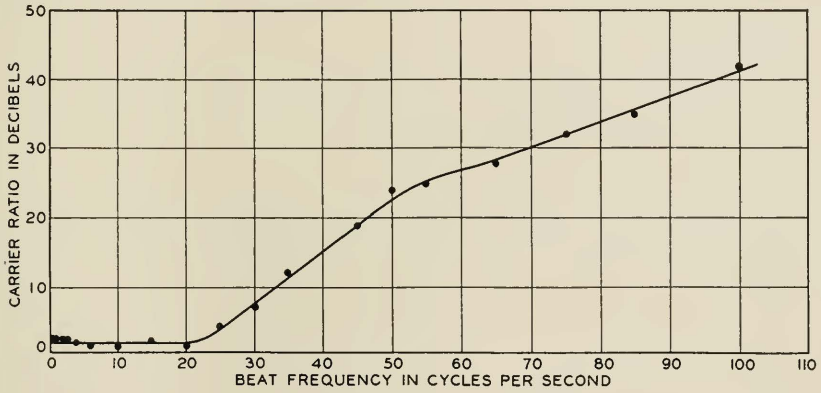


Fig. 5—Carrier ratio for perceptible flutter with a linear detector. Noise equivalent to 9.5 per cent modulation.

observations of this type. The noise level was constant at 10 db down from a 30 per cent modulated signal. By this it is meant that when the noise was impressed upon the receiver, together with a carrier the level of which had been fixed as described above, the audio-frequency output, as measured on a copper oxide level indicator, was 10 db below the audio output resulting from a 30 per cent modulation of the same carrier in the absence of noise.

A very interesting fact to be noted from this curve is that, for beat frequencies of less than about 20 cycles, the carriers must be very nearly equal before any flutter effect whatever may be detected. The average curve has been drawn through a value of 1.5 db. The observed values vary from this figure by not more than $\pm$ 0.5 db.

The right-hand portion of the curve is determined by the audibility of the beat-note, and its position will of course depend upon the masking effect of the noise background. Theory has indicated that the flutter frequency portion of the curve should drop with the noise level, but it is evident that, with such a small difference in carrier amplitudes as is indicated in the figure, the results would not be appreciably different were the noise level to be reduced. On the other hand a noise level which is down only 10 db from a 30 per cent modulated signal is equivalent to a modulation of nearly 10 per cent. This is an extremely objectionable noise level, so objectionable, in fact, that under the conditions of the tests it was very unpleasant to listen to. Consequently, it did not seem worth while to run curves similar to that of Fig. 5 for a number of different noise levels. Instead, a set of observations was made with a fixed carrier frequency difference of 2 cycles and a variable noise level. The results are indicated by the lower curve in Fig. 6.

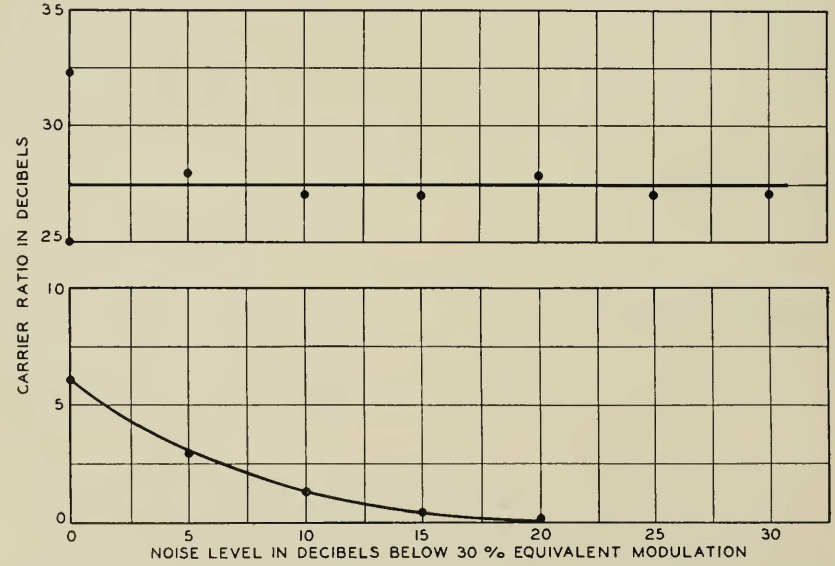


Fig. 6—Carrier ratio for perceptible flutter as a function of noise level. The upper curve is for the *square law* and the lower curve for the *linear detector.*

With a noise level equivalent to a 30 per cent modulation, a carrier ratio of only 2 : 1 is necessary to reduce the flutter to a barely detectable amount. At low noise levels, down 20 db or more from 30 per cent, the flutter could hardly be detected but there was noticeable a "bumping" sound which was due to the rather violent motion of the cone of the loud speaker at a frequency of 2 cycles. This was partially eliminated

by inserting a capacity in series with the voice frequency circuit of the speaker, but even when greatly reduced the bumping was detectable and was more important than any flutter which may have been present.

The Square Law Rectifier

Observations similar to those just discussed were made with a square law detector. In Fig. 7, the ordinates represent the carrier ratio neces-

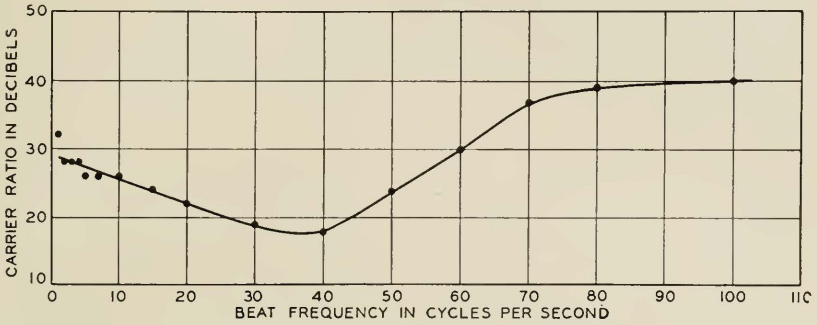


Fig. 7—Carrier ratio for perceptible flutter with a square law detector. Noise equivalent to 9.5 per cent modulation.

sary to reduce the flutter to a just detectable value, while the abscissae represent the beat frequency. The noise is 10 db down from an equivalent 30 per cent modulation. The curve is in striking contrast to that of Fig. 5. At very low frequencies a carrier ratio of 28 db is required when a square law detector is employed, while if the receiving set embodies a linear detector a ratio of 1.5 db is sufficient. The right-hand portions of the curves are fairly similar, since the carrier ratio is here dependent upon the audibility of the beat note and not upon flutter effects. The observations of which Fig. 7 is a record were made in the small sound-proof room. In Fig. 8 are shown two curves made in the open laboratory. In the upper curve the noise output was approximately 20 db down from that due to a 30 per cent modulated signal, while in the lower curve it was approximately 30 db down.

The theory which has been outlined indicates that in the case of the square law detector the flutter effects should be practically independent of noise level, and the curves shown in the last three figures bear out this prediction quite positively. Even more definite confirmation is furnished by the upper curve of Fig. 6, which shows the result of observations taken with a fixed beat frequency of 3 cycles. The two curves of this figure show the great superiority of the linear rectifier over the square law in receiving non-isochronous transmissions in the presence of a noise background.

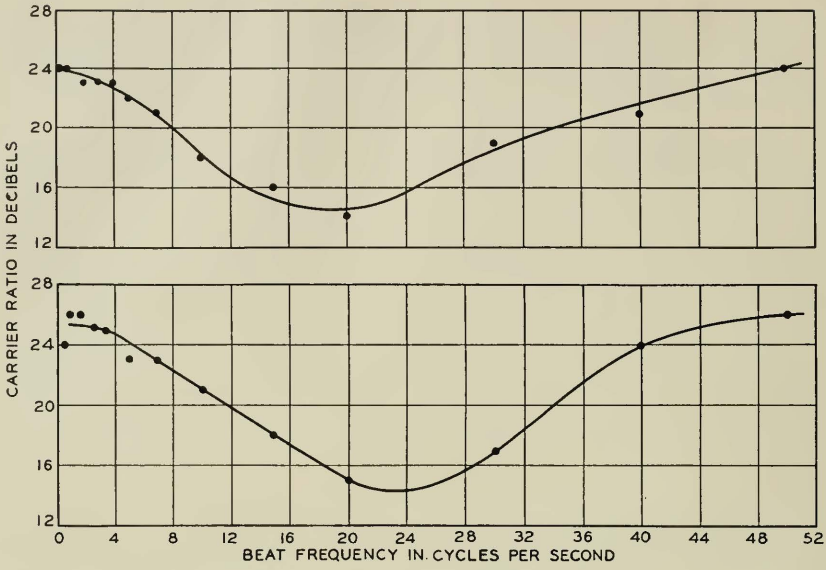


Fig. 8—Carrier ratio for perceptible flutter with a square law detector. Noise equivalent to 3 per cent modulation, for the upper curve, and to 0.95 per cent for the lower curve.

The Square Law Rectifier with Automatic Volume Control

It has been predicted that the use of automatic volume control in the receiving set should greatly increase the flutter effects observable with a linear rectifier, while with a square law device these effects should be the same for both automatic and manual control except, perhaps, at the frequencies of reduced efficiency of the gain control. An experimental check was made on the latter statement, the results of which are shown in Fig. 9. It will be noticed that this curve is very similar to the curves of Figs. 7 and 8.

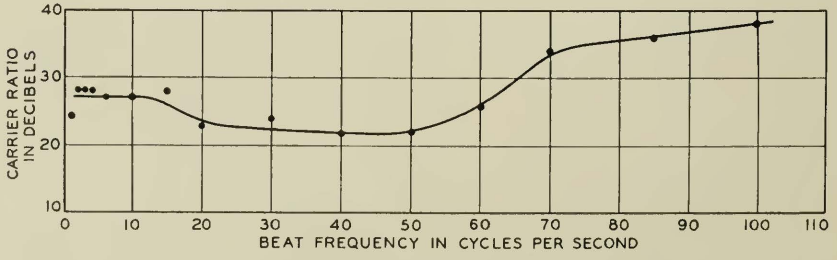


Fig. 9—Carrier ratio for perceptible flutter with automatic volume control. Noise equivalent to 9.5 per cent modulation.

The action of an automatic volume control, in keeping constant the level of the total carrier delivered to the detector, should become less pronounced as the beat frequency rises and should fail altogether when this frequency reaches the audible range. This reduction in efficiency of control may either increase, leave unaltered, or decrease the magnitude of the flutter, depending upon the amount of time delay involved in feeding back the controlling voltage. In the receiver used, the reduction in efficiency of the gain control occurred between 20 and 40 cycles. A comparison of Fig. 9 with Figs. 7 and 8 indicates that in this receiver the gain control tends to increase the flutter somewhat when the heterodyne frequency is within this range.

Interference of Undesired Program

When the interfering station transmits a program which is different from that of the desired station, serious interference may occur which is due primarily to the beats between the undesired sidebands and the desired carrier. If the carrier beat frequency is subaudible and there is little or no noise background, this will be the predominant form of interference. Its magnitude will depend upon the degree of modulation of the undesired signal, but is practically independent of the type of detector and gain control which are used. In the presence of considerable noise background it may or may not be more important than flutter effect.

In order to get some data on this point, observations were made with a square law detector and manual gain control. This represents about the worst condition, as far as flutter effect goes, but will be approximated by AVC receivers. At a fixed noise level the carrier ratio was determined at which the flutter could be noted, and also the ratio at which the program interference was detectable. This was done for receiver band widths of 7000 and 3500 cycles. The band width had no appreciable effect upon the program interference but exercised a very definite effect upon the flutter. Fig. 10 shows the results of the observations which were taken. The solid sloping curve represents the average of the observations on program interference, while the two horizontal curves show the carrier ratio at which the flutter was just detectable for the two bands widths used. The program interference was classed as audible when it could just be heard on the peaks of modulation. However, for considerable intervals of time it was entirely inaudible. Consequently, when the same carrier ratio was recorded for the flutter and for the program interference the former was actually the more annoying. In order to take account of this difference of character between the two types of interference it is necessary to

shift the program curve downward. Just how far it should be displaced is very hard to determine, as the amount will depend upon the type of program on the undesired station. Observations have indicated that the shift should amount to at least 7 db. The dashed curve in Fig. 10 has been drawn 7 db below the solid curve.

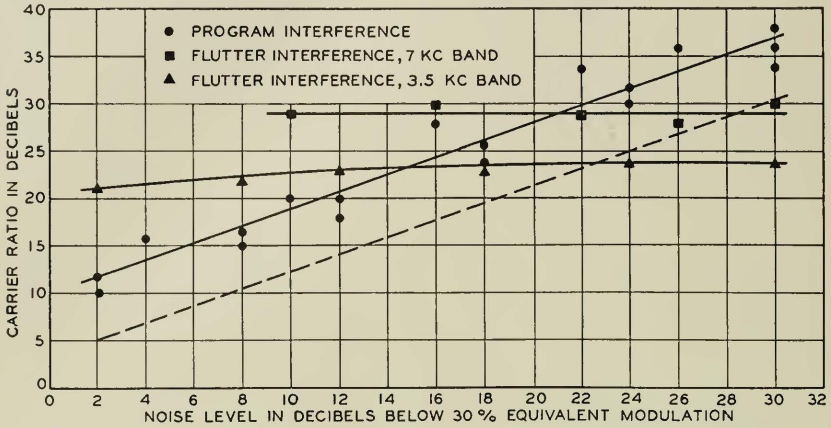


Fig. 10—Relative importance of program interference and flutter interference.

It will be noticed that with a band width of 3500 cycles the flutter curve crosses the program curve at a noise level equivalent to about 2 per cent modulation. (In every case the noise level was measured with the 7000 cycle band, regardless of what band was to be used in the listening tests. This should be kept in mind throughout the present discussion.) This means that at equivalent modulations of more than 2 per cent the flutter effect would be more objectionable than the program interference. However, at high noise levels, say 5 to 10 per cent, the listener would be sure to reduce the band width of his receiver to considerably less than 3500 cycles and this would reduce the relative importance of the flutter. Nevertheless, at very high noise levels the flutter is more important than the program interference. If the undesired station were to employ abnormally low modulation the program interference would be decreased and the relative importance of the flutter increased.

It is evident that the dependence of the flutter on band width, and the different reaction of individual observers as to what type of interference is the more objectionable, renders it impossible to make a definite statement as to the exact values of carrier ratio and noise field which will make the two types of interference equally important. But we can draw the useful conclusion that in cases of excessive noise, such

as may occur in rural areas without causing the listener to abandon attempts at reception, the flutter will be the more important. Consequently, an improvement in the service in such regions would be obtained by synchronizing the carriers of the two stations, even though they continue to transmit different programs.

Effect of Interfering Carrier on Desired Program

Even if the interfering wave were unmodulated and there were a neglibible noise background, there still remains the possibility of distortion of the desired program by the heterodyning action of the undesired carrier. In order to determine how important this effect is as compared with those which have been discussed, a modulated carrier (derived from WABC) and a weaker unmodulated wave were used. A beat frequency of about 3 cycles was maintained during the course of these observations.

With the linear rectifier it was found that a perceptible distortion of the desired program could not be detected on speech and jazz music until the weak carrier was brought within 1 db of the strong one. When the program consisted of music containing many sustained notes, such as occur in a violin solo and even in vocal solos, the cyclic variations in output level were more noticeable. In such a case a ratio of about 4 db was necessary to reduce the distortion to the detectable limit.

With the square law rectifier it was found that a carrier ratio of 10 db produced detectable distortion with any type of program. At a ratio of 16 db distortion could be detected only when the program contained sustained notes, and at 18 db could be noticed only when the notes were sustained for a considerable time.

The dependence of the permissible ratio upon the type of program led us to make a similar observation when the strong carrier was modulated 30 per cent with a pure tone of 400 cycles. Under such conditions it was necessary to reduce the interfering carrier to about 34 db below the strong one before the 3-cycle variation in the pure tone definitely vanished.

Conclusions

The studies which have been reported furnish quantitative data on the various types of interference which are encountered in shared channel broadcasting and show what relative levels of interfering carrier may be tolerated under various conditions.

In high grade service areas the program from the undesired station will be the most serious form of interference, provided the carrier beat

frequency is subaudible. If there is a moderate noise background present, it will tend to mask the program and will therefore permit of somewhat higher interfering field strength. However, if the interference is raised beyond a certain level, dependent upon the received band width, flutter effects will become pronounced. This will not be true with a linear detector and manual gain control, but in practice radio receivers which have linear detectors almost invariably have automatic volume control.

If the noise level is very high it may mask even rather loud program interference, and under such conditions the flutter effect is likely to be much the most serious source of trouble. This condition is of practical occurrence in outlying areas where a degraded service must be tolerated continually. In such regions shared channel broadcasting is limited in usefulness primarily by the flutter effects, and in extreme cases, by distortion of the desired program due to the heterodyning action of the interfering carrier. Both of these types of disturbance would be eliminated by synchronizing the carriers of the two stations, and it seems likely that control of the carrier frequencies to within $\pm$ 0.1 cycle might definitely extend the limits of the lower grade service areas of shared channel stations.

Acknowledgment

I wish to acknowledge my indebtedness to Mr. J. E. Corbin for his assistance in carrying out the experimental work which has been reported in this paper.

Wide-Band Open-Wire Program System *

By H. S. HAMILTON

Radio programs are regularly transmitted between broadcasting stations over wire line facilities furnished by the Bell System. Both cable and open wire facilities are employed for this service. Recently a new program transmission system for use on open wire lines has been developed which has highly satisfactory characteristics. A description of this open wire system and test results obtained with it are given in this paper.

THE simultaneous broadcasting of the same radio program from a large number of broadcasting stations, in different sections of the United States, has become of such everyday occurrence that the radio listening public takes it as an accepted fact and in many cases does not know whether the program is originating in the studio of a local broadcasting station or in a broadcasting studio in some distant city. The wire line facilities furnished by the Bell System for the interconnection of the radio stations, particularly the wire line facilities in cable, have such transmission characteristics that little detectable quality impairment is introduced even when programs are transmitted over very long distances.

This cable program system was described in a recent paper.[1] More recently a new program system for use on open-wire lines, which possesses transmission characteristics comparable with those of the cable system, was developed and an extensive field trial made involving two circuits between Chicago and San Francisco. This paper describes this new open-wire program system and gives the principal results of the tests made on the two transcontinental circuits.

In the paper referred to describing the cable system, the various factors and considerations involved dictating the grade of transmission performance that is desired for program circuits were discussed in considerable detail so they will not be reviewed here. The transmission requirements chosen as objectives for both cable and open wire are as follows:

Frequency Range

Frequency range to be transmitted without material distortion—about 50 to 8,000 cycles.

* Published in April, 1934 issue of *Electrical Engineering*. Scheduled for presentation at Pacific Coast Convention of A. I. E. E., Salt Lake City, Utah, September, 1934.

[1] A. B. Clark and C. W. Green, "Long Distance Cable Circuit for Program Transmission," presented at A. I. E. E. Convention, Toronto, June, 1930; published in *Bell Sys. Tech. Jour.*, July, 1930.

Volume Range

Volume range to be transmitted without distortion or material interference from extraneous line noise—about 40 db which corresponds to an energy range of 10,000 to 1.

Non-Linear and Phase Distortion

Non-linear distortion with different current strengths and phase distortion to be kept at such low values as to have negligible effect on quality of transmission even on the very long circuits.

The frequency range afforded by the new open-wire program circuits extends about 3,000 cycles higher and more than 50 cycles lower than the frequency range available with the open-wire [2] program circuits previously used. The extension of the frequency range at the upper end necessitates the sacrifice of one carrier telephone channel of carrier systems operating on the same wires with the program pair since the frequency band of the lowest carrier channel lies in this range. In order to minimize noise and the possibility of crosstalk, the phantoms of program pairs are not utilized and, of course, d.-c. telegraph compositing equipment is removed in order that the proper low-frequency characteristics may be realized.

Description of New Open-Wire System

In general, the amplifiers on the open-wire program circuits employ the same spacing as the telephone message circuit repeaters on the same pole lead. The average repeater spacing is about 150 miles but the repeaters may be located as close as 60 miles or may be as much as 300 miles apart depending on the location of towns and cities on the open-wire route and the gauge of the wires used. The upper diagram of Fig. 1 shows a typical layout of the new wide-band open-wire program system. Three types of stations are shown, a terminal transmitting station, an intermediate station which may be either bridging or non-bridging and a terminal receiving station.

The terminal transmitting station includes an equalizer for correcting for the attenuation distortion of the local loop from the broadcasting studio, an attenuator for adjusting the transmission level received from the local loop to the proper value, an amplifier for inserting the required gain, filters for separating the program and carrier channels, monitoring amplifier, loudspeaker and volume indicator for

[2] A. B. Clark, "Wire Line Systems for National Broadcasting," presented before the World Engineering Congress at Tokio, Japan, October, 1929; published in *Proc. I. R. E.*, November, 1929, and in *Bell Sys. Tech. Jour.*, January, 1930. F. A. Cowan, "Telephone Circuits for Program Transmission," presented at Regional Meeting of A. I. E. E., Dallas, Texas, May, 1929; published in *Transactions of A. I. E. E.*, Vol. 48, No. 3, pages 1045–1049, July, 1929.

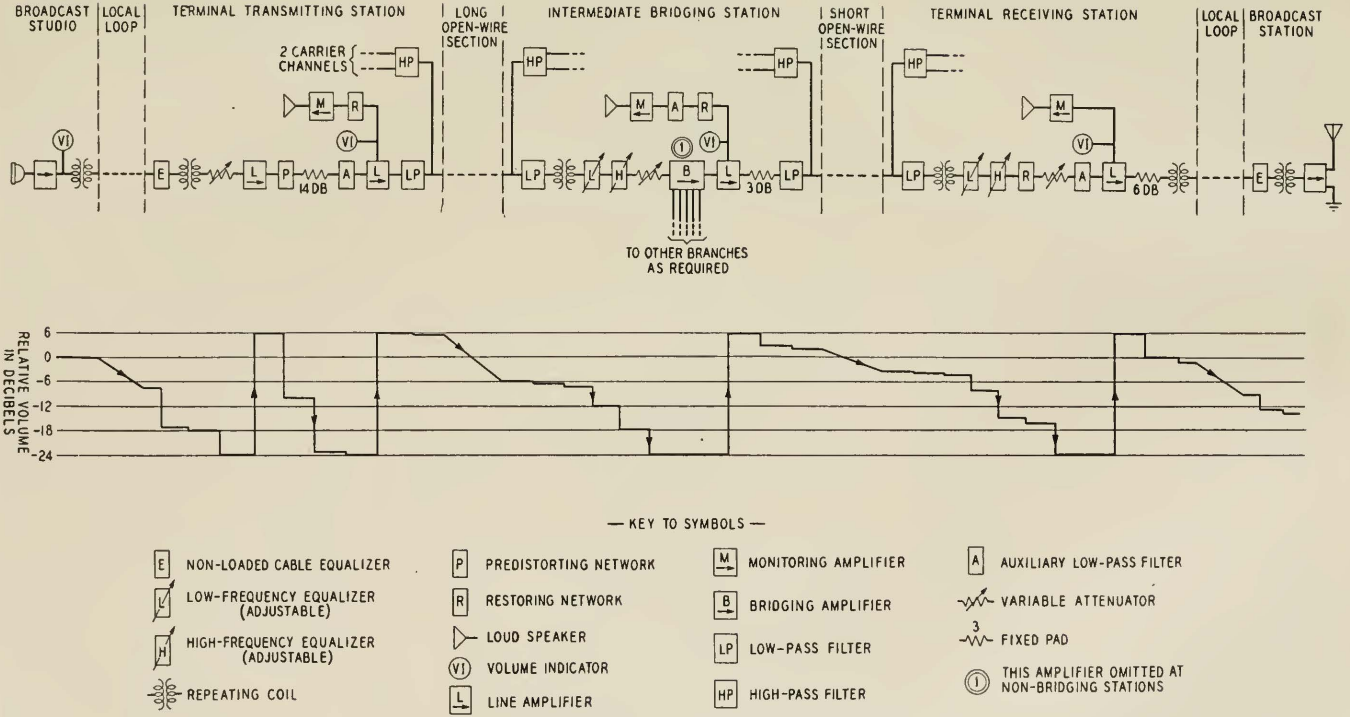


Fig. 1—Typical circuit layout and level diagram.

making the necessary operating observations and a predistorting network and associated amplifier.

At the intermediate station are included line filters for separating the carrier currents and program currents and directing them to their proper channels, two adjustable attenuation equalizers for correcting for the attenuation distortion of the line wires and associated apparatus, gain control attenuator, line amplifier and associated monitoring equipment. At intermediate stations where it is necessary to provide branches to radio stations or to other program circuits an amplifier of a special type having several outlets is inserted immediately in front of the line amplifier.

At a receiving terminal, the layout employed is very similar to that utilized at intermediate stations except that an additional low-pass filter and a restoring network are inserted ahead of the receiving amplifier.

A novel feature is provided in this program system for minimizing its susceptibility to interference at higher frequencies. It consists in predistorting the transmission at the sending end of the circuit so that currents above 1,000 cycles are sent over the line at a higher level than if this arrangement were not employed, thus increasing the signal-to-noise ratio at these frequencies. Such an increase in power at high frequencies is permissible without overloading in the line amplifiers in view of the fact that the energy content of the program material above 1,000 cycles is materially less than at the low frequencies and decreases rapidly as the frequency is increased. In order to restore the program material to the same relations it would have if it were not predistorted, a network is inserted at each point in the branches which feed the radio stations and at the receiving terminal. This network introduces attenuation and phase distortion which are complementary to those introduced at the sending end of the circuit by the predistorting network. The net reduction in high-frequency interference is equal to the discrimination introduced by the predistorting network in favor of these frequencies, and is therefore equal approximately to the loss of the restoring network at the same frequencies.

In the lower part of Fig. 1 is shown a level diagram, from which may be noted the losses and gains introduced by different parts of the system at a frequency of 1,000 cycles. The maximum volumes which are permitted in the various parts of the system are also indicated approximately by this diagram.

Line Facilities

As is well known the open-wire lines employed in telephone and program service do not have uniform attenuation for all frequencies,

the low frequencies being transmitted with much less loss than the high frequencies. Since the program circuits employ the same type of open-wire facilities that is used in the message circuits, three different gauges of wire with either of two pin spacings between wires may be used and the repeater sections may vary in length from 60 to 300 miles. This means that the attenuation frequency characteristic of a repeater section not only varies with frequency but also varies considerably in magnitude of attenuation depending on gauge of wire and length of repeater section.

On Fig. 2 are shown three pairs of characteristics which illustrate the loss-frequency characteristics of three lengths of 165-mil, 8-inch

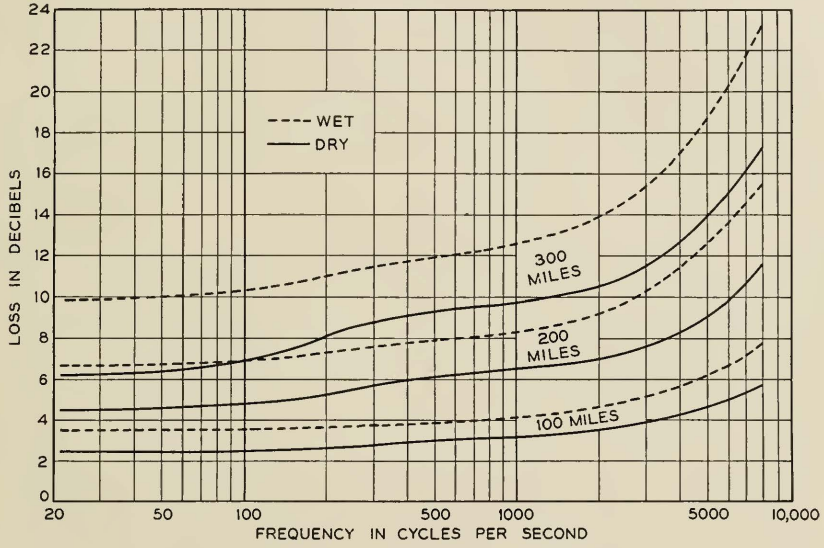


Fig. 2—Loss of 165-mil, 8-inch spaced pairs when inserted between 600-ohm resistances.

spaced circuits. The lengths chosen for purposes of illustration are 100, 200 and 300 miles, respectively. The solid line curves show the insertion loss-frequency characteristics of the circuits for average dry weather conditions when the circuits are connected between 600-ohm resistances. The dashed line curves indicate the wet weather insertion loss characteristics, that is, they indicate the loss-frequency characteristic which might obtain if the lines were very wet for the entire length of a repeater section.

For the purpose of comparing the attenuation frequency characteristics of the different types of open-wire lines, the curves shown on Fig. 3 have been prepared. These characteristics have been plotted so

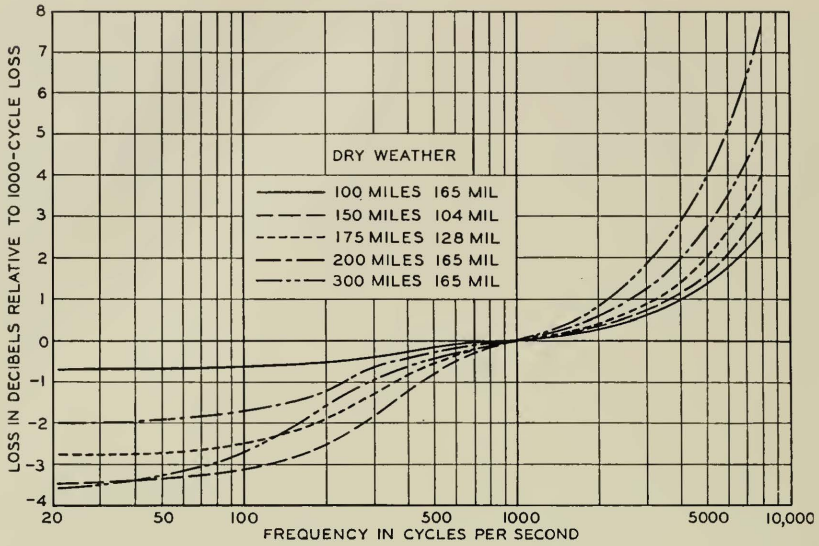


Fig. 3—Attenuation characteristics of 8-inch spaced open-wire pairs when inserted between 600-ohm resistances.

that all coincide at 1,000 cycles; thus a direct comparison of the difference in shape of the attenuation frequency characteristics may readily be observed.

Figure 4 shows resistance and reactance components of 165-mil and 128-mil 8-inch spaced open-wire lines. Note that, except at low frequencies, the impedances of the various open-wire lines are quite uniform throughout the frequency range and do not depart greatly from 600 ohms. For this reason and in consideration that the majority of telephone apparatus is designed for 600-ohm impedance, all units of this new program system, except the carrier line filters, have been designed to have an impedance of 600 ohms. In order to reduce reflection losses, particularly in the carrier range, the line filters have been designed to have an impedance on the line side somewhat lower than 600 ohms although the drop or office side impedance is 600 ohms.

Attenuation Equalizers

To furnish the necessary attenuation corrections for the three different gauges of lines, four adjustable attenuation correcting networks have been provided. One attenuation equalizer provides attenuation correction for high frequencies only and is common for all gauges. The three other equalizers provide low-frequency attenuation correction designed specifically for the particular gauge of circuit the

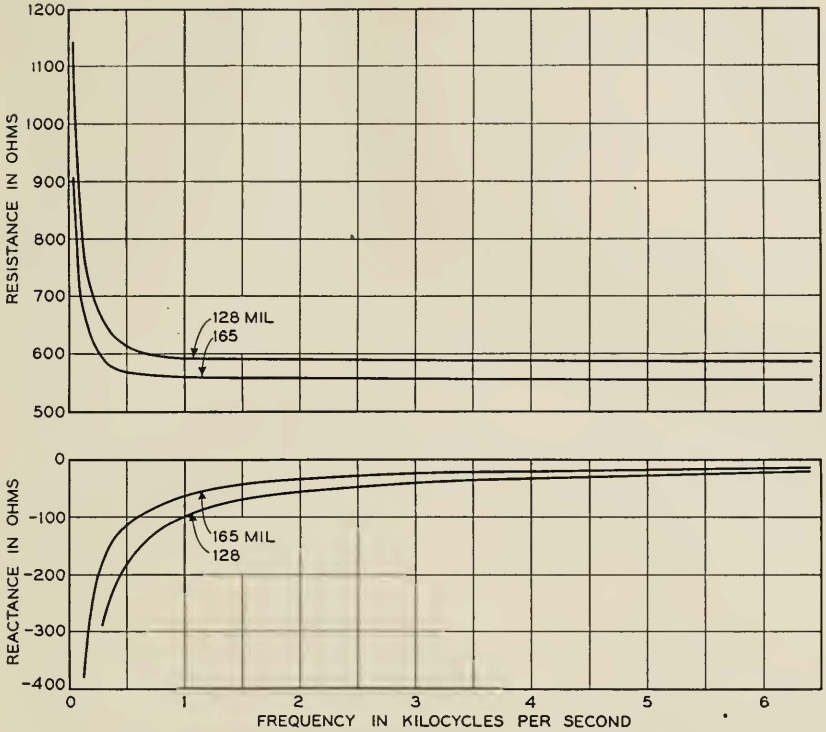


Fig. 4—Impedance of 8-inch spaced open-wire pairs.

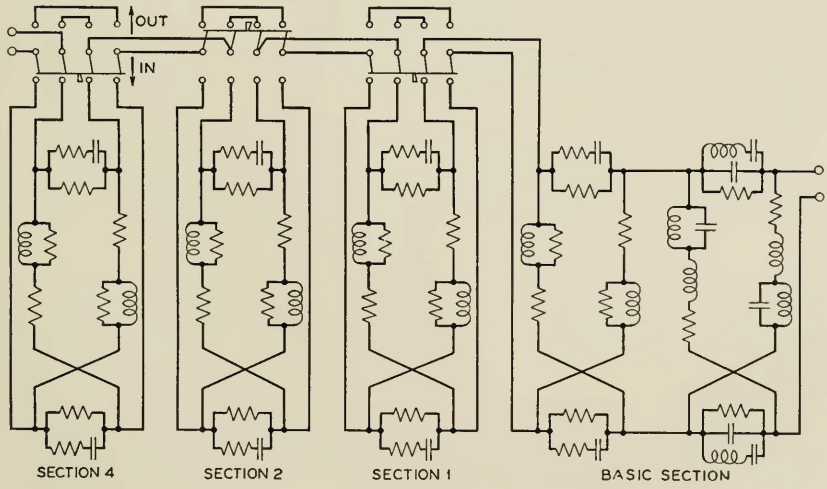


Fig. 5—Low-frequency attenuation equalizer.

equalizer is to be associated with and also include a fixed amount of high-frequency attenuation correction.

On Fig. 5 is shown a schematic diagram of one of the low-frequency attenuation equalizers. This consists of four sections of 600-ohm constant impedance type networks. One section referred to as a basic section introduces attenuation correction over the complete frequency range from 35 to 8,000 cycles for a particular minimum length of line, as for example, in the case of 165-mil circuits this is for 100 miles. The three other sections on the other hand furnish attenuation correction only for frequencies from approximately 1,000 cycles down to 35 cycles. Section 1 of the equalizer for 165-mil circuits puts in about ½ db more loss at low frequencies than it does at 1,000 cycles. Section 2 puts in double the amount of correction that is introduced by Section 1 and Section 4 introduces four times as much attenuation correction as Section 1. These three sections are controlled by switches so that any one or all of them may be cut in tandem with the basic section. The attenuation corrections afforded for the various adjustments of this equalizer are shown on Fig. 6.

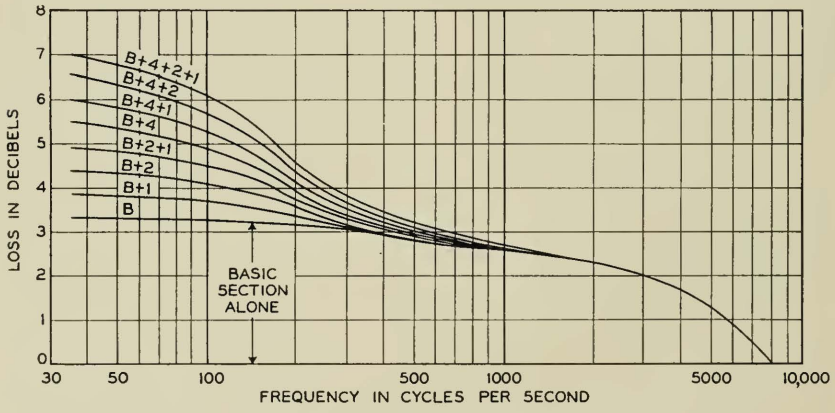


Fig. 6—Attenuation correction furnished by low-frequency equalizer for 165-mil. circuits.

The attenuation equalizers for 128-mil and 104-mil facilities are similar in construction to the one just described having different constants so as to furnish somewhat different attenuation correcting characteristics.

Figure 7 shows a schematic diagram of the high-frequency attenuation equalizer. This consists of four 600-ohm constant impedance type network sections which, as indicated, are controlled by switches so that any one or all of them may be cut in tandem with the program

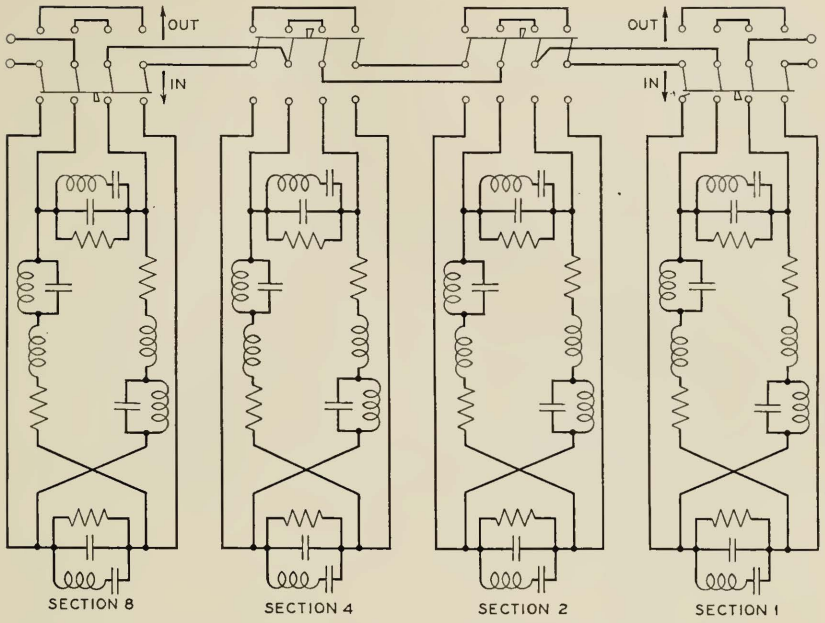


Fig. 7—High-frequency attenuation equalizer.

circuit as required. Figure 8 shows the loss-frequency characteristics of these four sections. As may be noted, the loss of the various sections is practically constant over the frequency range up to 1,000 cycles, decreasing from there on to a minimum value at 8,000 cycles.

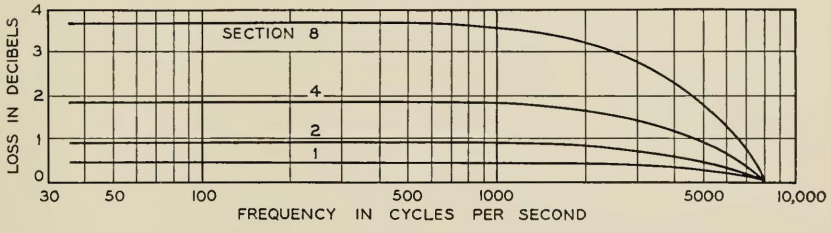


Fig. 8—Attenuation correction furnished by high-frequency equalizer.

Section 1, as may be noted, furnishes about ½ db attenuation correction. Section 2 is double that of Section 1, Section 4 is four times that of Section 1 and Section 8, eight times that of Section 1. These sections may be used in tandem so that attenuation correction for the high frequencies is, therefore, provided in steps of ½ db from zero to 7½ db.

An illustration of how the equalizers introduce the necessary attenuation correction is given on Fig. 9. The lower curve on this figure shows the loss of a 300-mile section of 165-mil circuit. The losses introduced by the particular sections of low and high-frequency equalizers that

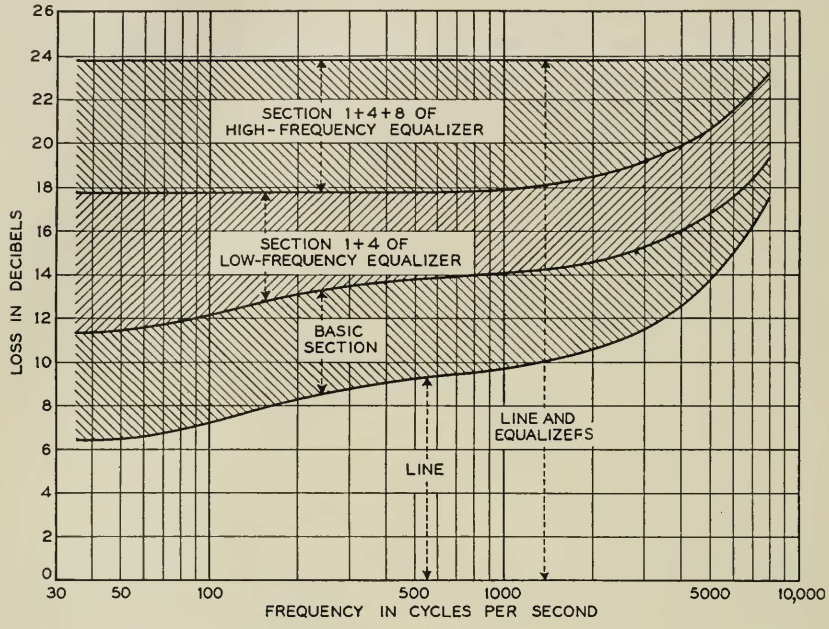


Fig. 9—Loss of 300-mile line section and associated equalizers.

would be required for this length of line are indicated by the cross-hatched areas, and the total line and equalizer loss is shown by the top horizontal line. Sufficient gain is introduced by the line amplifier to annul this loss.

Amplifiers

Two types of amplifiers are provided, one of which is used as a line or monitoring amplifier and the other which is used as a means for transforming one circuit into several circuits so as to feed various branches at points required.

For certain combinations of program circuits as many as 50 amplifiers may be connected in tandem. This necessarily imposes severe requirements on the transmission performances of the amplifiers particularly with reference to flatness of gain-frequency characteristics and phase distortion. By designing the coils used in the amplifiers

so as to have very high inductances the desired phase distortion requirements were met while at the same time the necessary flatness of gain characteristic was obtained at the low frequencies.

Figure 10 shows the transmission circuit of the line amplifier and monitoring amplifier. This device has a 600-ohm input and output impedance and consists of two stages of push-pull amplification. The potentiometer is a balanced slide wire having a continuous gain adjustment over a range of 6 db. A balanced input transformer serves to connect the potentiometer to the grids of the two push-pull vacuum tubes which function as the first stage of this amplifier. The first stage is connected to the second or power stage by means of resistance coupling which gives better results both as to phase distortion and low-frequency gain characteristics than if transformer or retard coil coupling were used. Resistances are provided in the grid circuits of the second stage so that the high-frequency characteristic may be adjusted as required. The power tubes are connected to an output transformer which has the unique feature of providing a monitoring outlet which is not materially affected by voltages produced at or beyond the line terminals. The transformer, as may be observed, consists of three balanced windings arranged as in the form of the well-known hybrid coil used in two-wire telephone repeaters, with the exception that the two low impedance windings are of unequal ratio, the line windings having many more turns than the monitoring windings. The ratio of the windings is such that the voltage at the monitoring terminals when said terminals are closed through 600 ohms is 30 db below the voltage at the line terminals. Resistances are inserted in series with the monitoring winding so that an impedance of 600 ohms will be presented at the monitoring terminals.

The average gain of the amplifier with the potentiometer set at its maximum position is 33 db. Of 100 amplifiers measured, the gain at 35 cycles averaged .10 db less than the gain at 1,000 cycles while from 100 to 8,000 cycles the gain was constant within .05 db. The delay at 50 cycles is approximately .6 millisecond greater than it is at 1,000 cycles. From 150 to 8,000 cycles the delay is substantially constant and is only a small fraction of a millisecond. The amplifier is capable of handling an output power 9 db above reference volume without noticeable distortion.

At several points along a program circuit taps or branches are provided so as to connect various broadcasting stations to the program circuit and also to connect to other program circuits which form part of a broadcasting network. Points where such connections or branches are made are commonly called bridging stations. At some points as

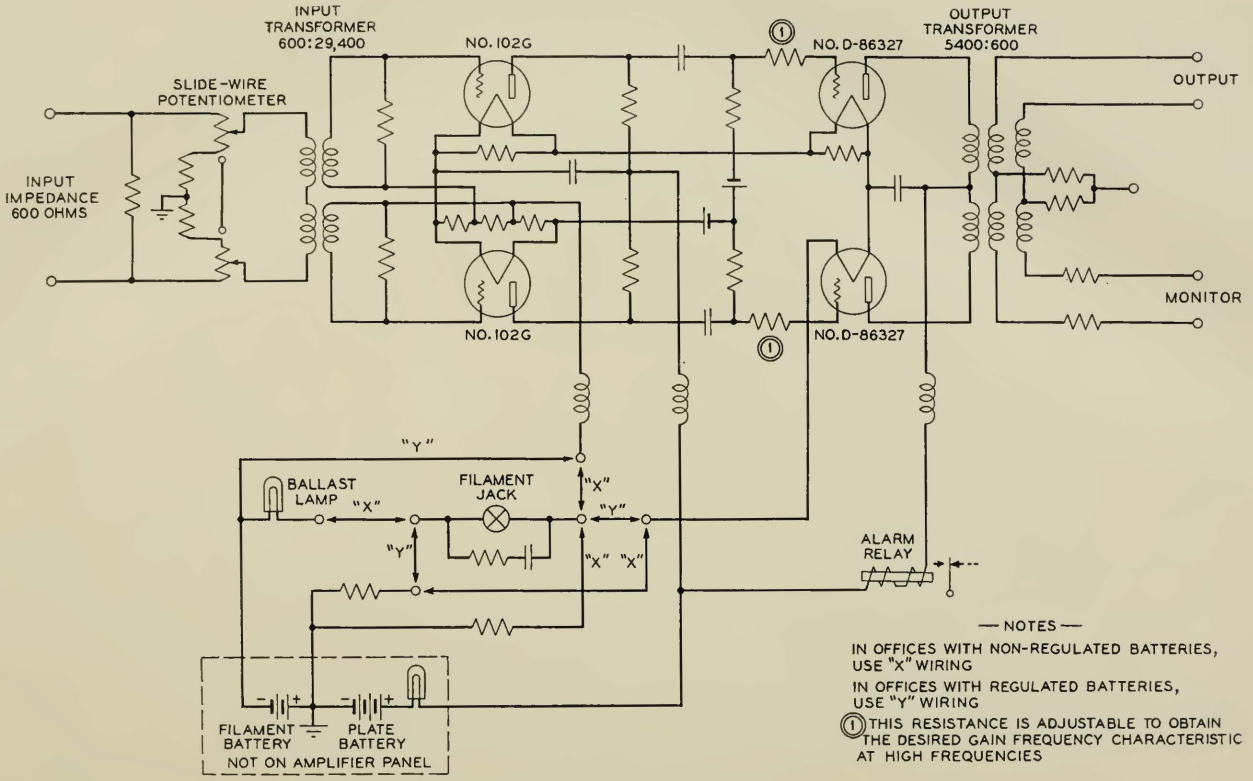


Fig. 10—Schematic of line amplifier.

many as six branches are supplied but generally only two or three taps are utilized.

To accomplish this branching out at a bridging station a resistance network multiple is provided, having six outlets. This network multiple is shown on Fig. 11. To annul the loss of the network a single stage amplifier is connected in front of it. This network multiple and amplifier are mounted on the same panel forming a single integral unit. The network multiple is so proportioned that if any one of the branches is accidentally opened or short-circuited the other branches are affected to only a minor degree. The amplifier is adjusted so that the gain from the input terminals to any of the output branches is zero. The bridging amplifier is normally inserted immediately in front of the line amplifier. As in the case of the line amplifier mentioned above, high inductance coils are utilized in order to keep phase distortion at a minimum. A resistance adjustment is provided in the grid circuit in order to adjust the high-frequency characteristic of this amplifier to the desired value.

The gain-frequency characteristic of the bridging amplifier is practically identical with the corresponding characteristic just described for the line amplifier, while the delay is even less.

Predistortion

The means utilized to accomplish the predistorted transmission referred to earlier includes the provision of a so-called predistorting network at the sending end of a program circuit and a restoring network in each branch which supplies a broadcasting station. The predistorting network introduces a large loss at low frequencies with a decrease in loss as the frequency is increased. By introducing suitable amplification immediately behind the predistorting network the resultant effect is to raise the high-frequency transmission relative to the low-frequency transmission by the difference in loss between the 1,000-cycle loss of the predistorting network and its higher frequency loss. The restoring network characteristic is the inverse of the predistorting network. These two networks are 600-ohm constant impedance type structures. The restoring network is shown schematically in Fig. 12. The predistorting network is generally similar to this, having different constants and a slightly different arrangement of elements. On Fig. 13 are shown the loss-frequency characteristics of the predistorting and restoring networks and a third characteristic which is the sum of these two. As may be noted this latter characteristic has a constant value throughout the frequency range.

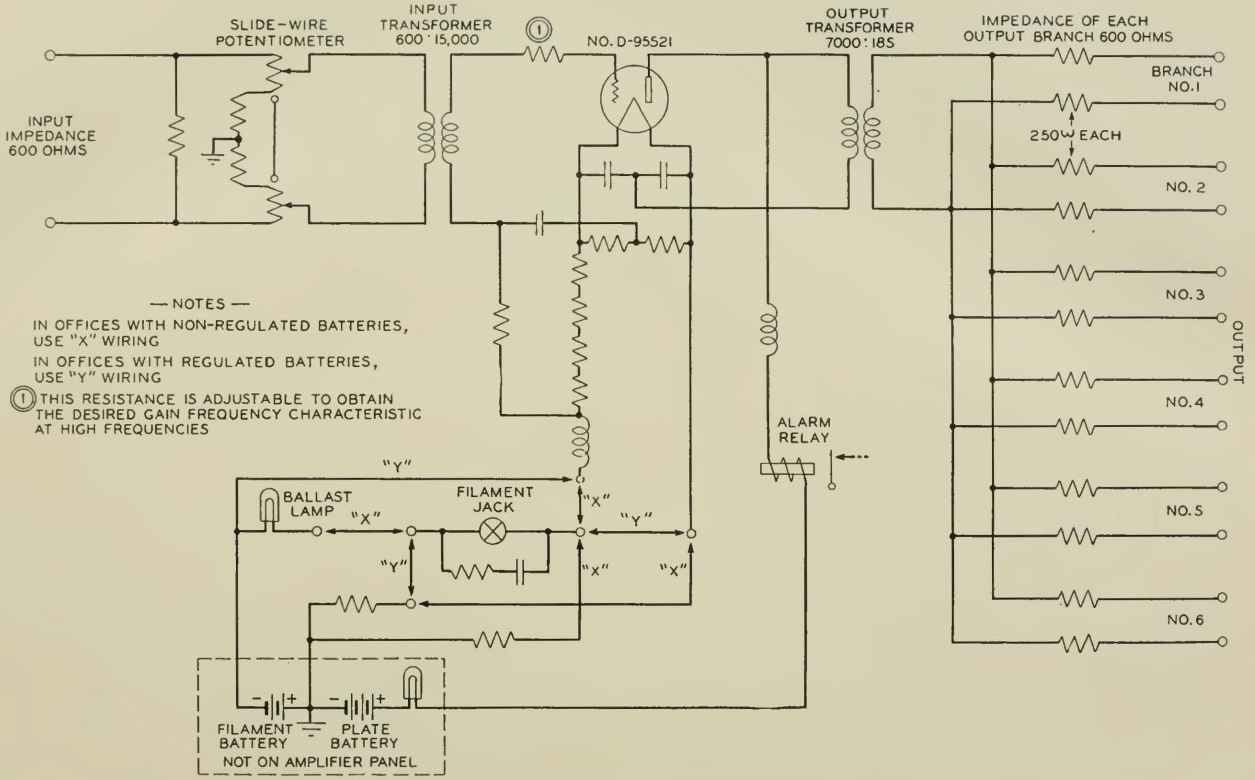


Fig. 11—Schematic of bridging amplifier.

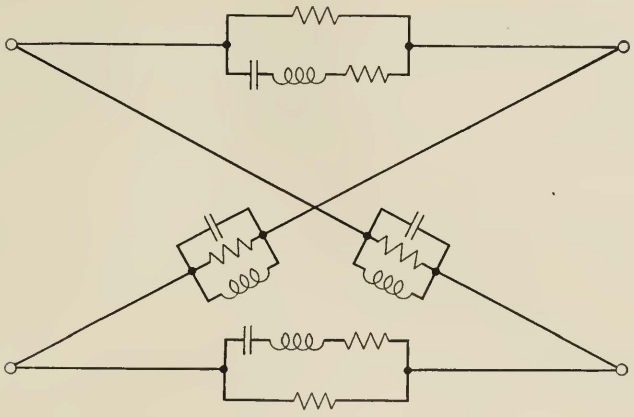


Fig. 12—Restoring network.

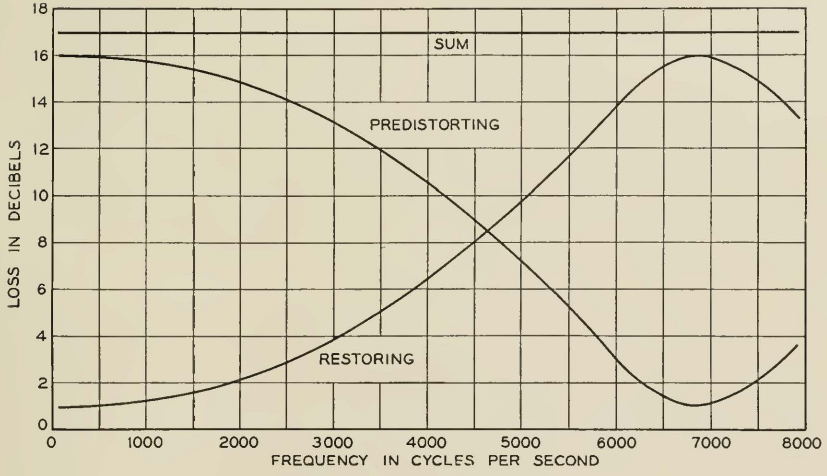


Fig. 13—Attenuation characteristics of predistorting and restoring networks.

Line Filters

As a rule, on open-wire circuits other transmission channels are provided on the same wires which carry the program transmission. These other channels operate at frequencies above the program range and in order to direct the various currents to their proper channels at a terminal or repeater station carrier line filter sets are inserted at the ends of the line wires. The carrier line filter sets include a low-pass and a high-pass filter. The low-pass filter, cutting off somewhat above 8,000 cycles, directs the program transmission to the program

apparatus and the high-pass filter which has a low end cutoff around 9,000 cycles directs the carrier transmission to its associated carrier equipment. Attenuation frequency characteristics of these filters are shown on Fig. 14. The low pass filter is of unusual design and is described at some length in a companion paper.[3]

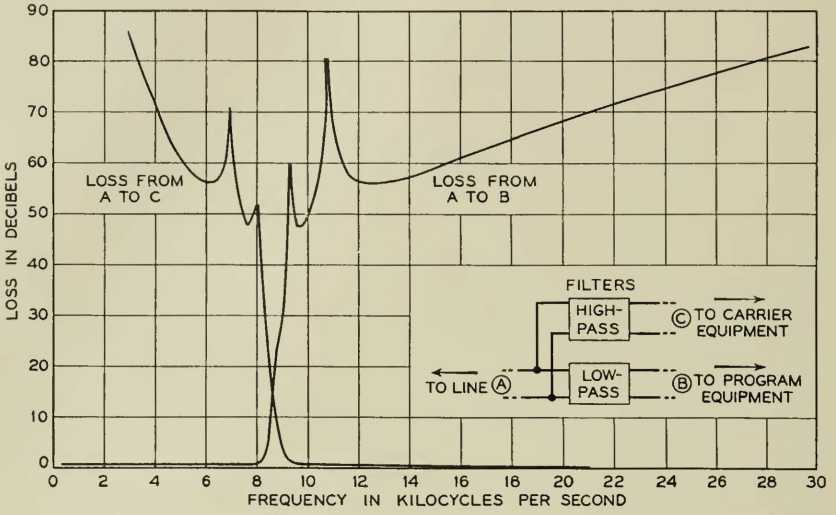


Fig. 14—Attenuation characteristics of line filter set.

Monitoring Features

A very important factor in the satisfactory operation of a program system is the provision of monitoring arrangements by which the operating forces are enabled to observe the quality of transmission, listen for extraneous interferences and observe indicating devices in order to make certain that the program is maintained at its proper volume.

Three types of aural monitoring facilities were provided on a trial basis for the new program system. The first type consists of a single unit loudspeaker operated by a suitable amplifier. With this loudspeaker system a good response characteristic from approximately 100 to 5,000 cycles is obtained, the low-frequency response depending, of course, on the size of the baffle used with the loudspeaker.

The second type of monitoring consists of two headset receivers arranged with a proper equalizing network circuit. This type of monitoring provides good response characteristics from approxi-

[3] A. W. Clement, "Line Filter for Wide-Band Open-Wire Program System," published in this issue of the *Bell Sys. Tech. Jour.*

mately 50 cycles to 8,000 cycles, enabling the observer to cover the entire program frequency range, thus permitting him to detect any extraneous interference which may be introduced even though this occurs at very low or very high frequencies.

The third type of monitoring consists of two loudspeakers and associated equalizing network with the loudspeakers mounted in a large baffle board. This arrangement affords a fairly uniform response from about 40 cycles to above 8,000 cycles. The particular type of monitoring which might be provided at the various stations would be governed by the service requirements involved.

To observe the volume on the program circuit, volume indicators are used. A new type of volume indicator was made available along with the new program system. This new device utilizes a full-wave copper oxide rectifier, has a much greater sensitivity range than that of the devices formerly used and possesses materially improved indicating characteristics. The volume indicator is connected across the monitoring terminals of the line amplifier, in which position it is bridged across a practically non-reactive 600-ohm impedance. Located thus it is also independent of line impedance affording more accurate results and obviating the necessity of correcting volume readings on account of line impedances. Also at this location it introduces no loss or phase distortion to the through program circuit.

The above constitutes a description of the major items employed in this program system. There are a number of other units, such as attenuators, repeating coils, etc., which will not be described in detail here but will be referred to as the need arises.

Typical Station Layouts

Due to the various requirements for different types of service and due in part to the different type of facilities, the general apparatus layouts and arrangements at different repeater stations are not always the same. Several of the more important general or typical layouts will be briefly discussed, however.

On Fig. 15 is shown a layout of a typical intermediate station where bridging is not required and where the gauge of the wires in the two directions is the same. As may be noted from this figure, switching facilities are provided so that the apparatus may be connected into the circuit so as to properly take care of either the east-west or west-east transmission. For this type of layout most of the apparatus is common to both directions of transmission. The fixed artificial lines or pads indicated by Note 1 on Fig. 15 are for the purpose of building out whichever line has the lower 1,000-cycle attenuation so that this

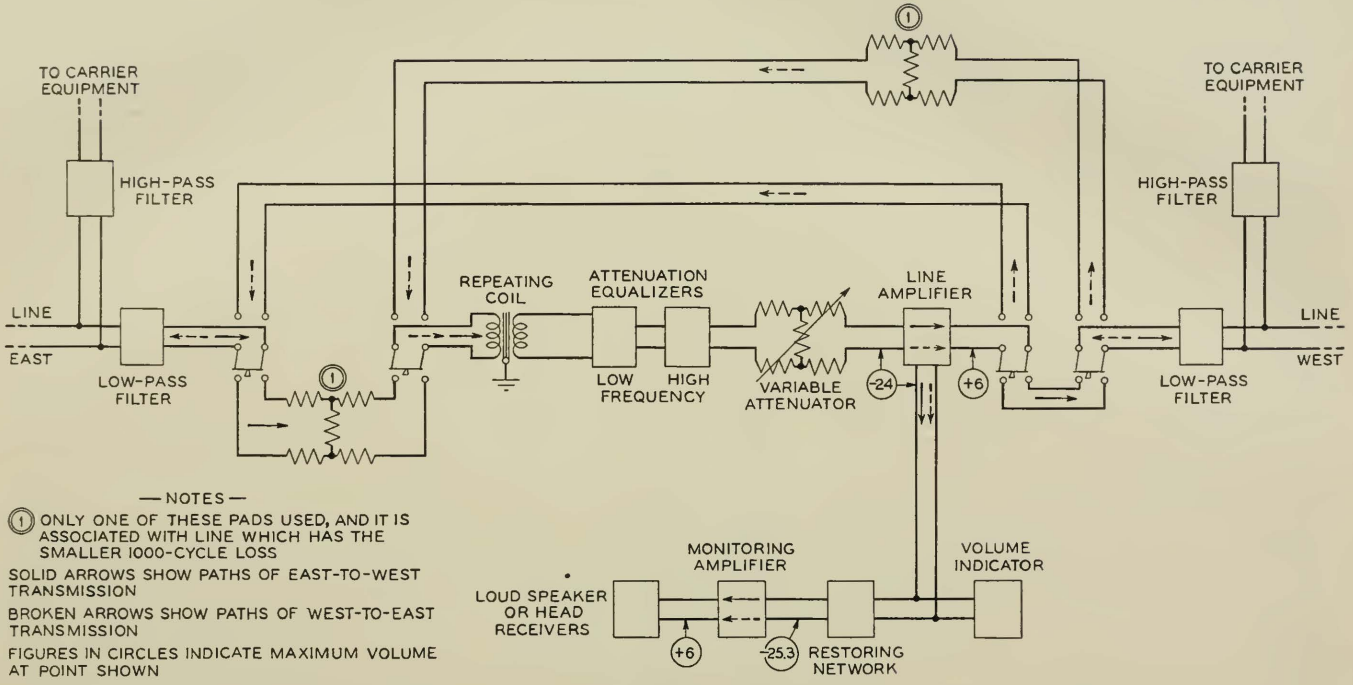


Fig. 15—Intermediate station layout.

line and associated pad will have the same 1,000-cycle loss as the other line. As indicated, only one of these pads is required. This building out of the shorter line minimizes attenuator adjustment when the direction of transmission is reversed. The line amplifier in this, as well as the other layouts to be discussed, is always set for a gain of 30 db.

On Fig. 16 is shown the layout of a typical intermediate non-bridging station where the gauges of the wires on the two sides of the repeater station are different. As mentioned earlier each gauge of wire has its own particular low-frequency attenuation equalizer. Consequently, where the gauges of the wires on the two sides of the repeater station are not alike, it is necessary to arrange the station layout so that the proper low-frequency equalizer will be associated with the proper direction of transmission. This association of apparatus may be readily observed from Fig. 16.

On Fig. 17 is shown the layout of a typical terminal station. This layout differs from the intermediate station layout largely in the fact that provision must be made for the introduction of predistortion when the terminal station is transmitting a program to the open-wire line and in the provision of a restoring network when the terminal station is receiving a program from the open-wire line. The general layout of the apparatus may readily be observed by reference to the figure. The monitoring facilities at this type of station, in general, differ from those provided at the normal intermediate station in that a two-unit loudspeaker is provided for use as desired.

On Fig. 18 is shown the layout of a typical intermediate bridging station where the gauge of the wires in the two directions is the same. This arrangement differs largely from the arrangement shown on Fig. 15 in that the bridging amplifier is inserted immediately ahead of the line amplifier so as to provide the necessary additional branches as required. The general circuit arrangements involved to take care of the different types of branches which may be encountered are indicated on this figure. The photograph, Fig. 19, shows the program equipment layout at an intermediate bridging station, which is of the type just discussed in Fig. 18, utilizing, however, only one branch circuit which is connected to a local broadcasting station.

In certain of the layouts just discussed, one apparatus unit designated as "Aux Filter" is shown which has not previously been mentioned. This is an 8,000-cycle low-pass filter somewhat similar to the low-pass line filter, except that it is not designed to operate in parallel with any high-pass filter. This filter is required at the transmitting and receiving terminals, in the branches feeding the radio station and

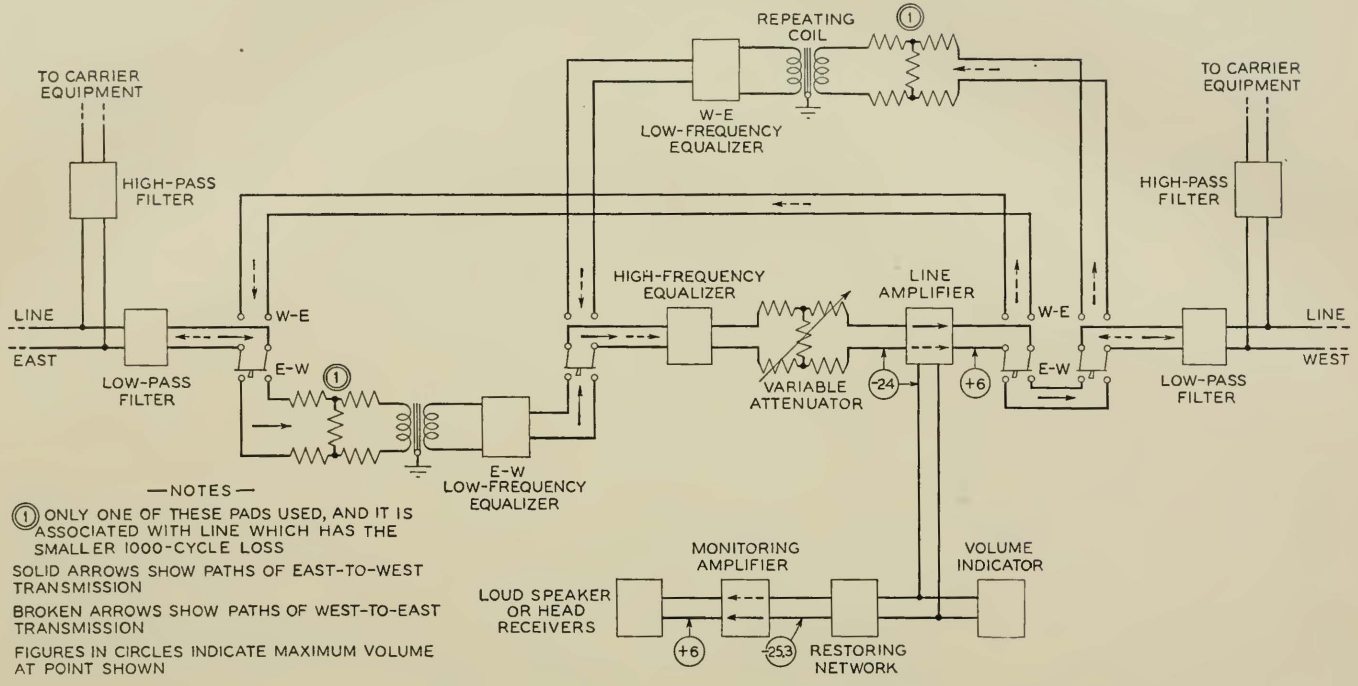


Fig. 16—Intermediate station layout; different gauges.

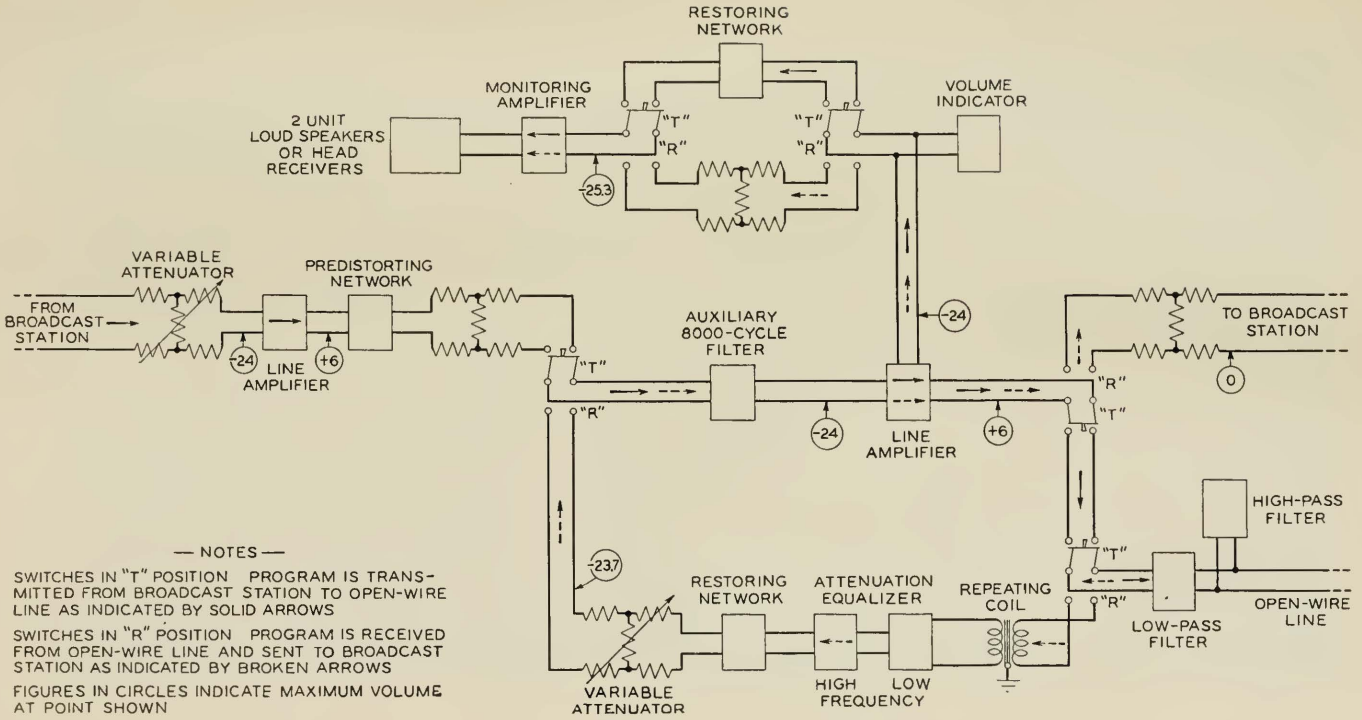


Fig. 17—Terminal station layout.

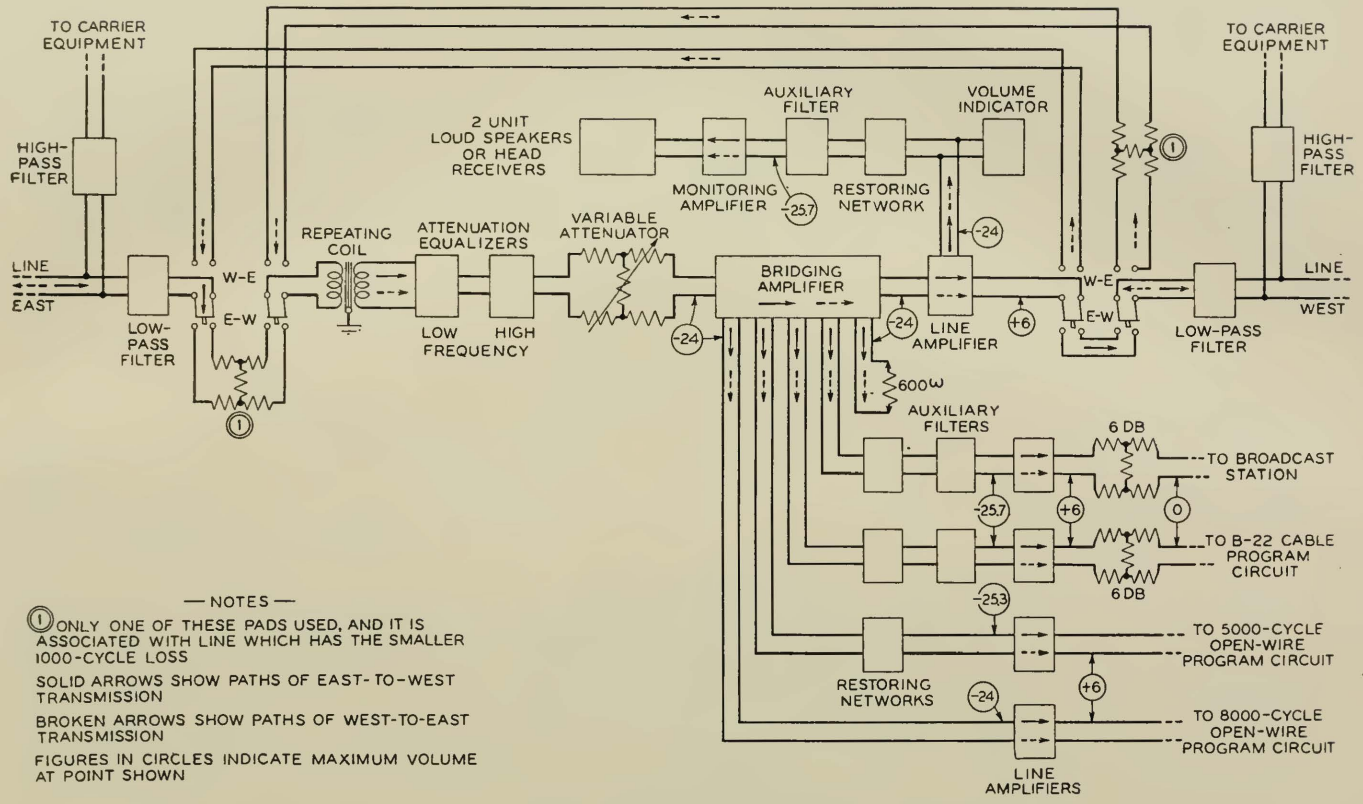


Fig. 18—Bridging station layout.

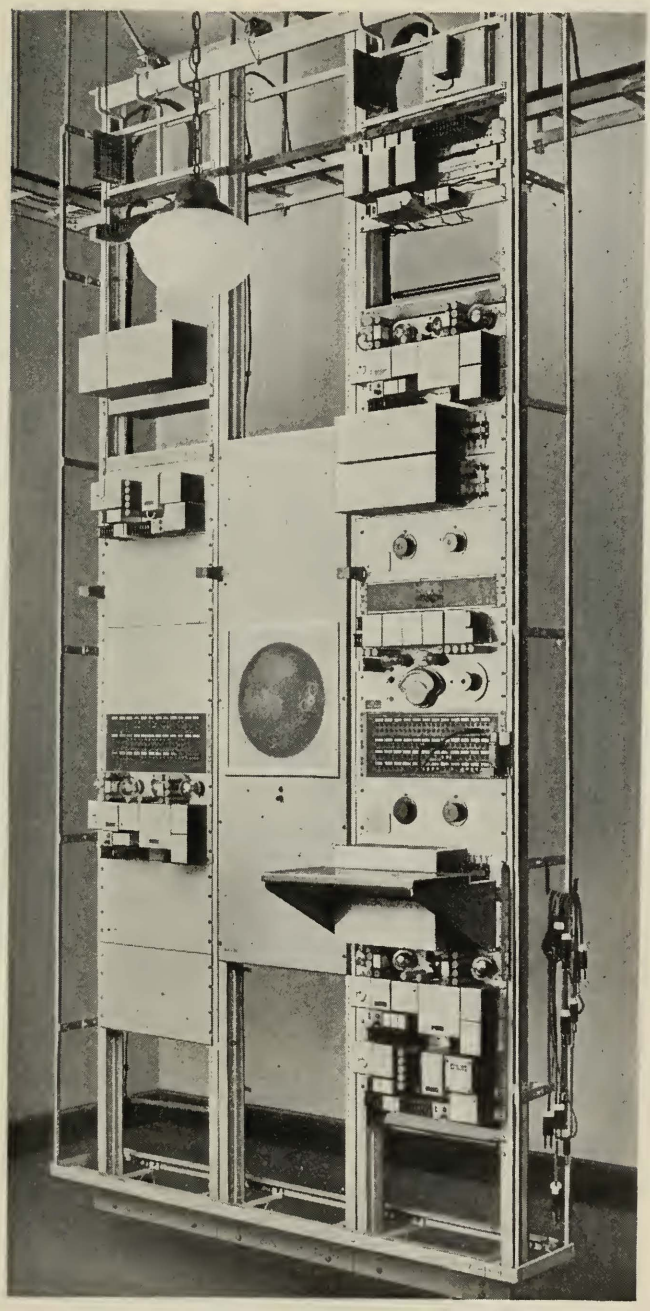


FIG. 19—Intermediate bridging station bay layout.

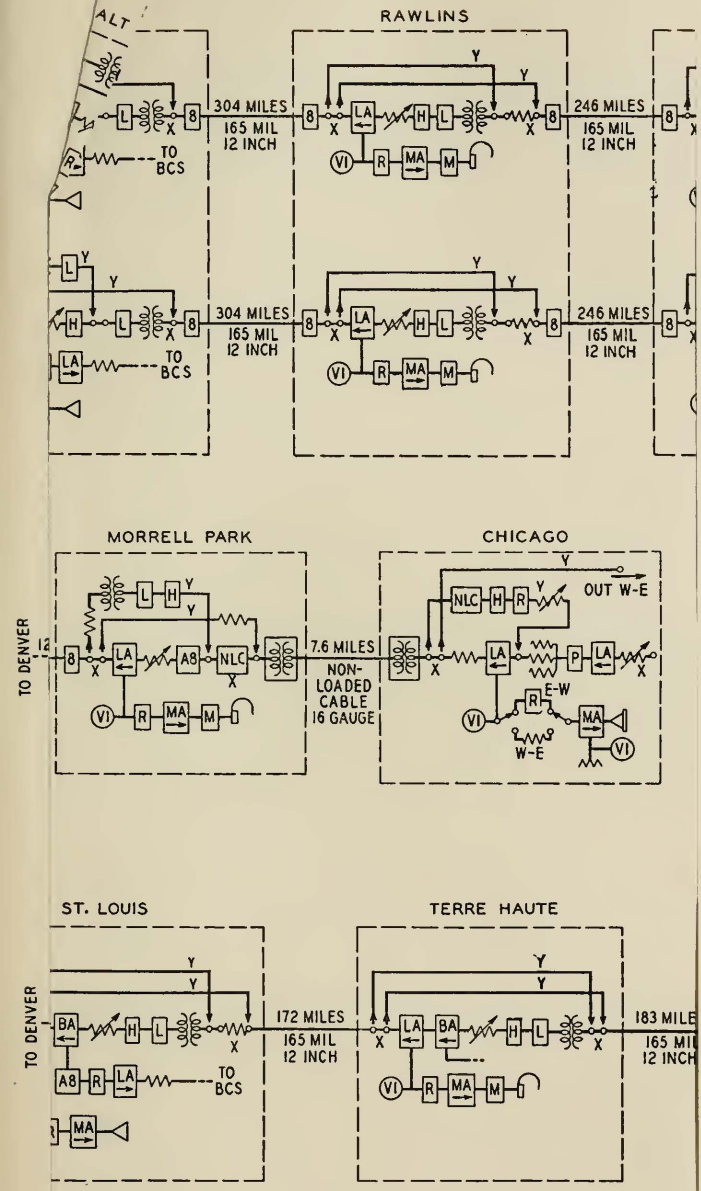
also in the high quality monitoring circuit to afford additional discrimination against unwanted high-frequency interference as, for example, interference from the carrier channels. This arrangement of splitting the filter requirements enables a less expensive type of line filter set to be employed.

Overall Performance

The initial application of this new program system was made on two transcontinental circuits between Chicago and San Francisco. One circuit, referred to as circuit 1, was routed through Omaha and Denver over the central transcontinental line. The other circuit, referred to as circuit 2, was routed via St. Louis and Kansas City to Denver and thence over the same pole lead as circuit 1. The layout of these two circuits is shown in Fig. 20. Circuit 1 was approximately 2,395 miles long and was routed through 17 repeater stations involving 23 amplifiers in tandem. Circuit 2 was approximately 2,689 miles long and was routed through 19 repeater stations involving 29 amplifiers in tandem. Both circuits were routed through B-22 cable facilities between Sacramento and Oakland, California, and non-loaded cable facilities in the transbay submarine cable between Oakland and San Francisco.

At San Francisco a listening studio was set up in the Grant Avenue office where the program circuits terminated. A two-unit loudspeaker with suitable connecting networks was set in a 7′ x 7′ baffle, the response of this loudspeaking system being practically uniform from about 40 cycles to above 8,000 cycles. The room in which the loudspeakers were located was acoustically treated so as to obtain the proper reverberation time. A powerful amplifier having a flat gain-frequency characteristic from 35 cycles to well above 8,000 cycles supplied the loudspeaker system. A high quality phonograph system for furnishing test programs was also installed at the Grant Avenue office. The records used were of the vertical cut type and included several recordings of a 75-piece orchestra as well as various solo and instrumental recordings. Two outside pickup points were used, one at the studios of one of the broadcasting companies at San Francisco and the other at a hotel. At both of these places the moving coil type of microphones was used and the latest type of high quality pickup amplifiers. The pickup system used at both these places had a response characteristic within about 2 db of being flat over the range of 35 to 10,000 cycles.

Figure 21 is a photograph showing the special equipment placed in the Grant Avenue office for carrying out the various overall tests and


RAWLINS
304 MILES
165 MIL
12 INCH
246 MILES
165 MIL
12 INCH
TO
BCS
304 MILES
165 MIL
12 INCH
246 MILES
165 MIL
12 INCH
TO
BCS
MORRELL PARK
CHICAGO
TO DENVER
7.6 MILES
NON-
LOADED
CABLE
16 GAUGE
OUT W-E
E-W
W-E
ST. LOUIS
TERRE HAUTE
TO DENVER
172 MILES
165 MIL
12 INCH
183 MILE
165 MI
12 INCH
TO
BCS

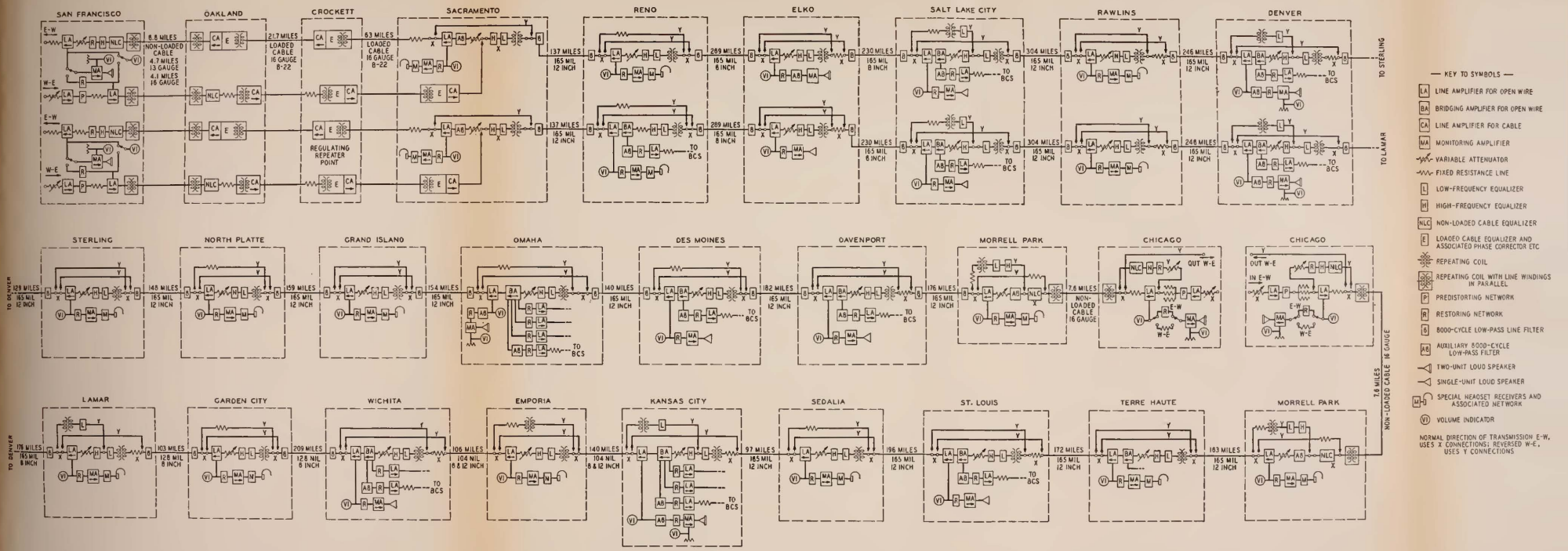


Fig. 20—Circuit layout for trial of wide-band open-wire program system.



Fig. 21—Special apparatus bay layout.



Fig. 21—Special apparatus bay layout.

also shows the new equipment provided at San Francisco on the two program circuits under discussion. The three right-hand bays accommodated the special equipment.

In making transmission measurements, the circuit under test was first split up in a number of sections and each section was then measured at four test frequencies, namely, 50, 100, 1,000 and 7,000 cycles. If the results were not within required limits the attenuators and equalizers were readjusted as required. The various sections were then connected together and the overall circuit measured at several frequencies. Figure 22 shows the transmission-frequency character-

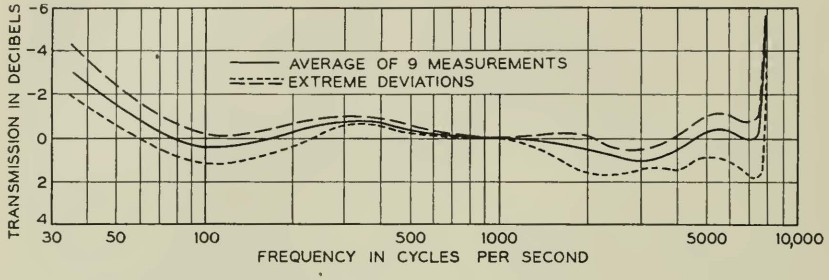


Fig. 22—Transmission frequency characteristics of circuit No. 1, Chicago to San Francisco.

istics of circuit 1. The solid line is the average of nine measurements while the dashed lines show the extreme deviations obtained for any of the nine measurements. Figure 23 shows corresponding data for

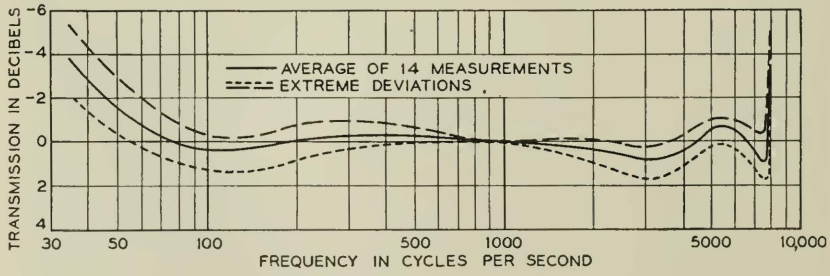


Fig. 23—Transmission frequency characteristics of circuit No. 2, San Francisco to Chicago.

circuit 2. For comparison purposes the average characteristics of the two circuits separately and the two of them connected in tandem making a loop circuit of over 5,000 miles are shown on Fig. 24.

Other measurements were made to determine whether non-linear effects were produced. For example, two frequencies were applied

to the circuit, one being measured and the other alternately cut off and on to determine whether one frequency adversely affected the transmission of the other or produced undesirable sum and difference products. Such distortion effects were found to be small. Measurements were made to determine whether the overall transmission varied with the load applied. With a testing power which was varied in magnitude from 50 milliwatts to .1 milliwatt, the transmission varied slightly more than 1 db, that is, with the heavy load the circuit loss was somewhat more than 1 db greater than at the light load.

A noise and crosstalk survey was made on these program circuits and on message circuits on the same pole lead. Observations were made at the terminals of the message circuits while a program was being transmitted on the program circuits to determine the amount of

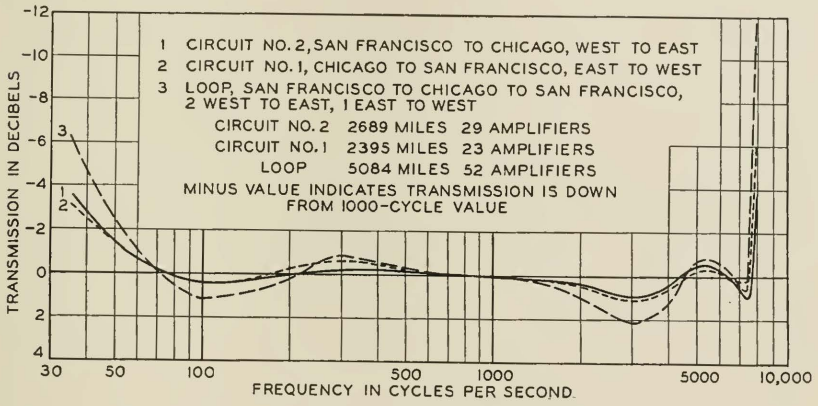


Fig. 24—Average transmission frequency characteristics.

interference introduced into the message circuits from the program circuits, and, conversely, observations were made on the program circuits while various paralleling message circuits were in use, and the resulting interference was recorded.

The noise or crosstalk volume on the program circuits was measured by means of a volume indicator, which had inserted between it and the circuit at the point of measurement a network having a loss-frequency characteristic such that the various frequencies affecting the meter reading were attenuated or weighted in much the same way that the ear weights the different frequencies. Crosstalk volume and noise on the message circuit were measured with an indicating meter in much the same manner except that the network used here had an attenuation frequency characteristic corresponding very nearly to that of the ear and an average telephone set. The network used on the program

circuits was referred to as a "program weighting network," while that used with the message circuit was the ordinary "message weighting network." The noise and crosstalk volume was then recorded in db referring to reference noise with either program weighting or message weighting. Reference noise is that amount of interference which will produce the same meter reading as 10^{-12} watt of 1,000-cycle power, which is 90 db below 1 milliwatt.

The results of this survey indicated that in consideration of the layout and levels of the existing message circuits and of the noise existent on these circuits and on the program circuits, the value for maximum program volume, should, under normal conditions, be + 3 referred to reference volume; that is, at this value the best balance between program to message crosstalk and program circuit noise would result. It was also determined that on very long sections, or on sections where all circuits were subjected to severe noise exposure, the maximum volume on the program circuits could be increased 3 db to improve the signal-to-noise ratio on the program circuits. This higher volume could be permitted in these cases since on the longer sections the message circuits also usually operate at higher levels, and on the especially noisy short sections the increased crosstalk to the message circuits will ordinarily be masked by the greater noise.

The average noise measured at San Francisco or Chicago at the circuit terminals at the reference volume point was 49 db above reference noise "program weighting" when the restoring network was included at the receiving terminal. The noise averaged 5 db higher than this with the restoring network removed. This value of noise is about 43 db below the maximum power of the program measured at the same point with the same measuring instrument. This, therefore, establishes a signal-to-noise ratio of about 43 db, thus permitting a volume range of approximately 40 db.

The various tests referred to gave statistical data concerning the transmission performance of the circuits from which it could readily be predicted that the circuits would transmit programs with very little impairment to quality. To substantiate this, very critical listening tests were made, comparing the quality of a program after it had been transmitted over various length circuits with the same program transmitted over a reference circuit which was distortionless over the frequency range for which the circuits were designed, namely, to 8,000 cycles. Figure 25 shows schematically the terminal arrangements employed at San Francisco for these listening, or, as they are more commonly called, comparison tests.

Various types of programs were used, such as speech, vocal and

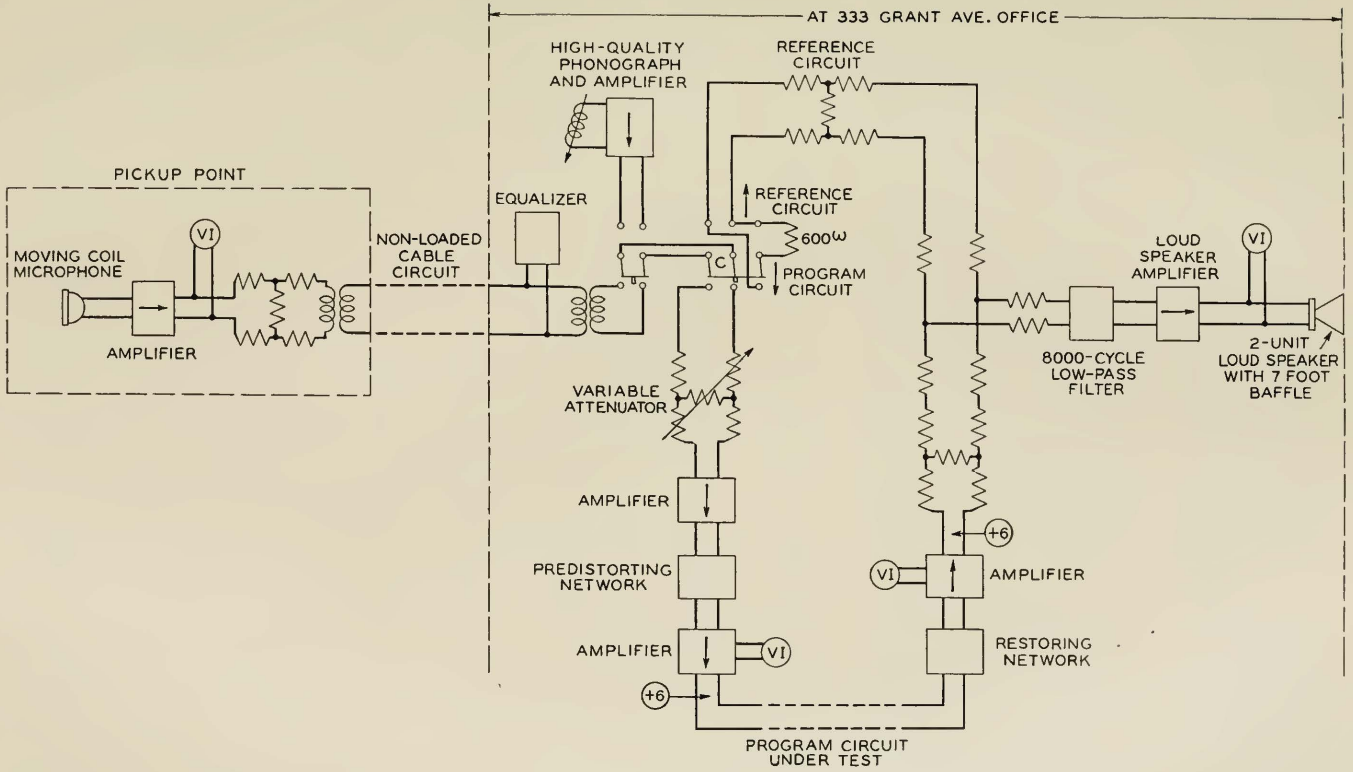


Fig. 25—Terminal arrangements at San Francisco for comparison tests.

instrumental selections, orchestral renditions, both classical and jazz. Quite a number of observers were employed, some of whom were present on several tests and a few on all tests. On tests made on a San Francisco–Denver–San Francisco loop involving 2,600 miles of circuit, no observer was able to consistently differentiate between the quality over the reference circuit and that over the program circuit. On tests made on the San Francisco–Chicago–San Francisco loop certain of the more experienced observers were able to differentiate between the circuits somewhat more than 50 per cent of the time, but this, it must be remembered, was on a direct comparison test. None of the observers could tell with any assurance which was the program circuit and which was the reference circuit if a few minutes were allowed to elapse between switches. On the Chicago loop 264 observations were made on direct comparison tests on which 60 per cent of the observations favored the reference circuit and 40 per cent favored the program circuit.

Included as part of the overall program, were tests to determine the volume range, maximum volume obtainable and speed with which the circuits could be reversed.

On the volume range tests a source of program was obtained and so regulated that it had a very narrow volume range. This was then applied to the circuit with the sending end gain adjusted so that the maximum volume applied at the repeater outputs was + 6. The sending end gain was then gradually decreased so as to apply a gradually decreasing volume to the circuit. This process was continued until the program volume was so weak that the line noise interfered with its satisfactory reception. The amount that the sending end gain was adjusted determined the volume range. The average value for several tests was slightly in excess of 40 db. The maximum volume was determined by switching a 10 db pad from the sending end to the receiving end of the circuit and listening to a transmitted program, noting the point at which there was a quality difference between the high volume and low volume condition. It was found that a slight difference could be detected when the maximum volume on the high volume condition was + 10, thus showing the circuit was capable of handling a maximum volume slightly lower than this value.

As mentioned earlier, switching means are provided at each station for reversing the direction of transmission. On the initial field tests it was demonstrated that the circuits could be reversed readily and at the same time maintain satisfactory overall characteristics. At the present time, on receipt of proper advance notice, the circuits are being reversed on commercial programs in approximately 30 seconds.

Conclusion

The above development provides a program transmission system applicable to open-wire lines which even for very long distances will provide transmission characteristics which should be adequate for program transmission for a number of years to come.

Acknowledgment

The author makes grateful acknowledgment to his associates in the Bell Telephone Laboratories, to members of the Long Lines Department of the American Telephone and Telegraph Company, and of the Pacific Telephone and Telegraph Company, for their cooperation in connection with the setting up of the circuits and participation in many of the tests, and to the National Broadcasting Company and the Columbia Broadcasting System for their assistance in making available the program pickup sources at San Francisco.

Line Filter for Program System *

By A. W. CLEMENT

Open wire circuits recently have been developed for transmitting radio broadcast programs with greater naturalness and over greater distances than heretofore.[1] The simultaneous utilization of these circuits for the transmission of broadcast programs and carrier telephone messages requires the use of line filters to restrict the program and carrier currents to the proper circuits. The low pass line filter developed for the program circuits and its contribution to the maintenance of good quality in the programs transmitted are described in this paper.

PROGRAM transmission systems operated on open wire telephone lines ordinarily are not assigned the exclusive use of the lines, but usually share them with other communication facilities. The wide-band system described in an accompanying paper [1] transmits currents in the frequency band extending from 35 to 8,000 cycles per second, while the lines over which it is routed possess useful transmission ranges extending from 35 to considerably above 30,000 cycles. In order that the range above 8,000 cycles shall not be wasted, carrier telephone systems utilizing these frequencies usually are operated on the same wires with the program systems.

Line filters are used at each terminal and repeater point in the program system to separate the program currents from the carrier currents and to guide each to the proper channel. They are operated in sets consisting of a low-pass filter and a high-pass filter connected in parallel at one end, the end that faces the line. The low-pass filter transmits the program currents freely while effectively excluding the carrier currents, and the high-pass filter transmits the carrier currents while excluding the program currents.[2]

The line filters are located in the open-wire program systems as shown in Figs. 1, 15, 16, 17, 18, and 20 of the accompanying paper by H. S. Hamilton.[1] The low-pass filter is in the direct path of the program currents and therefore has a number of features of special interest. It is the object of this paper to describe this filter and its contribution to the maintenance of good quality in the programs transmitted over the system.

* Published in April, 1934 issue of *Electrical Engineering*. Scheduled for presentation at Pacific Coast Convention of A. I. E. E., Salt Lake City, Utah, September, 1934.

[1] "Wide-Band Open-Wire Program System" by H. S. Hamilton, published in this issue of the *Bell. Sys. Tech. Jour.*

[2] "Telephone Transmission Networks" by T. E. Shea and C. E. Lane, published in *A. I. E. E. Transactions*, Vol. 48, 1929, pages 1031–1034.

This low-pass line filter, with its associated high-pass filter, makes it possible to use the open-wire lines simultaneously for wide-band program service and for commercial carrier telephone service, without impairing the quality of the program. It represents an improvement over older types of line filters, as well as an advance in the technique of equalization in filters. In cases requiring careful delay and loss equalization, it has been the usual practice to design the filter first to supply the required discrimination or filtering action, and then design a delay corrector to correct for the delay distortion in the filter, after which a loss equalizer is designed to correct for the amplitude distortion in both the filter and the delay corrector. The loss equalizer introduces a small delay distortion which usually can be anticipated and corrected in the delay corrector. In the wide-band program filter the functions usually performed by these three separate types of networks have been combined, with a consequent saving in cost and space.

Requirements to Be Met by Program Filter

To function effectively as a line filter, the low-pass filter must provide sufficient discrimination against carrier currents to make their effect completely inaudible in all the receivers connected to the program system. Discrimination varying from 46 to about 90 db is necessary to accomplish this end. Because of the presence of an auxiliary low-pass filter [1] which supplies considerable loss in the frequency ranges where the requirement is unusually severe, each line filter need furnish discrimination varying only from 40 to about 60 db.

From the standpoint of program quality, it is essential that the line filter, while furnishing the foregoing discrimination, shall not introduce any appreciable distortion into the program. This requirement would call for nothing unusual in the way of filter design if there were only a few filters in the system. Long open-wire program systems, however, may extend as far as 3,000 or 4,000 miles, and may contain as many as 50 low-pass line filters. A program that has traversed such a circuit still must be comparable in quality to a program that is broadcast from the point at which it originated. Since the system contains much other apparatus, such as equalizers and amplifiers, each low-pass line filter can be permitted to introduce not more than about 1/100 of the distortion that can be tolerated in the whole system, assuming 50 filters in the system.

There are two types of distortion that must be controlled very carefully in the program filter: these are (1) amplitude distortion, and (2) delay, or phase, distortion. Amplitude distortion is introduced

by a filter when its loss is not the same at all frequencies in the transmitted band, currents of some frequencies being attenuated more than others. The effect of amplitude distortion on the program is to change the relative intensities, or volumes, of tones of the frequencies at which distortion occurs, thus impairing the naturalness of the program. Amplitude distortion ordinarily can be corrected without much difficulty by means of suitable attenuation equalizers.

Delay distortion is introduced by a filter when different frequency components of a signal require different lengths of time for propagation through the filter. This type of distortion is related directly to the shape of the phase shift-frequency characteristic. The slope of this phase shift curve usually is taken as a measure of the delay introduced by the filter. Stated mathematically, the delay in seconds is taken as $\partial B/\partial\omega$, where B is the phase shift in radians and ω is $2\pi f$, f being the frequency in cycles per second. Thus if the phase shift of the filter is proportional to frequency, $\partial B/\partial\omega$, or the delay, is constant and there is no delay distortion. In this case the wave form of a signal transmitted through the filter remains unchanged, the signal being delayed in transmission an interval of time corresponding to the slope of the phase shift curve. If the slope of this curve is not constant over the transmitting band of the filter, however, delay distortion is introduced. In low-pass filters, the difference between the slope of the phase shift curve at a given frequency and the minimum slope of the curve is a measure of the delay distortion at that frequency.

A discussion of delay distortion in telephone apparatus, including filters, as well as a discussion of the effect of delay distortion on telephone quality, may be found in two recent articles on these subjects.[3, 4] Whereas the effect of amplitude distortion is to weaken or strengthen some of the tones in the sound being transmitted with respect to the other component tones, the effect of delay distortion is to introduce unnatural audible effects which may become so pronounced as to be annoying if the delay distortion be great enough.

Delay distortion is present in most filters used in communication work, but ordinarily not in such magnitude that its effect is noticeable. As a rule, it need be considered only when a large number of filters is used in a single circuit, as in the case of the program systems. Delay distortion is in general more difficult to correct than amplitude distortion. One of the unusual features of the low-pass line filter used in the wide-band program circuits is the means employed to keep it free from delay distortion.

[3] "Phase Distortion in Telephone Apparatus" by C. E. Lane, *Bell. Sys. Tech. Jour.*, July, 1930.

[4] "Effects of Phase Distortion on Telephone Quality," by J. C. Steinberg, *Bell Sys. Tech. Jour.*, July, 1930.

The filter consists of four parts, each with distinguishing functional characteristics. The separate parts, or sections, have image impedances such that when they are joined together no current is reflected at the junctions. Figure 1 shows the filter in block schematic form. Each part of the filter provides some of the attenuation required to exclude carrier currents from the program circuit, the attenuation of the complete filter being the sum of the attenuations of all parts. On Fig. 2 are shown the loss-frequency characteristics of the various sections and of the complete filter.

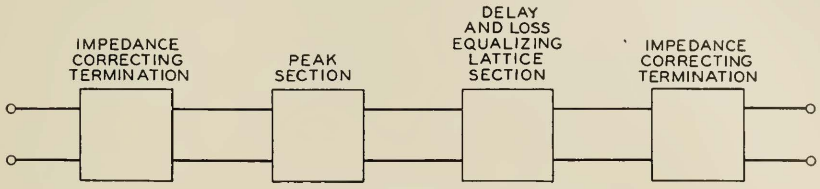


Fig. 1—Block schematic diagram of filter.

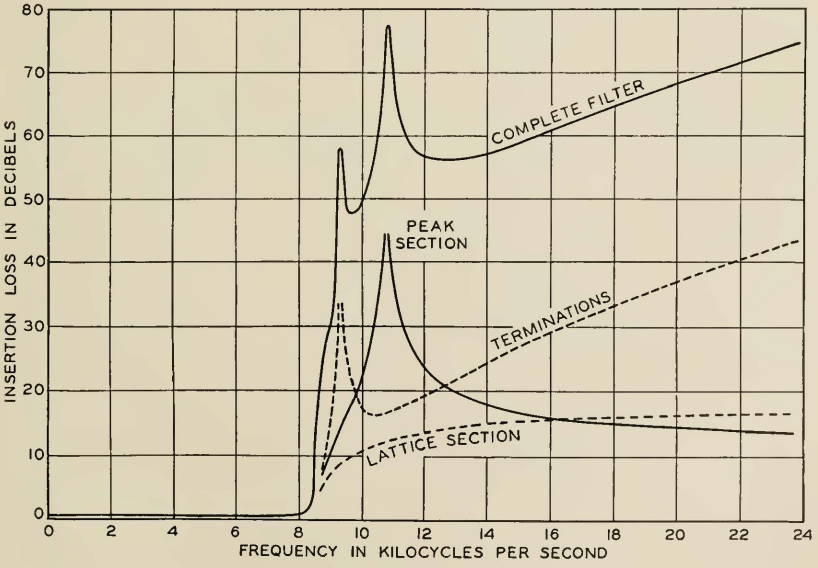


Fig. 2—Loss in filter and in component sections.

Delay Equalization

Likewise, the phase shift of the complete filter is the algebraic sum of the phase shifts of all sections. The phase shift of the filter exclusive of the delay and loss equalizing section is similar to that of the

usual ladder type low-pass filter. Over the lower frequencies of the transmitting band the phase shift-frequency characteristic is practically linear with frequency, but at the higher frequencies the slope of this curve increases gradually with frequency and becomes very large near the upper edge of the band. Phase shift varying in this manner introduces much more delay distortion than can be tolerated, and therefore has to be corrected. It is one of the functions of the delay and loss equalizing section, which is of the lattice type, to correct for this distortion. The phase shift of this lattice section is such that when it is added to that of the rest of the filter the total phase shift is very nearly proportional to frequency over the whole program band, and delay distortion thus is almost entirely eliminated.

The property of the lattice section by which its phase shift can be made to vary with frequency in the desired manner is expressed in the following characteristic equation, which holds only in the transmitting band and when the section is terminated in its image impedances:[5]

$$\tan \frac{B}{2} = \frac{Kf\left(1 - \frac{f^2}{f_2^2}\right)\sqrt{1 - \frac{f^2}{f_c^2}}}{\left(1 - \frac{f^2}{f_1^2}\right)\left(1 - \frac{f^2}{f_3^2}\right)}. \tag{1}$$

In this equation, B is the phase shift in radians; f is the frequency in cycles per second; f_1, f_2, f_3, and f_c are frequencies at which the phase shift of the section is successive multiples of π radians or 180 deg., f_c being also the cut-off frequency of the filter; and K is a constant controllable by assigning the proper values to the coils and condensers of the section. By assigning to f_1, f_2, and f_3 the values of frequency at which it is desired that the phase shift of the section shall be π, 2π, and 3π radians, respectively, and by giving K the proper value, the phase shift-frequency curve is made to approximate the ideal one which completely would correct the delay distortion of the filter. Figure 3 illustrates the building up of the phase shift characteristic.

The delay corresponding to the rate of change of the phase shift with frequency is plotted in Fig. 4. The average delay introduced by the filter is about 0.00035 sec. It may be noted that for frequencies below 7,500 cycles per second, the variation from this average does not exceed 0.000025 sec. Thus the delay due to 50 filters in a long program circuit does not deviate from the average in this frequency range by more than 0.00125 sec. Distortion of this amount ordinarily would not be detected by the average listener. Above 7,500 cycles

[5] U. S. Patent No. 1,828,454 to H. W. Bode.

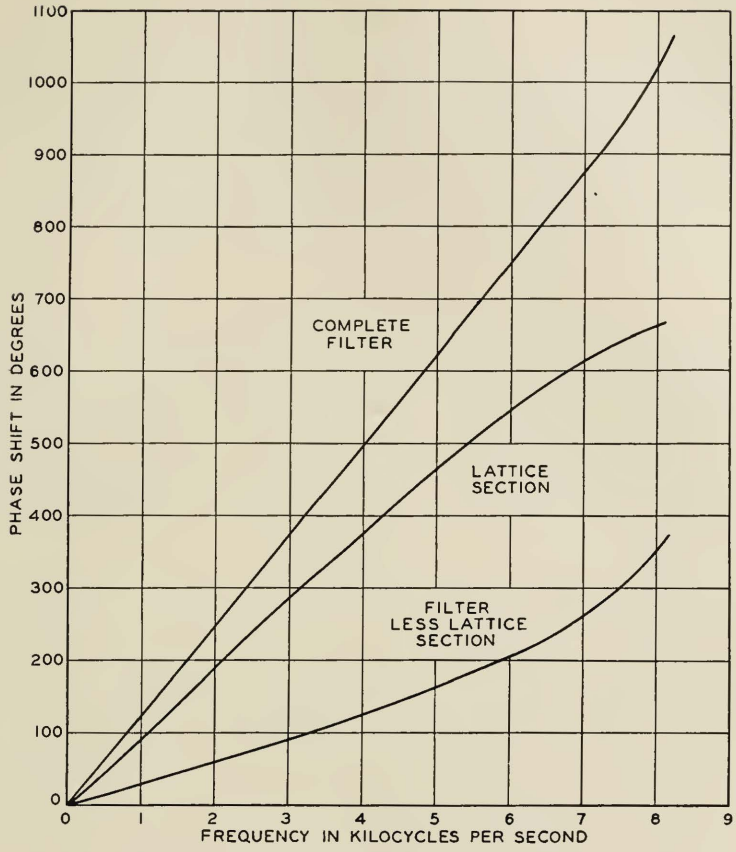


Fig. 3—Phase shift in filter and in component parts.

per second the delay gradually increases with frequency, rising quite rapidly outside the program band. The high attenuation at frequencies above the program range, however, eliminates any effect this distortion otherwise might have on the program.

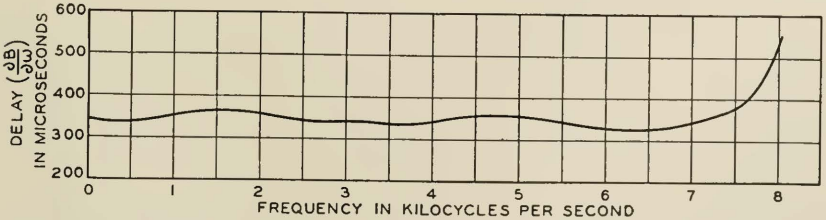


Fig. 4—Delay-frequency characteristic of filter. The ordinates of this curve are proportional to the slope of the upper curve of Fig. 3.

Loss Equalization

Another function of the lattice section is to make the loss of the filter constant in the program frequency band. In a dissipationless filter terminated in its image impedances (which is substantially the condition under which this filter is operated) the loss in the transmitting band is zero. The effect of dissipation is to introduce a loss which is given approximately in this band by the equation:

$$A_d = \frac{\omega}{2Q} \frac{\partial B}{\partial \omega} \tag{2}$$

where A_d is the loss due to dissipation, B is the phase shift of the non-dissipative filter, and Q is the average dissipation factor of the coils (dissipation in the condensers being negligible, ordinarily). The factor Q is equal to the average of the ratios $\omega L_e/R_e$, and $\omega/2Q$ in equation (2) therefore may be written $R_e/2L_e$, where R_e and L_e are the effective resistance and effective inductance, respectively, of the coils.

In the coils of the program filter, Q is about proportional to frequency over the lower portion of the program band, but above this range the factor $\omega/2Q$ increases with frequency. For the filter exclusive of the lattice section, the factor $\partial B/\partial \omega$ is also greatest at the higher frequencies, as may be seen from the lower curve in Fig. 3; hence this part of the filter introduces much more amplitude distortion than is permissible. For the lattice section alone, however, the factor $\partial B/\partial \omega$ is greatest at the lower frequencies, as is apparent from the middle curve of Fig. 3. Thus the natural tendency of dissipation in the lattice section is to compensate for the distortion in the other sections of the filter. This compensating tendency can be controlled to a considerable degree, since by equation (2) A_d is proportional to R_e. By proper adjustment of the effective resistance of the coils of the lattice section, its loss is made practically complementary to that of the rest of the filter, so that the loss of the complete filter is substantially constant throughout the program range.

The loss of the filter in the transmitting frequency band is shown in Fig. 5. The average loss below 7,000 cycles per second is about 0.53 db and the deviation from this average does not exceed 0.03 db. Considering again a circuit containing 50 filters, the deviation from the average loss introduced by the filters does not exceed 1.5 db in this range. Between 7,000 and 7,500 cycles per second the amplitude distortion per filter is about 0.10 db, and above 7,500 cycles the loss increases in such a way as to tend to mask the small delay distortion in this range.

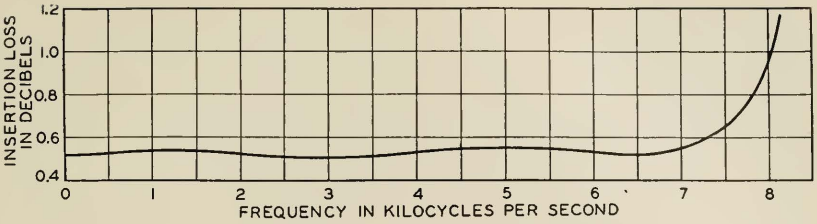


Fig. 5—Loss of filter in program frequency band.

Impedance Correction

In the discussion of the lattice section it was stated that its phase shift is given by equation (1) only when the section is terminated in its image impedance. To facilitate the design and simplify the filter structure, this section has been given an image impedance of the simplest type. This impedance, Z_I, varies with frequency according to the following equation:

$$Z_I = \frac{Z_o}{\sqrt{1 - \frac{f^2}{f_c^2}}}, \tag{3}$$

where Z_o is the "nominal impedance" of the filter, a constant equal approximately to the average impedance of the open-wire lines in the program band; and f_c is the theoretical cut-off frequency. Thus the image impedance rises with increasing frequency to a very high value near the cut-off; and, since the line impedance is practically constant except at very low frequencies, a large mismatch would result at the upper edge of the transmitted band if the lattice section were connected directly to the line. The impedance correcting sections at the ends of the filter are employed to avoid this mismatch. The properties of these sections are such that when they are inserted between the lattice section and the line or the office terminating apparatus, the impedance of the filter matches that of the line and the office apparatus, and the lattice section faces its own image impedance. In this manner, both internal and external reflections largely are avoided; and the phase shift of the lattice section has the proper value.[6]

The general theory on which the design of the impedance correcting sections is based is discussed at length in a recently published article.[7] In brief, the sections consist of two parts: a 4-terminal network to

[6] "Impedance Correction of Wave Filters," by E. B. Payne, *Bell. Sys. Tech. Jour.*, October, 1930.

[7] "A Method of Impedance Correction," by H. W. Bode, *Bell. Sys. Tech. Jour.*, October, 1930.

make the resistance of the filter approximately constant over the program band, and a 2-terminal network placed in shunt at the end to cancel the reactance of the filter in this band. The inductance and capacitance of the coils and condensers of the 4-terminal network are related to the coefficients of a power series expansion of the right-hand part of equation (3) in the manner explained in the article by H. W. Bode.[7] The 2-terminal shunt network at the apparatus end is designed so that, while canceling the reactance of the filter in the program band, it resonates just above the band to produce a peak or sharp maximum of attenuation. It thus supplies the sharp selectivity required to produce an abrupt change from free transmission of the program frequencies to high attenuation of the carrier frequencies.

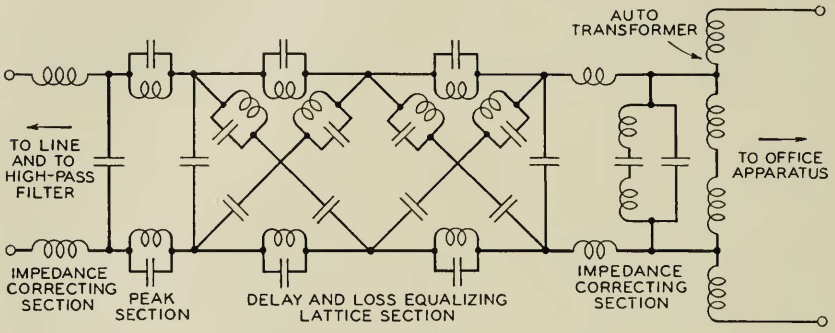


Fig. 6—Schematic diagram of filter.

At the line end, the impedance correcting section is designed for parallel connection with the high-pass line filter. The high-pass filter itself acts as the shunt reactance-canceling network.

The peak section shown at the left of the delay and loss equalizing section in Fig. 1 provides attenuation which rises rapidly with frequency above the program band in such a way as to add to the selectivity of the filter. It is a ladder section of a type often employed in filters for its selectivity.

The filter is designed to match the average impedance of the open-wire lines. The impedance of the office apparatus, however, is slightly higher than that of the lines and the filter. An autotransformer therefore is used at the end of the filter connected to the office apparatus, to effect the required change in impedance. A schematic diagram of the complete filter is shown on Fig. 6, the parts being marked for identification in accordance with the foregoing discussion.

Contemporary Advances in Physics, XXVIII
The Nucleus, Third Part *

By KARL K. DARROW

Transmutation, the major subject of the Second Part of this sequence on the nucleus, assumes again a leading role in the present article. Remarkable cases have been discovered since the first of the year, including a great number in which the impact of one nucleus upon another (or of a neutron on a nucleus) provokes an instantaneous transmutation which is followed after seconds, minutes or hours by the spontaneous breaking-apart of one of the resultant nuclei. One may say that these last are the nuclei of new kinds of radioactive elements, and the phenomena are often called "induced radioactivity"; but many of these new unstable elements differ from all radioactive bodies hitherto known in that they emit *positive* electrons. Some additional examples of transmutation are described at the end of this article.

Induced Radioactivity

UP to the end of last year (1933) it was taken for granted that transmutation is practically instantaneous: that when two nuclei collide, the ensuing fusion and disruption (if any there be) are ended within a time inappreciably short. Nowadays, however, many cases are being discovered, in which a disruption occurs a long time—several minutes or even hours, possibly not for days—after the collision. We must suppose that at the moment of the collision something happens, which entails the eventual disruption. In a very few cases we may be reasonably sure that this initial "something" is itself a transmutation, resembling those previously known in that it is instantaneous, but differing from them in that one of the resulting fragments is an unstable nucleus, of which the eventual spontaneous disruption is that which is observed. This may be the course of events in all cases, but it is also conceivable that in the collision one of the original nuclei may be put into an unstable state without the occurrence of an initial transmutation.

The first-to-be-known of these phenomena was discovered by M. and Mme. Joliot at the very start of 1934, when they exposed samples of aluminium (and boron and magnesium) to the bombardment of the 5.3-MEV alpha-particles from polonium, and after a few minutes of exposure removed them from the bombarding beam and placed them

* In this issue is published the first section of "The Nucleus, Third Part." The paper will be concluded in the October, 1934 issue.

"The Nucleus, First Part" was published in the July, 1933 issue of the *Bell Sys. Tech. Jour.* (**12**, pp. 288–330), and "The Nucleus, Second Part" in the January, 1934 issue (**13**, pp. 102–158).

beside a Geiger-Müller counter.[1] Hundreds of counts per minute disclosed the emergence of fast-flying particles from the samples. The number per minute fell off *exponentially* (Fig. 1) with the lapse of time: a very important feature, for this is the law of radioactivity. The exponential decline implies that the nuclei which were destined to emit these particles were formed at the moments of collisions and existed intact for periods of time—"lifetimes"—not the same for all but distributed in a perfectly random fashion. Such a decline is

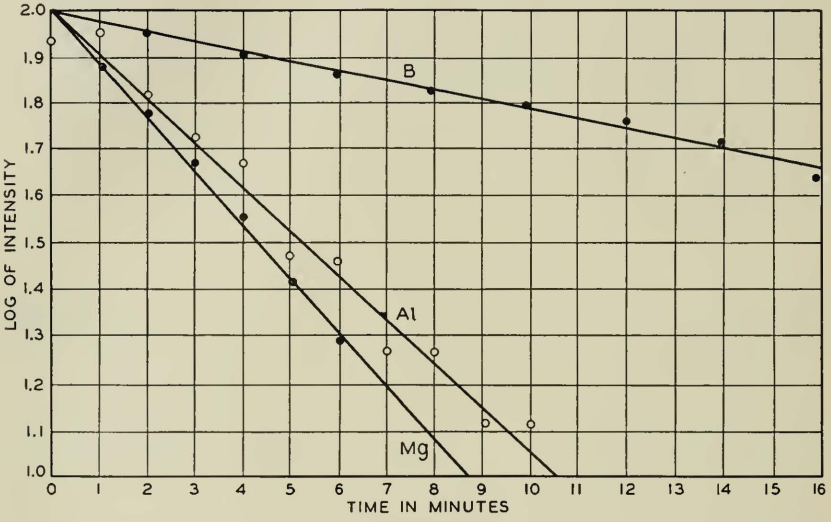


Fig. 1—Exponential decay of the radioactivity induced by boron, aluminium, magnesium with alpha-particles: semi-logarithmic plot. (F. Joliot & I. Curie-Joliot, *Journal de Physique.*)

characterized by a singe constant, the "half-period," or lapse of time during which the rate of emission of particles drops to one-half of its initial value. The half-periods in the three cases examined by the Joliots are different: boron 14′, magnesium 2′ 30″, aluminium 3′ 15″. This is a welcome feature wherever it occurs, as when two substances exhibit different half-periods the effect cannot be ascribed to any contamination common to both.[2]

Since thus there are not only delays between the bombardment and the ultimate disruption, but also (at any rate in the tested cases) a

[1] "The Nucleus, Second Part," p. 119; the "Geiger-Müller" counter has a thin wire for its inner electrode, while most of those called simply "Geiger counters" have needle-points, though the earliest counters invented by Geiger were of the former type.

[2] Cases are on record in which several different elements have exhibited decay-curves, each the sum of two exponentials, one having a half-period characteristic of the element and the other a half-period common to all samples; the latter is then ascribed to a common admixture.

random distribution of the lengths of these delays, it is customary and proper to refer to these phenomena as "induced radioactivity."

Examples of induced radioactivity have already been provoked with all of the four known agents of transmutation: alpha-particles acting on B, Na, Mg, Al and P,—protons acting on boron and carbon—deutons acting on boron and carbon and a number of others—neutrons acting on a large variety of elements. The half-periods reported when neutrons are the agents have ranged from a few seconds to a couple of days, while in all other cases they are of the order of a few minutes.

The nature of the ejected particles resulting from the ultimate disruption is of course of the greatest importance. The Joliots found them to be positive electrons or orestons [3] in their pioneering experiments, and this was confirmed by Ellis and Henderson at the Cavendish Laboratory; the tests have been made by applying magnetic fields to tracks made visible in the Wilson chamber or to beams of particles on their way to photographic or other detectors, and are doubtless to be regarded as conclusive, though no details have yet been published. Induced radioactivity provoked by α-particles, in the few cases so far known, thus results in the emission of orestons.[4] This seems also to be the rule when it is provoked by deutons or protons, as is shown by splendid Wilson-chamber photographs (Figs. 2, 4) obtained by Anderson when samples of various elements (boron in the form of B_2O_3, carbon, aluminium, beryllium) were first bombarded for several minutes and then put right into the chamber itself. The tracks of the particles springing from the samples have the specific aspect of electron-tracks,[5] and in the imposed magnetic field of 800 gauss they have a curvature of which the sense proves the particles to be positive. On the other hand it is stated by Fermi that the radioactivity induced by impacts of neutrons involves the emission of *negative* electrons, though in his very brief reports there is no intimation as to how this is shown.

For each individual case it is important to inquire whether the half-period is independent of such circumstances as the kinetic energy K_0 of the impinging particles. If so, it is sufficient to postulate a single kind of unstable nucleus resulting from the collisions; otherwise, not. This has been investigated in the cases of radioactivity induced by alpha-particle impact; the Joliots reduced K_0 from 5.3 to 1 MEV, without observing any change in the half-perod.

[3] As an occasional alternative to "positive electron" I adopt Dingle's beautiful word "oreston" (Orestes, in Greek mythology, was the brother of Electra).

[4] Excepting that the Joliots have lately reported that magnesium emits electrons of both signs, which they attribute to different isotopes.

[5] "The Nucleus, First Part," p. 303.



Fig. 2—Induced radioactivity resulting from bombardment of carbon by 0.9-MEV protons: tracks of positive electrons. (C. D. Anderson.)

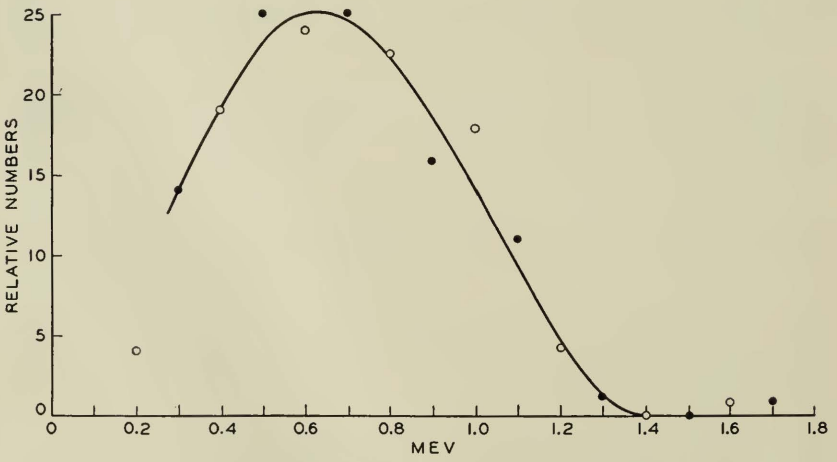


Fig. 3—Distribution-in-energy of positive electrons of the induced radioactivity resulting from bombardment of carbon by 0.9-MEV protons. (Anderson & Neddermeyer, *Physical Review*.)

One next inquires whether all of the orestons resulting from a given type of impact spring off with the same energy. Experience with natural radioactivity shows that while alpha-particles are emitted either with a single definite energy or with one of several definite discrete energies characteristic of the particular process, negative electrons (beta-particles) are always emitted with a very wide and



Fig. 4—Induced radioactivity resulting from bombardment of boron oxide by 0.9-MEV deutons: tracks of positive electrons, some springing from gas adjoining the target, as though a radioactive gas had diffused out of the boron oxide block. (Anderson.)

continuous distribution-in-energy. Short as is the time which has elapsed since January last, and weak as are the beams of positive electrons resulting from induced radioactivity, it is already assured that in several cases at least it is the latter rule which is followed and not the former. The best distribution-curves are those derived at Pasadena from a statistical study of oreston-tracks made visible in a Wilson chamber and curved by an imposed magnetic field; they refer to radioactivity provoked by 0.9-MEV protons falling on carbon, and by 0.9-MEV deutons falling on Be, B, C and Al. I reproduce one of these curves as Fig. 3 (another curve obtained with 0.7-MEV protons

falling on carbon is indistinguishable from it). In one of the cases of radioactivity induced by alpha-particle impact, Ellis and Henderson at the Cavendish Laboratory observed a continuous distribution of energies of the positive electrons ranging between 1 and 2.5 MEV.

In all of these cases of delayed transmutation, nothing is observed of the ultimate disruption excepting the emergence of the electron; the other fragments apparently do not receive energy enough to make a track or reach a detector, and our knowledge is thus forcedly incomplete as it is with most other examples of transmutation. In respect to the initial process occurring at the collision, the prospect of attaining complete knowledge seems even dimmer. We are not without some guidance, for when alpha-particles impinge on aluminium or boron, certain particles are expelled with apparently no delay, and these may be fragments resulting from that initial process. There is, however, an *embarras de choix;* both protons and neutrons are expelled in each of these cases; if one is a fragment resulting from the same process of which an unstable nucleus of half-period 3′ 15″ is another fragment, then the other must be due to something entirely different. Actually Ellis and Henderson inferred from their data that in the case of aluminium, the number of protons produced by a given bombardment is fifty times as great as the number of unstable nuclei which eventually eject orestons. This obliges us to assume that the initial process out of which the delayed transmutation arises is either the one which produces the neutrons, or else some other producing no fast-moving particle at all.

Decision between these alternatives is made from a most notable experiment of the Joliots, sufficient indeed by itself to settle the nature of the initial process. To introduce it in the way in which it suggested itself to them, I make the tentative assumption that the initial process is a case of what is called [6] "disintegration by capture with emission of a neutron," and that the residue of this process is the unstable nucleus. Embodying this assumption in equations of "nuclear chemistry" written after the fashion of those in the Second Part with atomic number for a subscript preceding the symbol of each element (so that ${}_0n$ and ${}_1e$ become the proper symbols for a neutron and an oreston) we have for boron and for aluminium:

$$ {}_2\mathrm{He} + {}_5\mathrm{B} = {}_7\mathrm{N} + {}_0n, \quad \textit{followed by} \quad {}_7\mathrm{N} = {}_6\mathrm{C} + {}_1e, $$
$$ {}_2\mathrm{He} + {}_{13}\mathrm{Al} = {}_{15}\mathrm{P} + {}_0n, \quad \textit{followed by} \quad {}_{15}\mathrm{P} = {}_{14}\mathrm{Si} + {}_1e. $$

The unstable nucleus, if it is surrounded by its proper quota of orbital

[6] "The Nucleus, Second Part," pp. 147–148, 155.

electrons, should then possess the chemical properties of nitrogen in the former case, phosphorus in the latter.

The important experiment of the Joliots consisted in showing that when a sample of boron (or aluminium) is first exposed to alpha-particle bombardment and then to such chemical processes as would remove nitrogen (or phosphorus) commingled with the boron (or aluminium), the induced radioactivity is itself removed and carried away. I quote *verbatim:* "We have irradiated the compound BN. By heating boron nitride with caustic soda, gaseous ammonia is produced. The activity separates from the boron and is carried away with the ammonia. This agrees very well with the hypothesis that the radioactive nucleus is in this case an isotope of nitrogen. When irradiated Al is dissolved in HCl, the activity is carried away with the hydrogen in the gaseous state, and can be collected in a tube. The chemical reaction must be the formation of PH_3 or SiH_4. The precipitation of the activity with zirconium phosphate in acid solution seems to indicate that the radio-element is an isotope of phosphorus."

The assumed equations are thus substantiated in a very striking way. These experiments are in a sense the first chemical identifications of any product of transmutation; I say "in a sense," because while this nitrogen and this phosphorus are identified by virtue of chemical properties, they are detected only by virtue of their radioactivity.[7]

Some striking photographs, taken at Pasadena with an expansion-chamber containing a block of boron oxide previously bombarded by alpha-particles, show many tracks of positive electrons springing from points in the air of the chamber (Fig. 4). It is inferred that the unstable nuclei formed from the boron (not from the oxygen, since bombardment of SiO_2 has no effect) are carbon nuclei which unite with electrons to form carbon atoms and then with oxygen atoms to form molecules of CO or CO_2 having a natural tendency to diffuse out of the solid mass. The radioactivity may be driven completely out of the solid block in short order by heating to 200° C. The radioactive particles are unable to pass through a liquid-air trap.

[7] Inserting mass-numbers into the equations, one finds that since Al has but the one known isotope 27, the value 30 is indicated for the mass-number of "radio-phosphorus," as Joliot calls it; while since boron has two isotopes 10 and 11, the two values 13 and 14 are indicated for radio-nitrogen, with no certain evidence to dictate a choice between them. Ordinary stable phosphorus has no known isotope 30, and ordinary stable nitrogen has no known isotope 13, but the vast majority of its atoms are of mass-number 14. It seems natural that a very unstable isotope should have a different mass-number from any of the known and stable ones, and this may be a valid argument for inferring that it is B^{10} rather than B^{11} which is concerned in the induced radioactivity of boron; but there is nothing to prohibit us from supposing that there may be an unstable isotope of nitrogen agreeing in mass-number with the one which is durable.

The result of bombarding carbon with deutons might be expected to be the same as that of bombarding boron with alpha-particles, it being natural to assume the reactions:

$$_1H^2 + {}_6C^{12} = {}_7N^{13} + {}_0n^1, \quad \textit{followed by} \quad {}_7N^{13} = {}_6C^{13} + {}_1e.$$

The half-period of the delayed disruption has been determined at Pasadena as 10.3 minutes. This does not agree with that observed by the Joliots when alpha-particles are projected against boron. The disagreement is not so welcome as agreement would have been, but does not in the least invalidate the foregoing equations, since it is perfectly conceivable that two different unstable nuclei with different half-periods might both have the atomic number 7 and the mass-number 13. Bombardment of carbon with protons leads to delayed disruptions with the same half-period of about ten minutes, and this is not so easy to understand as it may seem, since the obvious notion that the proton and the C^{12} nucleus simply merge into a nucleus N^{13} which later on explodes leads into difficulties with the principles of conservation of energy and conservation of momentum.

As to the way in which the number of observed disruptions varies with the kinetic energy K_0 of the impinging particles, there are data relating to the bombardment of aluminium by alpha-particles. The Joliots varied K_0 from 5.3 MEV downwards; they report that the number of positive electrons diminishes with falling K_0, becoming imperceptible for boron at about 3 MEV, for Mg and Al at 4 to 4.5 MEV. Ellis and Henderson varied K_0 from 5.5 upward to 8.3 MEV, by using alpha-particles emitted from other radioactive bodies than polonium; they found the number of orestons steadily increasing with rising K_0, rising in the ratio 15 : 1 as K_0 was raised from 5.5 to 7 MEV, and showing signs of approaching a maximum not far beyond K_0 = 8.3 MEV.

The positive electrons emitted in induced radioactivity are frequently—perhaps generally—accompanied by high-frequency photons, of which energy-measurements may hereafter show that they are due to the coalescence of positive with negative electrons to form light.

I close this section by listing the elements which have been observed to display induced radioactivity after bombardment by one or other of the four agents of transmutation, and add those which have been tested without positive results, in order to show the scope of the experiments. In certain cases positive results have been obtained by some observers and not by others, but this may signify simply a weaker bombarding stream or a less sensitive detector in the apparatus of the latter.

Bombardment by alpha-particles: B, Mg, Al (Joliots, Ellis & Henderson); Na, P (Frisch); negative results with H, Li, Be, C, N, O, F, Na, Ca, Ni, Ag (Joliots).

Bombardment by deutons: Li, Be, B, N, C, O, F, Na, Mg, Al, Si, P, Cl, Ca (Henderson, Livingston & Lawrence, with 3-MEV deutons); Li, Be, B, C, Mg, Al (Crane & Lauritsen, with 0.9-MEV deutons).

Bombardment by protons: B, C (Crane & Lauritsen); C (Cockcroft, Gilbert & Walton with 0.6-MEV protons); C (?) (Henderson *et al.*, with 1.5-MEV protons). Negative results by Henderson *et al.* with 1.5-MEV protons on all but C among the elements listed above before their names.

Bombardment by neutrons: F, Na, Mg, Al, Si, P, Cl, Ti, V, Cr, Fe, Cu, Zn, As, Se, Br, Zr, Ag, Sb, Te, I, Ba, La, U (Fermi); F, Mg, Al (Dunning and Pegram).

Other Cases of Transmutation

It is not altogether safe to separate cases of "induced radioactivity" from "other cases of transmutation," inasmuch as most of the latter class have been observed under conditions where it was impossible to tell whether or not there was a delay between collision and disruption, and perhaps some of them belong in the former class. Of certain transmutations one may say that if there is such a delay, the law of conservation of momentum must be suspended for the duration thereof, resuming its sway only at the moment of the disruption. Nevertheless I should not wish to affirm that for the processes mentioned in this section or in the Second Part the delay is always literally zero.

Early in this year was first achieved, at the Cavendish Laboratory by Oliphant, Shire and Crowther, what had been the aim of many physicists for over a decade: the separation of a metal, normally consisting of more than a single isotope, into films each comprising atoms of practically a single isotope only, and thick enough for physical experiments. This was performed with lithium, and when protons and alternatively deutons were projected against films of Li^6 and alternatively Li^7, the four resulting sets of observations settled the attributions of the various groups of fragments previously observed when ordinary blocks of lithium had been bombarded. The origin of the two long-range groups of paired alpha-particles described in the Second Part was precisely as had been suspected: they proceed from the interactions:

$$ {}_1H^1 + {}_3Li^7 = 2\,{}_2He^4 + (T_1 - T_0); \quad {}_1H^2 + {}_3Li^6 = 2\,{}_2He^4 + (T_1 - T_0), $$

where $(T_1 - T_0)$ stands for the amount of energy transformed in each

reaction from energy-of-rest-mass to kinetic energy, equal to about 17 MEV in the first case and to about 23 MEV in the second. The continuous distribution of alpha-particles up to range 7.8 cm (Fig. 9 in the Second Part) is due to impacts of deutons against Li^7, and thus may still be attributed to a transmutation in which three nuclei,—a neutron and two alpha-particles—spring from the merger of a deuton with a Li^7 nucleus. Of the other attributions I shall presently speak.

The transmutations arising from the impact of deutons on deuterium are in some ways unique. They are the first to be known in which the two colliding particles are identical, both being H^2 nuclei; one of them appears to be much the most abundant yet observed, in the sense that a given number of bombarding particles produces an unprecedentedly great number of detectable fragments; each of them results in the formation of a nucleus long sought but never certainly detected till 1934.

The better-known of these reactions is described by the equation,

$$ {}_1H^2 + {}_1H^2 = {}_1H^3 + {}_1H^1 + (T_1 - T_0). \qquad (1) $$

It is both somewhat amusing and somewhat annoying to realize that this is not a transmutation at all in the formerly-proper sense of the word, since there is no change of one element into another! the hydrogen isotope of mass-number 2 is changed into hydrogen isotopes of mass-numbers 1 and 3 respectively; it will be desirable to enlarge the scope of the term "transmutation" to cover cases like this one. The H^1 nuclei resulting from this reaction were vividly demonstrated by Tuve and Hafstad when they projected deutons into divers gases in an ionization-chamber—air, carbon dioxide, ordinary hydrogen, and deuterium successively; there were no emerging protons (of range superior to 3.5 cm, the minimum observable) from any of the three first named, but from the last there was the "very large yield" of one proton per several thousand impinging deutons. Another estimate of yield has been supplied from the Cavendish school, by Oliphant Harteck and Rutherford; theirs refers to impacts by deutons of energy 0.1 MEV, a value considerably smaller than those of Tuve's research; they find that the number of protons coming forth from a thick layer of deuterium is of the order of a millionth of the number of such deutons entering the layer. The estimates do not seem incompatible, especially as the Cambridge people find the number of fragments to be mounting very rapidly as the deuton-energy T_0 increases;[8] and they show that any possibility of a slight admixture

[8] The "thick layers" are films of certain compounds of hydrogen in which a large proportion of the usual H^1 atoms have been replaced by H^2 atoms. The curve

of deuterium with any other substance must be very carefully considered and assessed, whenever that other substance is bombarded with a beam containing deutons and it is observed that protons are produced.

The range of the protons due to the foregoing reaction is about 14 cm when T_0 is low—0.1 MEV or thereabouts—and rises with T_0. Translate its minimum value into the corresponding kinetic energy (obtaining about 3 MEV); compute the momentum of the proton—this, save for a minor correction due to the relatively small momentum of the impinging deuton, should be opposite in direction and equal in magnitude to the momentum of the other fragment of the transmutation, the nucleus H^3. Thence compute the kinetic energy of this other fragment, and estimate thence its presumable range; owing to our lack of experience with such particles the estimate may not be very exact; Oliphant, Harteck and Rutherford arrive at the figure 1.74 cm. Now, the protons of 14-cm range of which I have been speaking are not the only fragments to be observed when deutons impinge on deuterium. There are also particles of a much less range; these are equally numerous with the 14-cm protons, and expansion-chamber photographs by Dee have shown that a track of the one variety is likely to be paired with a track of the other, after the fashion of the paired tracks due to the transmutations H + Li = 2He (Figs. 14 and 15, Second Part); and their range of about 1.6 cm. is taken by the Cavendish people as being in substantial agreement with the estimate aforesaid. It is this interlocking of concordant observations which speaks so strongly for the rightness of this description of the reaction, and therefore for the existence of the hitherto-unknown isotope H^3 of hydrogen.

Meanwhile it has been discovered at Princeton that the new isotope can be generated by maintaining a self-sustaining discharge in gaseous deuterium: a way of achieving transmutation several times attempted in past years, but never (so far as I know) with proved success. Out from the discharge tube (where the voltage is 50,000 to 80,000) some of the ionized atoms and molecules shoot through a hole in the cathode into another and very large chamber filled with deuterium in which they disperse themselves, thus having opportunities for transmutation in both this chamber and the tube. A sample of the gas is afterwards

of number-of-fragments *vs.* T_0 shows the peculiar shape common to such curves when obtained with thick layers, which suggests that as T_0 is raised the increase in the number of transmutations is at first partly due to an increase in the probability of transmutation at an impact, but later entirely due to the fact that the faster particles enter farther into the layer and have more opportunities of striking nuclei before their energy is gone than do the slower (The Nucleus, Second Part, p. 141). The theory of such curves has, however, never been worked out.

extracted and is ionized in a separate chamber; the charge-to-mass ratios of its ions are determined by an especial type of deflection-apparatus. Search is made for ions having the charge-to-mass ratio of a singly-ionized molecule of mass about 5, such as could be a molecule H^2H^3. Such ions occur. To discover them, however, is not the same thing as to prove the existence of H^3, since so far as anyone can tell from their charge-to-mass ratios (as measured with the accuracy attainable in these experiments) these ions might have the constitution $H^2H^2H^1$—there being some of the isotope H^1 in the gas. How to make such discriminations is one of the major problems in the analysis of the ions found in gases. In this case it happens to be known that in ordinary hydrogen, the ratio of the number of triatomic to that of diatomic molecular ions is proportional to the density of the gas. Now in these experiments, the ratio of the number of mass-5 ions to the number of mass-4 ions is the sum of two terms, one proportional to the gas-density and the other independent of it. The latter term is taken as the measure of the amount of H^2H^3, therefore of H^3, in the gas. A like study made with deuterium none of which had been exposed to the discharge indicated a very small amount of H^3, about one atom in two hundred thousand of H^2; the discharge enhanced this ratio fortyfold in an hour.

To return to the work at the Cavendish Laboratory: the lesser-known of the two reactions which may occur when deutons meet is probably described by the equation,

$$ {}_1H^2 + {}_1H^2 = {}_2He^3 + {}_0n^1 + (T_1 - T_0) \qquad (2) $$

and is a transmutation in the strictest sense of the word, helium as well as neutrons [9] appearing out of hydrogen. I refer to it as lesser-known, because although the neutrons have been observed the helium nuclei have not been. This lack of evidence withholds a desirable support from the equation, but does not contradict it; for on measuring the momentum of the neutron, equating it to that of the hypothetical He^3 nucleus and estimating the range of the latter, this range turns out to be so small as to make detection difficult. We are not, however, without other evidence for He^3; when protons are projected against lithium, particles of ranges 1.15 cm. and 0.68 cm. appear,[10] and the observations made with monisotopic films show that Li^6 is involved in their origin: if we suppose

$$ {}_1H^1 + {}_3Li^6 = {}_2He^4 + {}_2He^3 + (T_1 - T_0) \qquad (3) $$

[9] Harkins has suggested the name "neuton" for the element of which neutrons are the ultimate particles.

[10] Kirchner has lately observed an 0.9-cm. group.

the equation is supported by the facts that the ranges of the two groups stand to one another in the ratio computed by assuming equality of momenta, that particles of one are found to be paired with particles of the other, and that they ionize about as much as alpha-particles of equal range.

The rest-masses of the two new nuclei are estimated by putting, in equations (1), (2) and (3), the best available values for T_0 (the kinetic energy of the impinging deuton, that of the other H^2 nucleus being negligible) and T_1 (the sum of the kinetic energies of both fragments resulting from the reaction). The results are: for the rest-mass of H^3, 3.0151 from (1); for the rest-mass of He^3, 3.0166 from (3). To derive the latter from (2) is not so precise, the energy of neutrons being harder to evaluate than that of charge-bearing particles; Oliphant, Harteck and Rutherford prefer to say merely that the result is not incompatible with that from (3).[11]

These are the fourth and fifth of the nuclei (counting the neutron as one) in order of increasing mass. The departures of their masses from the adjacent integer are abnormally great for light nuclei, and their packing-fractions (First Part, p. 318) are the greatest yet known excepting that for H^2, and fall neatly by the upper branch of the curve of packing-fraction *vs.* mass-number (Fig. 8 of the First Part). The contrast between the packing-fractions 55 of He^3 and 5 of He^4 is especially striking. The new nuclei are the first isobars to be discovered of mass-number less than 40, and the first pair to be discovered of which the masses are distinguishable.

Cockcroft and Walton have studied at length the fragments emerging from lithium, boron and carbon bombarded by deutons. Lithium supplies a group and boron a group of protons which may result from the transformation of the lighter into the heavier isotope according to the schemes,

$$ {}_1H^2 + {}_3Li^6 = {}_3Li^7 + {}_1H^1 + (T_1 - T_0), \qquad (4) $$

$$ {}_1H^2 + {}_5B^{10} = {}_5B^{11} + {}_1H^1 + (T_1 - T_0), \qquad (5) $$

but the two members of each equation (in which all the rest-masses are known by deflection-experiments) do not agree very well. Carbon supplies a group and boron two more groups of protons which cannot be made to fit into such a scheme without postulating emission of gamma-rays to achieve the balancing of masses—an emission for which,

[11] These results are computed by assuming that the values of the rest-masses of H^1, H^2, He^4, Li^6, Li^7 and n^1 given by Aston, Bainbridge and Chadwick are exact, and that no additional fragment (such as a gamma-ray photon) of appreciable energy is emitted at the transmutation.

it is true, independent evidence exists in the case of carbon. Boron supplies a group of alpha-particles which may be due to the reaction,

$$_1H^2 + {}_5B^{10} = 3{}_2He^4 \qquad (6)$$

and which comprises the most energetic subatomic particles yet known, those of the cosmic rays excepted (12.3 MEV, range 15 cm.). Blocks of various heavier elements emit both alpha-particles and protons, of which the amounts both relative and absolute vary tremendously with heat-treatment, degassing, and other circumstances, so that evidently they cannot altogether proceed from the element constituting most of the block, and their origins furnish a severe problem for research.

Electrical Wave Filters Employing Quartz Crystals as Elements

By W. P. MASON

This paper discusses the use of piezo-electric crystals as elements in wave filters and shows that very sharp selectivities can be obtained by employing such elements. It is shown that by employing crystals and condensers only, very narrow band filters result. By using coils and transformers in conjuction with crystals and condensers, wide-band-pass and high and low-pass filters can be constructed having very sharp selectivities. The circuit configurations employed are such that the coil dissipation has only the effect of adding a constant loss to the filter characteristic, this loss being independent of the frequency. Experimental curves are given showing the degree of selection possible.

In the appendix, a study is made of the modes of motion of a perpendicularly cut crystal, and it is shown that all the resonances measured can be derived from the elastic constants and the density of the crystal. The effect of one mode of motion on another mode is shown to be governed by the mutual elastic compliances of the crystal. By rotating the angle of cut of the crystal, it is shown that some of the compliances can be made to disappear and a crystal is obtained having practically a single resonant frequency over a wide range of frequencies. Such a crystal is very advantageous for filter purposes.

Introduction

FILTERS for communication systems must pass, without appreciable amplitude distortion, waves with frequencies between certain limits, and must attenuate adequately all waves with somewhat greater or smaller frequencies. To do this efficiently, the change from the filter loss in the transmission region, to that in the attenuation region, must occur in a frequency band which is narrow compared to the useful transmission band. At low frequencies, ordinary electrical coil and condenser filters can perform this separation of frequencies well because the percentage band widths (ratio of band width to the mean frequency of the band) and the percentage separation ranges (ratio of the frequency range required, in order that the filter shall change from its pass region to its attenuated region, to the adjacent limiting frequency of the pass band) are relatively large.

For higher frequency systems, such as radio systems, or high frequency carrier current systems, the band widths remain essentially the same, and hence the percentage band widths become much smaller. Here separation by coil and condenser filters becomes wasteful of frequency space. For these filters, owing to the relatively low reactance-resistance ratio in coils (this ratio is often designated by the letter Q) the insertion loss cannot be made to increase faster than a

certain percentage rate with frequency. Hence an abrupt frequency discrimination cannot be obtained between the passed frequency range and the attenuated frequency range. In present radio systems, double or triple demodulation is often used to supplement the selectivity of filter circuits.

If, however, elements are employed which have large reactance-resistance ratios, filters can be constructed which have small percentage bands and which attenuate in small percentage separation ranges. Such high Q elements are generally obtainable only in mechanically vibrating systems. Of these elements, probably the most easily used is the piezo-electric crystal, for it possesses a natural driving mechanism.

It is the purpose of this paper to describe work which has been done in utilizing these crystals as elements in filters.[1] Since the quartz crystal appears to be the most advantageous piezo-electric crystal, all of the work described in this paper is an application of this type of crystal. The possibilities and limitations are discussed and experimental data are given showing that these filters are realizable in a practical form.

Piezo-electric Crystals and Their Equivalent Electrical Circuits

When an electric force is applied to two plates adjacent to a piezo-electric crystal, a mechanical force is exerted along certain directions which deforms the crystal from its original shape. On the other hand deformations in certain directions in the crystal produce a charge on the electric plates. Hence the crystal is a system in which a mechanical electrical coupling exists between the mechanical and electrical parts of the system.

Quartz crystals, particularly when vibrating along their smallest dimension, as they do for high frequency oscillators, have a large number of resonances which do not differ much in frequency from the principal resonance. While this is no great disadvantage for an oscillator, since an oscillator can pick out the strongest resonance and utilize it only, the large number of resonances is a great disadvantage when using

[1] The development of ideas in the direction of employing crystals as elements of selecting circuits dates back to Cady who in patent—Re. 17,358 issued July 2, 1929, original filed January 28, 1920—showed various types of tuned circuits of which crystals formed a part. Subsequently Espenschied in patent 1,795,204, issued March 3, 1931, filed January 3, 1927—patented broadly the use of crystals as elements of true filter structures. More recently a patent of the writer's—1,921,035 issued August 8, 1933, filed Sept. 30, 1931—describes the use of crystals in lattice structures, and this patent, together with several others pending, forms the basis for the filters discussed in this paper. It is only within the last few years that filter structures including crystal elements have been practically realized.

the crystal to select currents over a wide band of frequencies and to reject currents whose frequencies lie outside this pass region. Hence it is advantageous for filter uses to obtain a crystal which has substantially a single prominent resonance over a wide range of frequencies. Such a vibrating element can usually be obtained only by making the dimension along which the crystal is vibrating, large compared to the other dimensions, and this fact determines the best cut of crystal to use.

Two principal types of orientations have usually been employed in cutting quartz crystals. The first type is the so-called Curie or perpendicular cut in which the crystal is so cut that its major surfaces are parallel to the optical axis and perpendicular to an electrical axis. Such a crystal is shown by Fig. 1. The second type of cut is the parallel

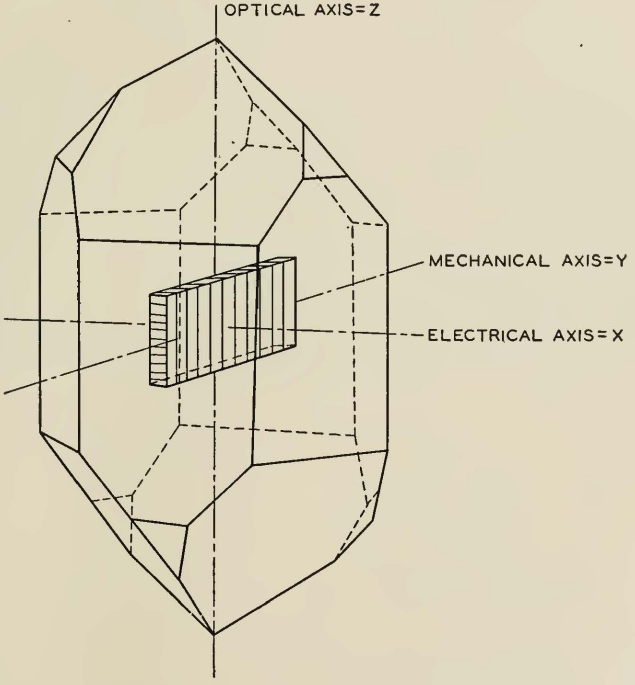


Fig. 1—Orientation of a Curie or perpendicular cut with respect to native crystal.

or 30-degree cut in which the major surfaces of the crystal plate are parallel to both the optical and electrical axes. Since this cut results in a crystal vibrating along its smallest dimension, it is not of much interest for filter uses.

When using a crystal as part of an electrical system, it is desirable

to know the constants of an electrical circuit which has the same impedance characteristic as the crystal. If attention is confined to the single prominent resonance, the electrical circuit representing the crystal is as shown by Fig. 2. Some theoretical consideration has been

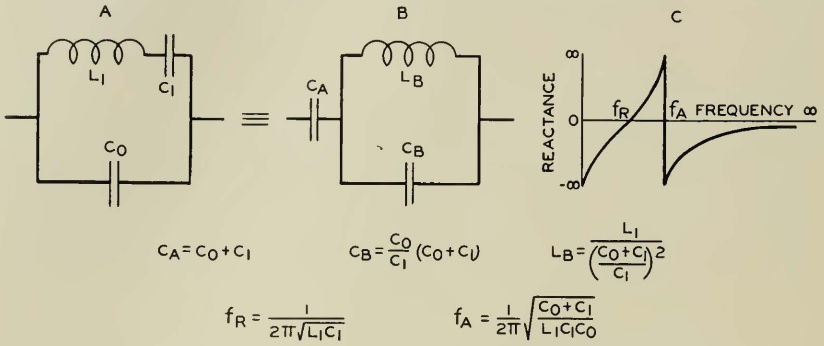


Fig. 2—Equivalent electrical circuit of piezo-electric crystal.

given to the electrical network representing perpendicularly cut crystals by Cady,[2] Van Dyke,[3] Dye,[4] Vigoureux [5] and others. Assuming a quartz plate to have only plane wave motion, Vigoureux has investigated the motion in such a plate, and has shown that there will be resonances at odd harmonics of a particular frequency determined by the length and mechanical constants of the plate. In the neighborhood of the fundamental resonance of the crystal, with the electrical plates placed on the crystal surfaces, he finds the equivalent circuit shown by Fig. 2A, the elements of which in practical units have the following values:

$$C_0 = \frac{l_0 l_m K}{4\pi l_e \times 9 \times 10^{11}} = \text{capacitance in farads},$$

$$C_1 = \frac{l_0 l_m 8 E d_{12}^2}{\pi^2 l_e \times 9 \times 10^{11}} = \text{capacitance in farads}, \qquad (1)$$

$$L_1 = \frac{l_e l_m \rho \times 9 \times 10^{11}}{8 l_0 E^2 d_{12}^2} = \text{inductance in henries},$$

where l_0, l_m, l_e are respectively the lengths of the optical, mechanical, and electrical axes in centimeters,

K = specific inductive capacitance = 4.55 for quartz,
E = Young's modulus = 7.85×10^{11} for quartz,
d_{12} = piezo-electric constant = 6.4×10^{-8} for quartz,
ρ = density = 2.654 for quartz.

[2] W. G. Cady: *Phys. Rev.* XIX, p. 1 (1922); *Proc. I. R. E.* X, p. 83, (1922).
[3] K. S. Van Dyke: Abstract 52, *Phys. Rev.*, June, 1925; Proc. I. R. E., June, 1928.
[4] D. W. Dye: *Proc. Phys. Soc.*, XXXVIII (5), pp. 399–453.
[5] P. Vigoureux: *Phil. Mag.*, Dec., 1928, pp. 1140–53.

Inserting these values, the element values in terms of the dimensions become

$$\begin{aligned} C_0 &= 0.402 \times 10^{-12}\, l_m l_0/l_e, \\ C_1 &= 0.289 \times 10^{-14}\, l_m l_0/l_e, \\ L_1 &= 118.2\, l_m l_e/l_0. \end{aligned} \tag{2}$$

From these values it is seen that there is a fixed ratio between these two capacitances [6] of

$$r = C_0/C_1 = 140. \tag{3}$$

As will be evident later, this ratio limits the possibilities of the use of quartz crystals in filter circuits.

Experiments with quartz crystals, with electrodes contiguous to the crystal surfaces and with the optical and electrical axes small compared to the mechanical axis, show that these values are approximately correct. The value of C_0 checks the above theoretical value quite closely. The value of C_1 obtained by experiment is somewhat larger than that given by equation (2) and the value of the inductance somewhat smaller. The ratio of C_0/C_1 has been found as low as 115 to 1, but a value of 125 to 1 is about all that can be realized, when account is taken of the distributed capacitance of the holder, connecting wires, etc.

When either of the dimensions along the electrical or optical axes becomes more than a small fraction of the dimension of the mechanical axis, the plane wave equations given above no longer hold accurately. This is due to the fact that a coupling exists between the motion along the mechanical axis and other modes of motion. For an isotropic body, one is familiar with the fact that when a bar is compressed or stretched it tends to stretch or compress in directions perpendicular to the principal direction of motion. This state of affairs may be described by saying that the modes of motion are coupled. In a crystal this same relation exists and in addition, due to its crystalline form, a shearing motion is set up whose shearing plane is determined by the mechanical and optical axes and whose motion is parallel to the mechanical axis. In fact the shearing motion is more closely coupled to the mechanical axis motion that is the extensional motion. As long as the optical axis is small compared to the mechanical axis, this coupling action manifests itself as a decreased stiffness along the mechanical axis, but if a condition of resonance is approached for the motion along the optical axis, the mode of motion is entirely changed. This effect is

[6] In a paper contributed recently to the *Institute of Radio Engineers*, it is shown that this ratio limitation is a consequence of a fixed electro-mechanical coupling between the electrical and mechanical systems of the crystal.

analyzed in the appendix and is quantitatively explained in terms of the elastic constants of the crystal. On the basis of this explanation, an investigation is also given in the appendix, of crystals cut at different orientations, and a crystal having many advantages for filter uses is derived.

Some experimental data [7] have been taken for perpendicularly cut crystals for various ratios of axes. Figure 3 shows the principal reso-

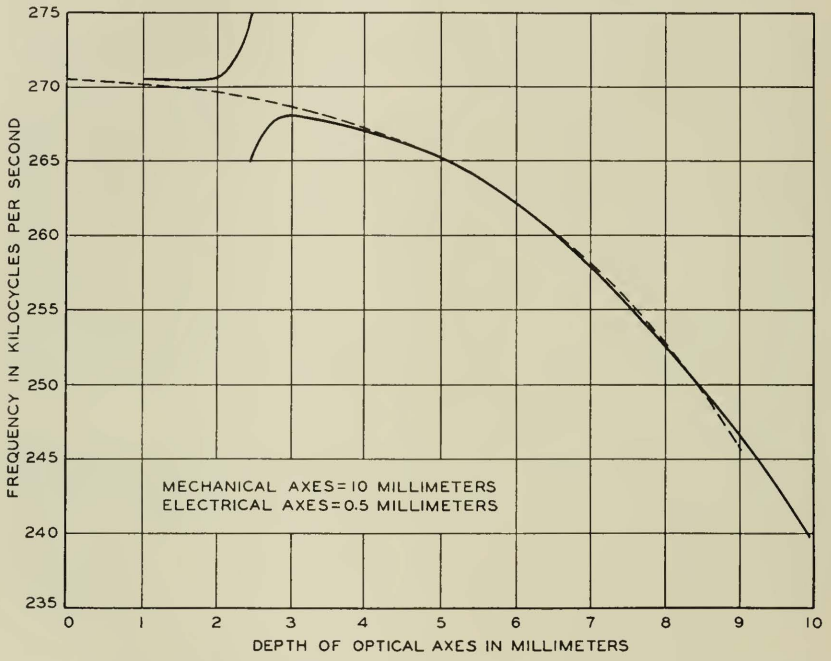


Fig. 3—Principal resonant frequency of a perpendicularly cut crystal as a function of the width of the crystal.

nant frequency (the frequency for which the electrical impedance is a minimum) for a series of crystals whose mechanical axes are all 10 millimeters, whose electrical axes are 0.5 millimeter, and whose optical axes vary from 1 to 10 millimeters. As will be observed, increasing the length of the optical axis in general lowers the resonant frequency. The discontinuity in the curve for the ratio $l_0/l_m = .23$ is discussed in detail in the appendix.

[7] The experimental data shown by Figs. 3 and 4 have been taken by Mr. C. A. Bieling while the temperature coefficient curve of Fig. 5 was measured by Mr. S. C. Hight.

The solid curve of Fig. 4 shows a measurement of the ratio, r, of the capacitances in the simple representation of the crystal shown by Fig. 2A. This ratio is measured by determining the resonant and anti-

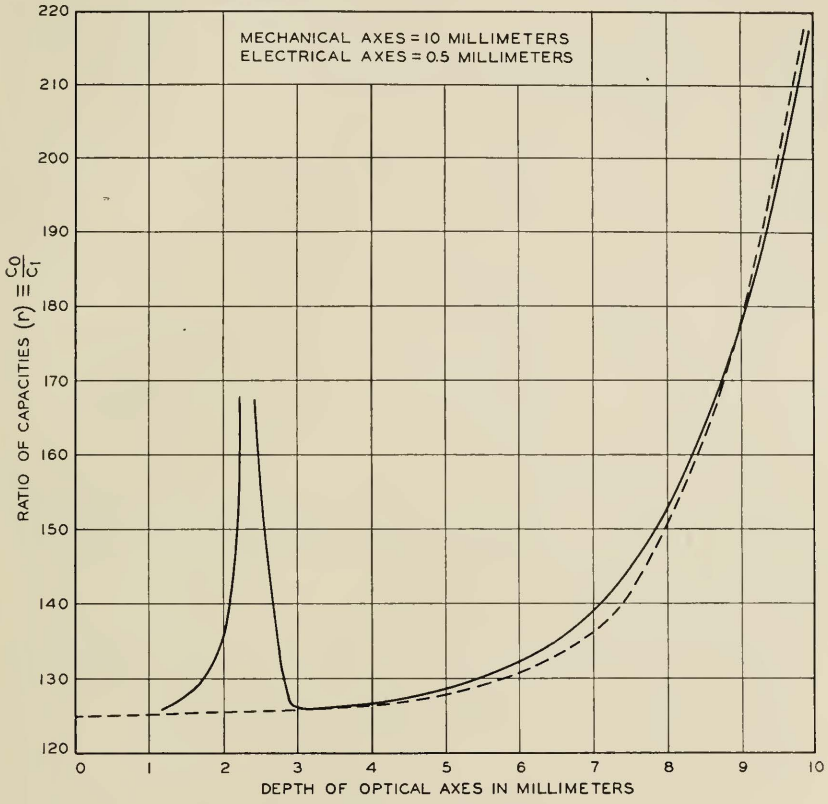


Fig. 4—Ratio of capacitances of a perpendicularly cut crystal.

resonant frequencies of the crystal. r is related to these by the formula

$$f_A^2/f_R^2 = 1 + 1/r, \tag{4}$$

where f_A is the anti-resonant frequency and f_R the resonant frequency.

Figure 5 shows a measurement of the temperature coefficient of the resonant frequency for the same set of crystals. It will be noted that as the optical axis increases in depth, the temperature coefficient increases and that crystals with smaller dimensions along the optical axis in general have much smaller coefficients. Increasing the thickness along the electrical axis has the effect of decreasing the tempera-

ture coefficient and in fact for certain ratios of the three axes the coefficient approaches zero.

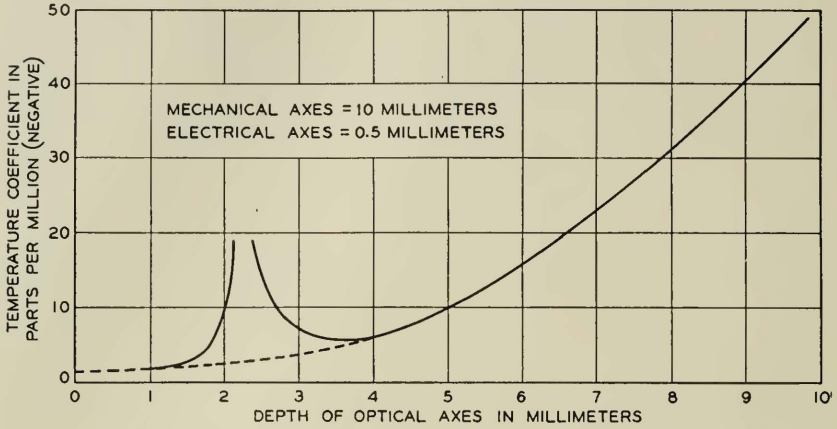


Fig. 5—Temperature coefficient of a perpendicularly cut crystal.

When the crystals are used in filters, two quantities are usually specified, the resonant frequency of the crystal and the capacitance of the series condenser. The shunt capacitance of the crystal is usually incorporated with an electrical capacitance which is specified by other considerations. The resonant frequency is determined principally by the mechanical axis length. The capacitance of the series condenser is determined by the ratio of the area to the thickness or by $l_0 l_m / l_e$. The third condition is given by the fact that the length of the optical axis should be kept as small as possible in order that any subsidiary resonances shall be as far from the principle resonance as possible. The curves of Figs. 3 and 4 and the fact that the resonant frequency of a given shaped crystal varies inversely as the length, can be used to determine the dimensions of the crystal. It is obvious that the crystal should not be used in the region $0.2 < l_0/l_m < 0.3$ on account of the two prominent resonant frequencies.

Use of Crystals and Condensers as Filter Elements

Considering crystals as representable by the simple electrical circuit shown on Fig. 2*A*, these circuits can be utilized as elements in electrical networks. They may, of course, be used in a network employing any kinds of electrical elements. Since, however, their Q is high, it would be advantageous not to employ any electrical elements which do not also have a high Q, in order that the dissipation introduced by these elements may not be a matter of consideration. The

Q's of the best electrical condensers may be as high as 10,000, which is of the same order as the crystal Q, and hence such elements can be employed advantageously with crystals. It is the purpose of this section to discuss the possibilities and limitations of filter sections employing crystals and condensers only.

The simplest types of filter sections are the ladder type networks illustrated by Fig. 6. If crystals and condensers only are employed in

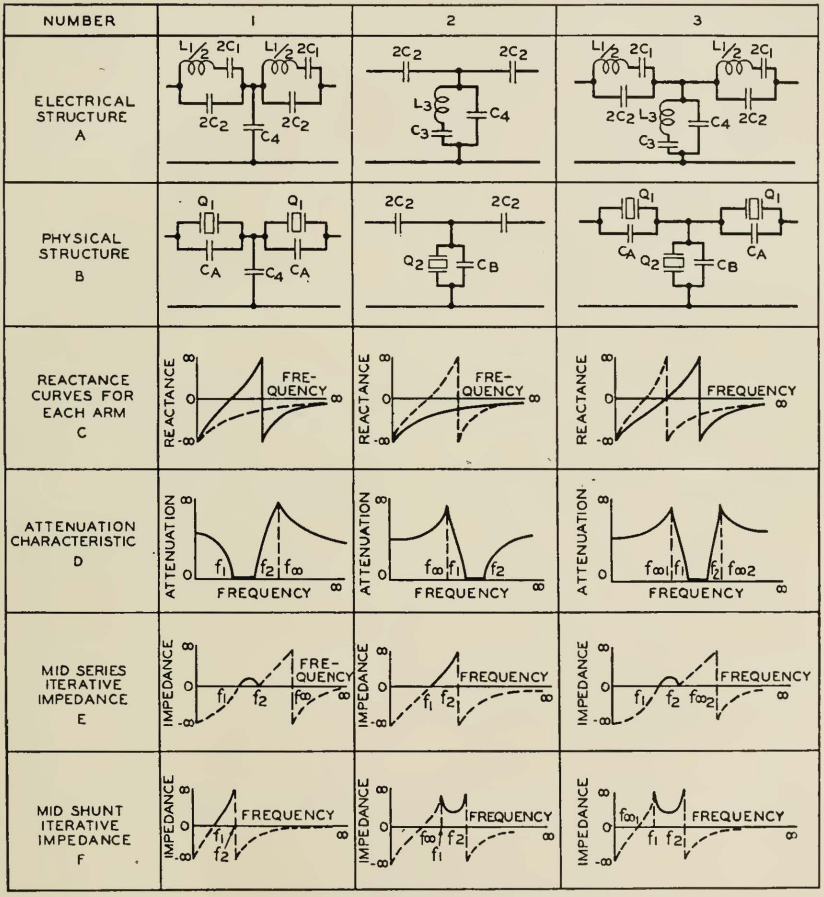


Fig. 6—Ladder networks employing crystals and condensers.

this type network, there are only three types of single band sections possible, all being band-pass filters. Figure 6 shows these sections, the impedance characteristic of each arm, the attenuation characteristics of these networks considered as filters, and their iterative impedances.

These attenuation characteristics and their limitations are at once found from a consideration of the impedance frequency curves for each arm shown by Fig. 6*C*. For a ladder type network it is well known [8] that a pass band will exist when

$$0 \leqq \frac{Z_1}{4Z_2} \leqq -1, \tag{5}$$

where Z_1 is the impedance of the series arm and Z_2 the impedance of the shunt arm. Hence, considering the first filter of Fig. 6, it is obvious that the lower cut-off fc_1 will come at the resonant frequency of the crystal. The upper cut-off fc_2 will come some place between the resonant and anti-resonant frequency, the exact position depending on the amount of capacitance in shunt. The anti-resonant frequency will be a point of infinite attenuation since the filter will have an infinite series impedance at this frequency.

With the restriction on the ratio of capacitances of the crystal noted in the previous section, it is easily shown that the ratio of the anti-resonant frequency to the resonant frequency is fixed and is about 1.004. Hence, we see that the ratio of $f\infty$ to fc_1 can be at most 0.4 per cent. The band width must be less than this since fc_2 must come between $f\infty$ and fc_1. A similar limitation occurs for the second filter of this figure, for which case the separation of $f\infty$ and fc_2 is at most 0.4 per cent. For filter number 3, a somewhat larger frequency separation between the points of infinite attenuation results, it being at most 0.8 per cent. The addition of any electrical capacitance in series or shunt with any of the crystals results in a narrowing of the band width.

It is seen then that there are two limitations in the types of filters obtainable with crystals and condensers in ladder sections. One, there is a limitation on the position of the peak frequencies and two, there is a limitation for the band width of the filters.

By employing the more general lattice type of filter section shown on Fig. 7, the first of these limitations can be removed. By means of this type of section it is possible to locate the attenuation peak frequencies at any position with respect to the pass band, but the pass band is limited in width to at most 0.8 per cent.

For a lattice filter a pass band exists when the impedances of the two arms are related by the expression [9]

$$0 \leqq \frac{Z_1}{Z_2} \leqq -\infty, \tag{6}$$

[8] See, for example, page 190 in book by K. S. Johnson, "Transmission Circuits for Telephonic Communications."

[9] "Physical Theory of the Electric Wave Filter," G. A. Campbell, *B. S. T. J.*, November, 1922.

where Z_1 is the impedance of the series arm (either 1, 2 or 3, 4 of Fig. 7*A*) and Z_2 the impedance of the lattice arm (either 1, 3 or 2, 4 of Fig. 7*A*). Hence, if one pair of branches has a reactance whose sign

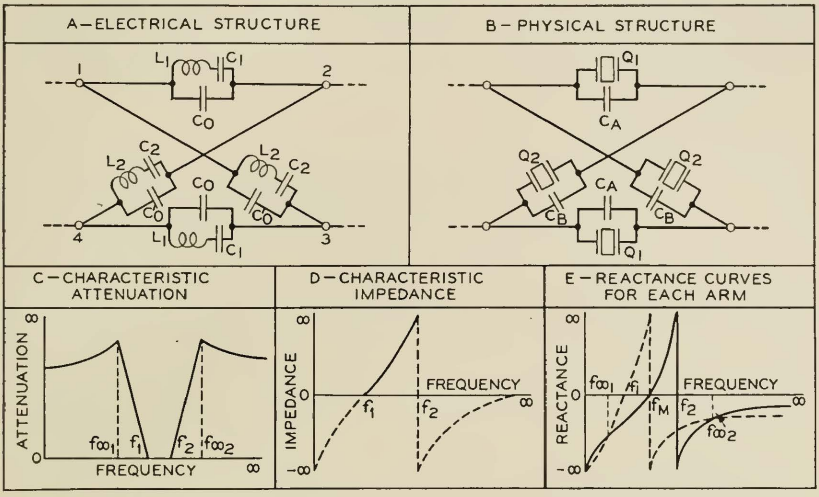


Fig. 7—Lattice network employing crystals and condensers.

is opposite to that of the other pair, a pass band exists, while if they are the same sign an attenuated band exists. Since the lattice is in the form of a Wheatstone bridge an infinite attenuation exists when the bridge is balanced, which occurs when both pairs of arms have the same impedance.

Let us consider a lattice filter with a crystal in each arm as shown by Fig. 7*B*. The crystals form two pairs of identical crystals, two alike in the series arms and two alike in the lattice arms. In order that a single band shall result it is necessary that the anti-resonant frequency of one arm coincide with the resonant frequency of the other as shown by Fig. 7*E*. It is obvious that the band width will be twice the width of the resonant region of the crystal or at most 0.8 per cent. Since the attenuation peaks occur when the two arms have the same impedance, they may be placed in any desired position by varying the impedance of one set of crystals with respect to the other. If crystals alone are used, these peaks will be symmetrical with respect to the pass band, but if in addition condensers are used with these crystals, the peaks may be made to occur dissymmetrically. In fact they may be made to occur so that both are on one side of the pass band. A narrower band results when capacitances are used in addition to crystals since the ratio of capacitances becomes larger. This may be utilized to control the width of the pass band to given any value less than 0.8 per cent.

The use of more crystals than four, in any network configuration employing only quartz crystals and condensers can be shown to result in no wider bands than 0.8 per cent, although higher losses can be obtained by the use of more crystals. Hence by the use of quartz crystals and condensers alone, a limitation in band width to 0.8 per cent is a necessary consequence of the fixed ratio of capacitances C_0/C_1 of equation (3).

Filter Sections Employing Crystals, Condensers and Coils

As was pointed out in the last section, filters employing only crystals and condensers are limited to band pass sections whose band widths do not exceed about 0.8 per cent. This band width is too narrow for a good many applications and hence it is desirable to obtain a filter section allowing wider bands while still maintaining the essential advantages resulting from the use of sharply resonant crystals. Such filters can be obtained only by the use of inductance coils as elements. Since the ratio of reactance to resistance of the best coils mounted in a reasonable space does not exceed 400, attention must be given to the effect of the dissipation.

The effect of dissipation in a filter is two-fold. It may add a constant loss to the insertion loss characteristic of the filter, and it may cause a loss varying with frequency in the transmitting band of the filter. The second effect is much more serious for most systems since an additive loss can be overcome by the use of vacuum tube amplifiers whereas the second effect limits the slope of the insertion loss frequency curve. Hence, if the dissipation in the coils needed to widen the band of the filter has only the effect of increasing the loss equally in the transmitting band and the attenuating band of the filter, a satisfactory result is obtained. The question is to find what configuration the coils must be placed in with respect to the crystals and condensers in order that their dissipation will not cause a loss varying appreciably with frequency.

Not every configuration will give this result, as is shown by the following example. The equivalent circuit of the crystal shown by Fig. 2*A* can be transformed into the form shown by Fig. 8*A* where the ratio $C_1/C_0 = 125$. This gives the same reactance curve as before, limited to a width of 0.4 per cent. Now suppose that we add an electrical anti-resonant circuit in series with the crystal—Fig. 8*B*—resonating at the same frequency and having the same constants as the anti-resonant network representing the crystal. If this circuit were dissipationless we could combine the two resonant circuits into one with twice the inductance and half the capacitance of that for the crystal alone

and hence the capacitance ratio would be 1/2 (125) or 62.5. The band width possible would then be twice that of the crystal alone. However, when the effect of dissipation is considered it is found that not much has been gained by employing the anti-resonant circuit. For the re-

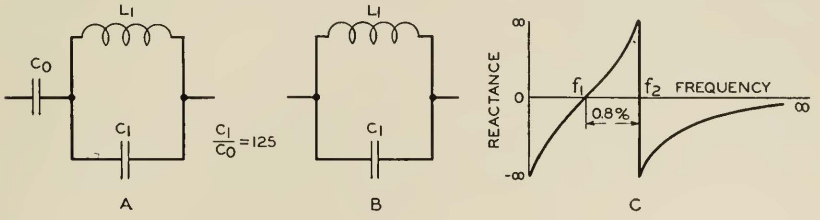


Fig. 8—Use of an anti-resonant circuit to broaden the resonance region of a crystal.

sistance, at resonance of the crystal and electrical circuit combination, will be the resistance of the electrical resonant circuit since that of the crystal is small compared to the electrical element. Hence we have doubled the impedance of the anti-resonant circuit and have the resistance of the electrical circuit. Hence the ratio (Q) of reactance to resistance of the anti-resonant circuit is double that of the electrical element alone. Even this Q, however, is insufficient to make a narrow band filter whose band width is 1.6 per cent (twice that possible with a crystal alone) and hence no useful purpose is served by combining a crystal with an electrical anti-resonant circuit.

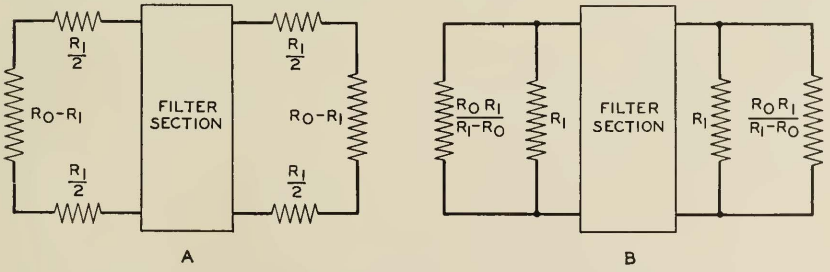


Fig. 9—Circuit showing resistances on the ends of filter sections.

Suppose, however, that all of the dissipation of the filter section be concentrated at the ends of the sections, either in series or in parallel with the filter as shown on Fig. 9. Then provided these resistances are within certain limits, they can be incorporated in the terminal resistances of the filter by making these resistances either smaller or larger for series or shunt filter resistances respectively. Between sections the resistances on the ends of the filter can be incorporated with other

resistances in such a way as to make a constant resistance attenuator of essentially the same impedance as the filter. For a series coil, this can be done by putting a shunt resistance between sections, while for a shunt coil it can be done by putting series resistances between sections. If this is done the whole effect of the dissipation is to add a constant loss to the dissipationless filter characteristic, this loss being independent of the frequency.

Since the lattice type network provides the most general type of filter network, attention will first be directed to this type of section employing inductances. It is easily proved that if any impedance is in series with both sides of a lattice network, as shown by Fig. 10*A*, then

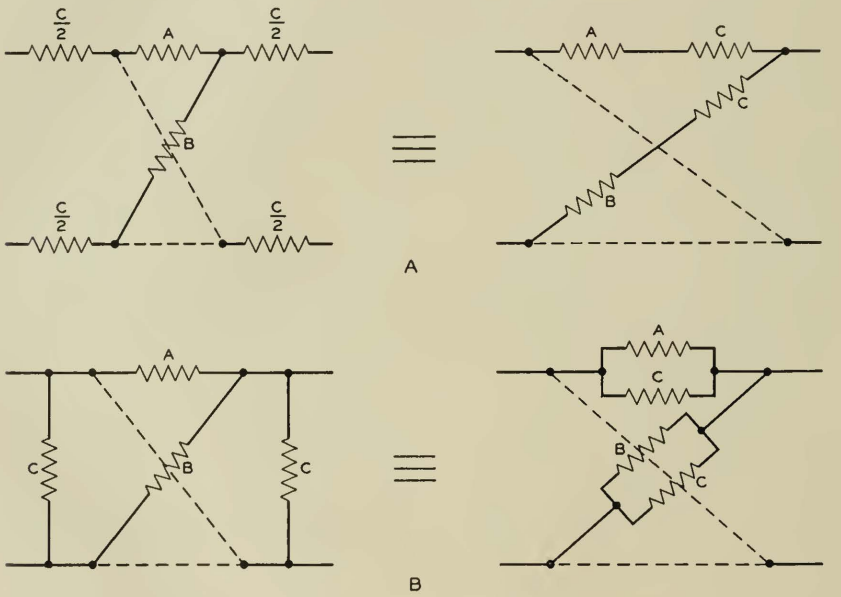


Fig. 10—Two network equivalences.

this is equivalent to placing this impedance in series with each arm of the lattice network as shown. Similarly, if a given impedance shunts the two ends of a lattice network, as shown by Fig. 10*B*, a lattice network equivalent to this is obtained by placing the impedance in shunt with all arms of the lattice. We are then led to consider a lattice network which contains coils either in series or in shunt with the arms of a lattice network, these arms containing only crystals and condensers, since the dissipation will then be effectively either in series or in shunt with the lattice network section.

If an inductance is added in series with a crystal the resulting re-

actance is shown by the full line of Fig. 11; the dotted lines show the reactance curves for the individual elements. It is evident that the resonant frequency of the crystal is lowered, the anti-resonant point remains the same, and an additional resonance is added at a frequency

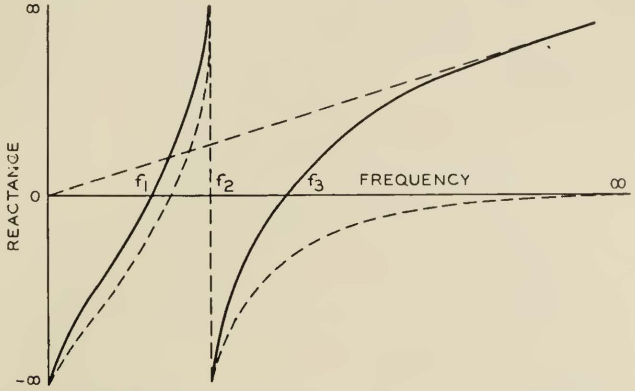


Fig. 11—Impedance characteristic of crystal and coil in series.

above the anti-resonant frequency. For a crystal whose ratio of capacitances r is about 125 it is easily shown by calculation that if the resonances are evenly spaced on either side of the anti-resonant frequency the percentage frequency separation between the upper resonance and the lower resonance is in the order of 9 per cent.

Suppose now that this element is placed in the series arm of a lattice network and another element of similar character is placed in the lattice arm, the second element having its lowest resonance coincide with the anti-resonance of the first element, and having the anti-resonance of the second element coincide with the highest resonance of the first element. This condition is shown by Fig. 12*C*. This network will produce a band-pass filter whose band extends from the lowest resonance of the series arm to the highest resonance of the lattice arm, a total percentage frequency band width of 13.5 per cent. By designing the impedances correctly the impedances of the two arms can be made to coincide three times so that there is a possibility of three attenuation peaks due to this section as shown by Fig. 12*D*. The loss introduced by the filter is equivalent to that introduced by three simple band-pass sections. Ordinarily the coils in the two arms are made equal so that their resistances are equal and for this case one of the peaks occurs at an infinite frequency. Since the resistances are equal, then by the theorem illustrated by Fig. 10*A* these resistances can be brought out on the ends and incorporated with the terminal

resistances, with the result that the dissipation of the coils needed to broaden the band has only the effect of adding a constant loss to the filter characteristic, this loss being independent of the frequency.

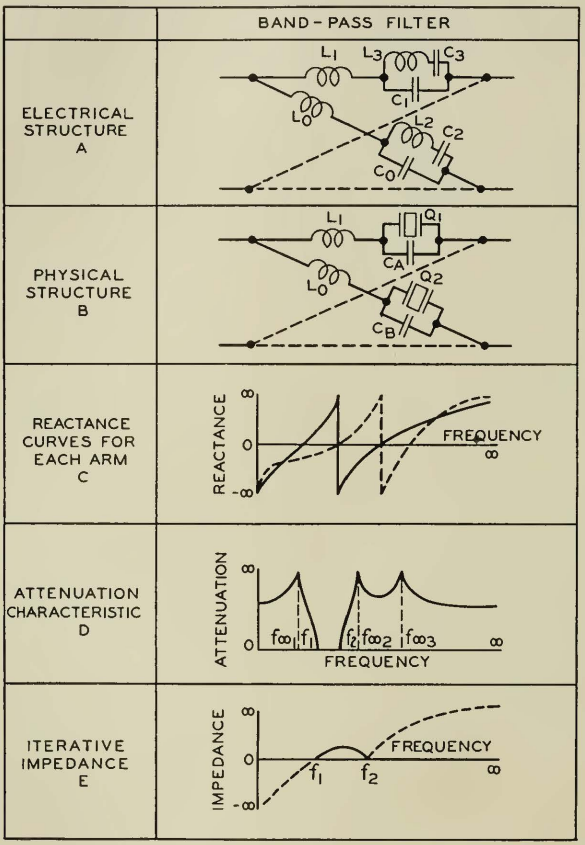


Fig. 12—Lattice network band-pass filter employing series coils.

To vary the width of the band below the 13.5 per cent band obtained with crystals only, added capacitances can be placed in parallel with the crystals increasing the ratio r. This results in a smaller separation in the resonant frequencies and hence a narrower band width. By this means the band width can be decreased indefinitely, although the dissipation caused by the coils introduces large losses for band widths much less than 1/2 per cent. By this means, however, it is possible to obtain band widths down to the widths which can be realized with crystals alone. On the upper side electrical filters can be built whose widths are as small as 13.5 per cent, hence this method fills

in a range not practical with electrical filters, or with crystals alone.

Another important characteristic of the filter is its iterative impedance. For a lattice filter this is given by [9]

$$Z_I = \sqrt{Z_1 Z_2},$$

where Z_1 is the impedance of the series arm and Z_2 that of the lattice arm. For a dissipationless filter, this is shown by Fig. 12*E*, as can be easily verified by a consideration of the reactance curves of Fig. 12*C*.

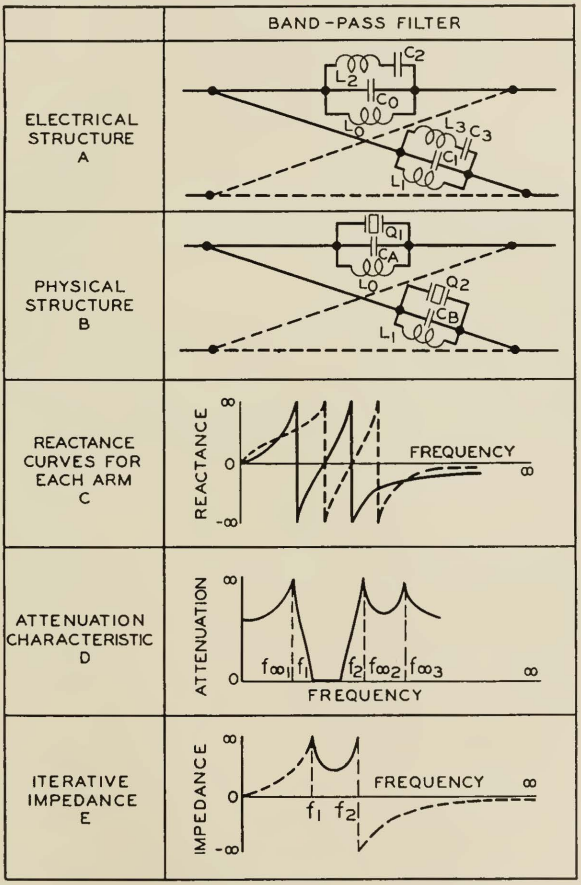


Fig. 13—Lattice network band-pass filter employing parallel coils.

This type of filter results in a relatively low impedance, for example about 600 ohms for a filter whose mid-band frequency is 64 kilocycles and whose band width is that shown on Fig. 19. Since the band width is decreased by adding more capacitance, it is evident that smaller

percentage band width filters will have lower impedances than the wider ones. For example, the filter whose characteristic is shown by Fig. 20, has an iterative impedance of 25 ohms.

It is evident that a still wider band can be obtained with the section discussed above by making the two resonances of Fig. 11 dissymmetrical. If the lower one is brought in closer to the anti-resonant frequency the top one extends farther out in such a manner that the total percentage frequency separation is greater than 9 per cent. If one element of this type is combined with one whose lower resonance is brought farther away from the anti-resonance than is the upper resonance, a filter whose pass band is greater than 13.5 per cent is readily obtained. On the other hand as the band is widened by this means, the cross-over points of the impedances of the two arms are of necessity brought very close to the cut-off frequencies, so that such a filter would introduce most of its loss very close to the cut-off frequencies. This type of characteristic might be useful in supplementing the loss characteristic possible with electrical elements, but by itself would not produce a very useful result.

We have so far discussed the characteristics which can be obtained by placing coils in series with crystals. An equally useful result is obtained by placing coils in shunt with crystals as shown by Fig. 13*B*. This arrangement results in a band-pass filter capable of giving the same band width as the first type discussed above. The only difference

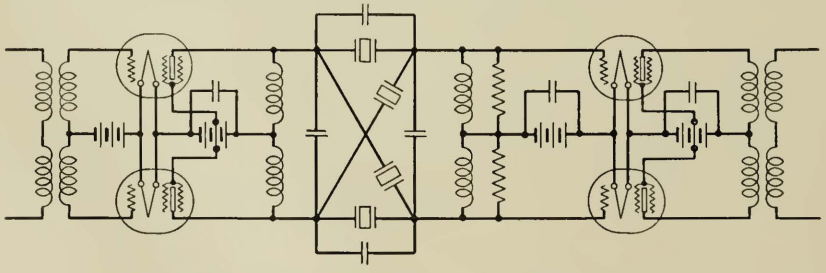


Fig. 14—Band-pass filter used between vacuum tubes.

occurs in the iterative impedance which will be as shown by Fig. 13*E*. For narrow band widths this type of filter has a very high iterative impedance. For example, for a one per cent band width, using ordinary sized coils and crystals, the iterative impedance may be as high as 400,000 ohms. Such filters can be used advantageously in coupling together high impedance screen gird tubes without the use of transformers. One such circuit is shown schematically by Fig. 14.

Filters made by using either series or shunt coils in conjunction

with condensers and crystals make very acceptable band-pass filters capable of moderate band widths. It is often desirable to obtain low and high-pass filters having a very sharp selectivity. The filter of Fig. 12 can be modified to give a high-pass characteristic by leaving out the coils in the series or lattice arms of the network. However, it will be found that the cross-over points in the impedance curve of necessity come very close to the pass band and hence no appreciable loss can be maintained at frequencies remote from the pass band. A broader and more useful characteristic is obtained by using a transformer having a preassigned coefficient of coupling, in conjunction with crystals and condensers, as the element for broadening the separation of resonances. Such an element is shown by Fig. 15*A*. As is well known, a trans-

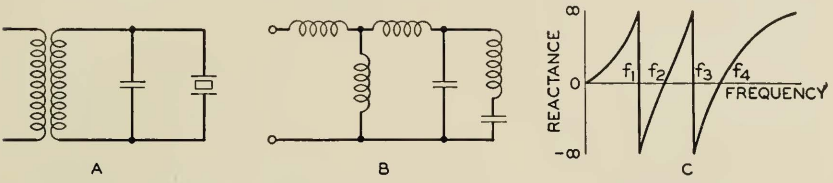


Fig. 15—Impedance characteristic of a transformer, condenser, and crystal.

former with a specified coupling can be replaced by a T network of three inductances as shown by Fig. 15*B*. The impedance characteristic, as shown by Fig. 15*C*, has two anti-resonant frequencies f_1 and f_3, and two resonant frequencies f_2 and f_4.

Suppose now that an element of this type is placed in one arm of the lattice and a similar element having a condenser in series with it is placed in the other arm as shown by Fig. 16*A*. If the elements are so proportioned that the anti-resonances of one arm coincide with the resonances of the other arm and vice versa, as shown by Fig. 16*B*, the impedances of the two arms are of opposite sign till the last resonance. Hence, a low pass filter results. It is possible to make the two impedance curves cross five times, so that an attenuation corresponding to five simple sections of low-pass filter results. Other arrangements of the resonances are also possible and are advantageous for special purposes. For example, as shown by Fig. 16*C*, we can make the last resonance and anti-resonance of both arms coincide, and the other resonances of one arm coincide with the anti-resonances of the second arm. This arrangement results in a low-pass filter having an attenuation corresponding to three simple low-pass filter sections and an impedance which can be made nearly constant to a frequency very near the cut-off frequency. This is advantageous for obtaining a filter with

a sharp cut-off, for otherwise the mismatch of impedance near the cut-off frequency causes large reflection losses which prevent the possibility of obtaining a sharp discrimination.

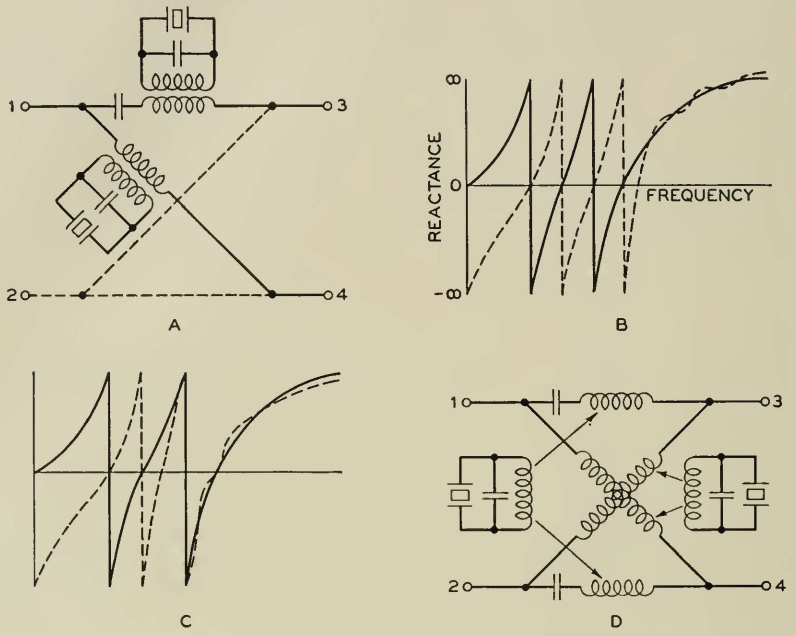


Fig. 16—Lattice network low-pass filter employing transformers, condensers, and crystals.

The effect of dissipation in the transformer on the loss characteristic is not so easy to analyze in this case as in the case of a series coil. The effect can be obtained approximately as follows. Of the three coils of Fig. 15*B* representing the transformer, the shunt coil has the least dissipation since no copper losses are included in this coil. For an air core coil, the Q of this shunt coil becomes very high and its dissipation can be neglected. The resistance of the primary winding can be incorporated in the terminal resistance as in the series coil type of filter and hence will cause only an added loss. The resistance of the secondary will be in series with the crystal and condenser, and for a reasonably good coil is of the same order of magnitude as the crystal resistance at resonance. Hence its effect will be much the same as cutting the Q of the crystal in half, so that instead of a crystal whose Q is 10,000, we use one whose Q is 5000 and a dissipationless coil. We see then that the Q of the crystal is still the most important factor in determining the sharpness of cut-off in the filter as in the previous

ones described, and hence a very sharp selectivity can be obtained with this circuit. It is possible to save elements in this filter by using two primaries for each coil, putting one primary in one series or lattice arm and the other in the corresponding series or lattice arms as shown by Fig. 16*D*. Only half the number of elements per section are required.

By replacing the series condenser of the series arm of Fig. 16*A* by a parallel condenser, it is possible to change the filter from a low-pass to a high-pass filter. Condensers in series, or in parallel with both arms result in wide band-pass filters. It is possible to obtain a wider pass band with this type of filter than with the single coil type since the resonances will be spread over a wider range of frequencies.

In a good many cases it is desirable to have unbalanced filter sections rather than the balanced type which results from the use of a lattice network. This is particularly true for high impedance circuits for use with vacuum tubes. Since the lattice type section is the most general type, it gives the most general characteristics obtainable. The filter sections described here can in some cases be reduced to unbalanced bridge T sections by well known network transformations, with, however, more restrictions on the type of attenuation characteristics physically obtainable.

A very simple bridge T network, which is equivalent to a lattice network of the kind shown on Fig. 13, with two crystals replaced by condensers, is shown on Fig. 17. This section employs mutual induct-

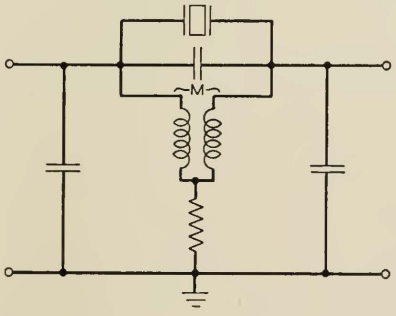


Fig. 17—Single crystal bridge T band-pass filter.

ance, and the resistance [10] shown is necessary in order to balance the arms of the equivalent lattice. This type of network is able to reproduce some of the characteristics of the lattice filter, but is not so general and is, moreover, affected by the dissipation of the coil to a larger extent than the equivalent lattice.

[10] The use of this resistance was suggested by Mr. S. Darlington and practically all the work of developing this filter has been done by Mr. R. A. Sykes.

Experimental Results

A number of filters have been constructed, during the past four years, which employ quartz crystals as elements. Figure 18 shows the measured insertion loss characteristic of a narrow band filter [11] employ-

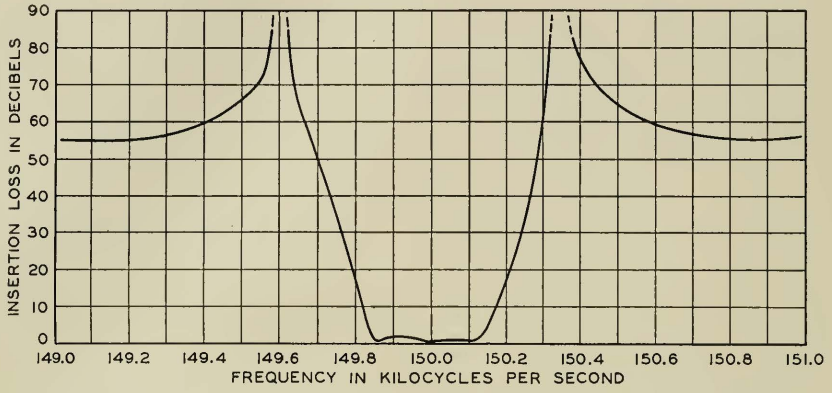


Fig. 18—Measured insertion loss characteristic of a narrow band-pass filter.

ing only crystals and condensers. This filter employs two sections of filter No. 3 of Fig. 6. It will be noted that in spite of the very narrow band width, the insertion loss in the transmitted band is quite small.

A number of the broader band filters employing coils as well as condensers and crystals have also been constructed. The frequency range so far developed extends from 36 kilocycles to 1200 kilocycles. Figure 19 shows the insertion loss characteristic of a band-pass filter whose mid-frequency is 64 kilocycles and whose band width is 2500 cycles. The insertion loss rises to 75 db, 1500 cycles on either side of the pass region. This filter was constructed from two sections of the band-pass type described in Fig. 12. A similar insertion loss characteristic, but shifted to a higher frequency, is shown by Fig. 20. The insertion loss in the center of the band for this higher frequency filter is considerably larger due to the smaller percentage band width. It is interesting to note that practically all of this loss is due to the dissipation introduced by the coils. The useful percentage band width is about one-half per cent and the filter reaches its maximum attenuation

[11] The filters whose characteristics are shown on Figs. 18 and 21 were designed and constructed by Messrs. C. E. Lane and W. G. Laskey. The author wishes to call attention to the fact that they and others associated with them in the Laboratories have made considerable progress in connection with the practical difficulties encountered in the design and construction of these filters such as working out the high precision element adjustment methods required, in methods of mounting, and in shielding methods.

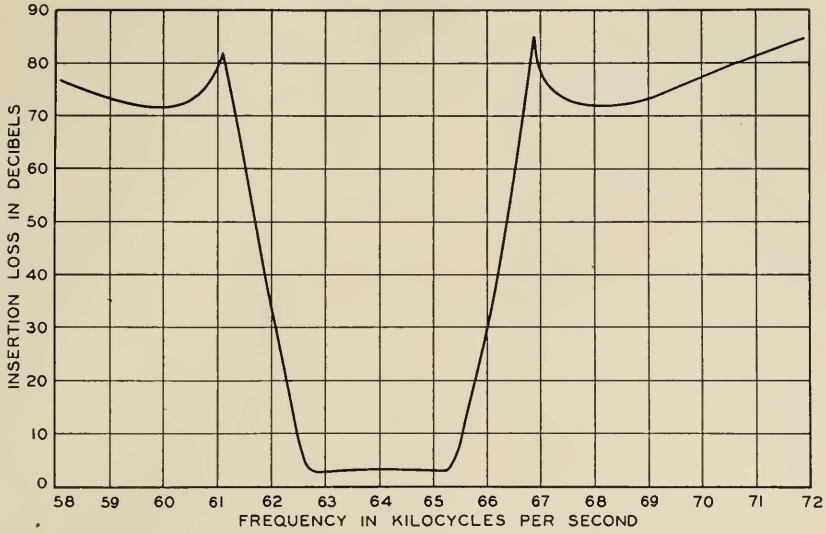


Fig. 19—Measured insertion loss characteristic of a band-pass filter.

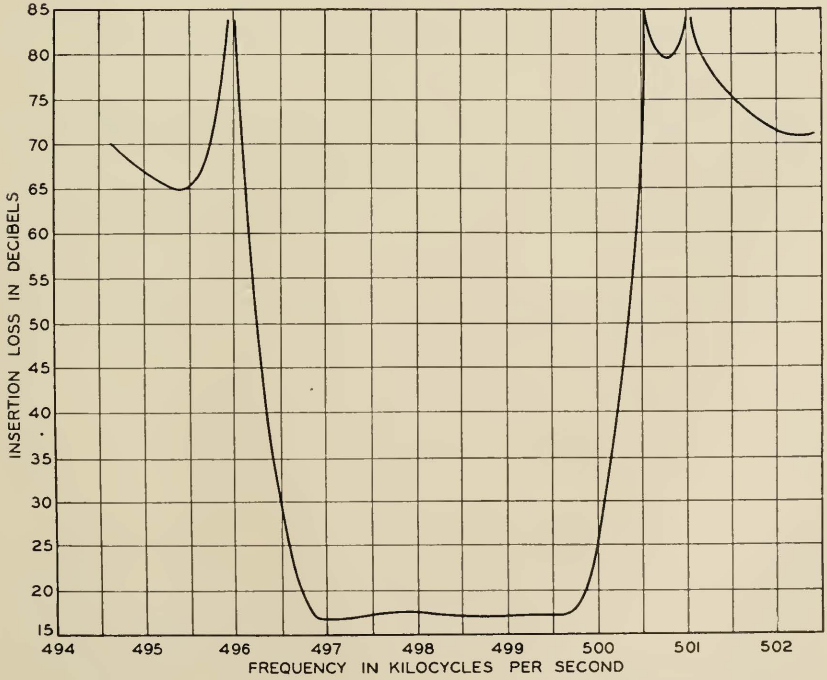


Fig. 20—Measured insertion loss characteristic of a band-pass filter at a high frequency.

in less than one-fourth per cent frequency range on either side of the pass band.

Figure 21 shows the insertion loss characteristics of a filter employed

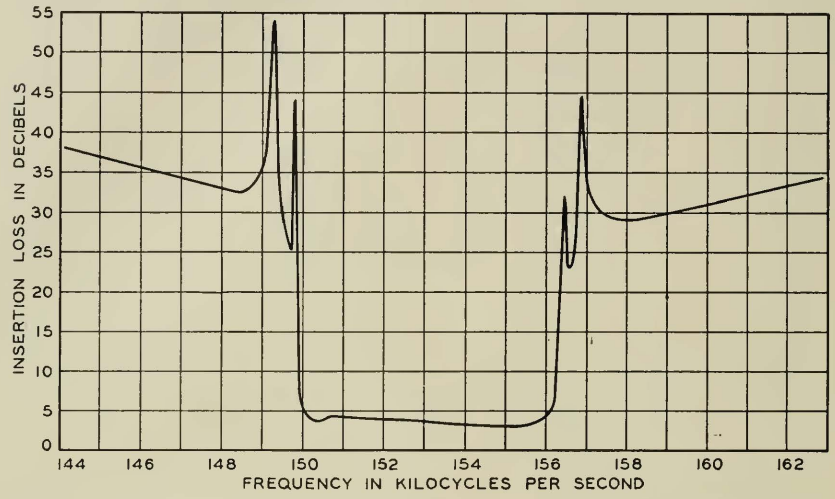


Fig. 21—Measured insertion loss characteristic of a band-pass filter used in a single side band radio receiver.

in an experimental radio system for separating the two sidebands of a channel at a high frequency. Here the separation is effected in about 0.15 per cent frequency range. With the best electrical filters the frequency space required for such a separation is about 1.5 per cent.

Figure 22 shows the insertion loss characteristic of a high-pass filter

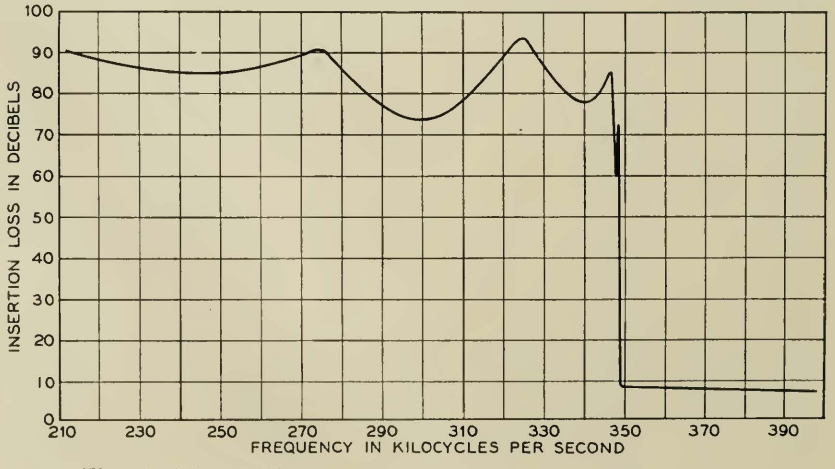


Fig. 22—Measured insertion loss characteristic of a high-pass filter.

constructed by using the circuit of Fig. 16*D*, modified by using a parallel condenser rather than a series condenser. The filter obtains a 65 db discrimination in less than a 0.12 per cent frequency separation.

Figure 23 shows a characteristic obtained by employing a filter of

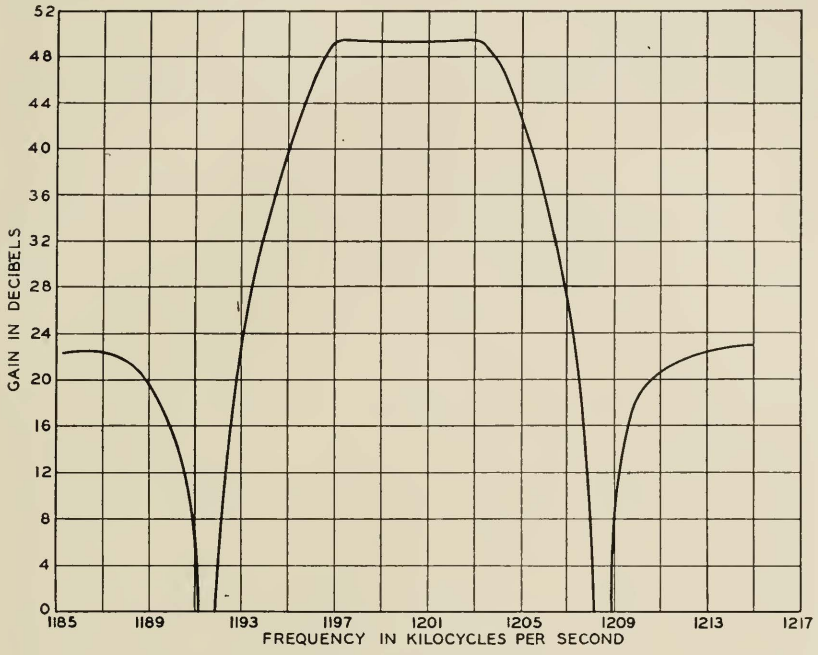


Fig. 23—Measured insertion loss characteristic of a single crystal bridge T filter.

the type shown by Fig. 17, together with a screen grid vacuum tube. The result is plotted as the gain of the circuit since this gives the most significant result for this type of circuit.

APPENDIX

The Modes of Vibration of a Perpendicularly Cut Quartz Crystal

Introduction

Quartz crystals have been cut into two principal types of orientations with respect to the natural crystal faces. The first type is the so-called Curie or perpendicular cut in which the crystal is so cut that its major surfaces are perpendicular to an electrical axis and parallel to the optical axis. Figure 1 shows such a cut. The second type is the so-called parallel or 30-degree cut in which the major surfaces are parallel

to both the optical and electrical axes. In this appendix a study is made of the modes of motion of a perpendicularly cut crystal. The effect has been studied of rotating the direction of the principal axis while still maintaining the principal surfaces perpendicular to the electrical axis. Such a crystal is designated as a perpendicularly cut crystal with an angle of rotation θ.

The perpendicularly cut crystal has received considerable theoretical and experimental consideration especially from Cady,[2] Van Dyke,[3] Dye [4] and Vigoreux.[5] They have assumed that the crystal has a plane wave vibration, and have calculated the frequencies of resonance in terms of the elastic constants and the density of the crystal, and have derived equivalent electrical networks for giving their electrical impedance. Such representations indicate that there should be one resonance for the crystal, the frequency of which is inversely proportional to the length and independent of the width of the crystal. As long as the length of the mechanical axis is large compared to that of any other axis, this prediction agrees with the experiment, but when the length of the other axes become comparable with that of the mechanical axis, the prediction is no longer fulfilled by experiment. It has long been recognized that this deviation is due to the failure of the plane wave assumption. Rayleigh [12] has given a correction for taking account of lateral motion, which is applicable to an isotropic medium. In a crystal, shear vibrations may be set up as well and for this case Rayleigh's correction can only be regarded as qualitative. Also if the other sets of resonance frequencies are to be investigated, account must be taken of the resonances of the other modes of vibration, and their reaction on the mode to be studied.

In this appendix experimental results have been obtained showing the frequencies of resonance found in perpendicularly cut crystals of various shapes and orientations. These frequencies are correlated with the elastic constants of the crystal and are shown to be completely accounted for by them. A coupled circuit representation is developed which is capable of predicting the main features of the principal vibration, including the change of frequency with the shape and orientation of the crystal, and the temperature coefficient curves.

Experimental Determination of the Resonant Frequencies

In order to investigate the modes of motion in a perpendicularly cut crystal in which the main axis coincides with the mechanical axis of the crystal, a set of measurements has been made on crystals whose

[2, 3, 4, 5] Loc. cit.

[12] Rayleigh, "Theory of Sound," Vol. I, Chapter VII, page 252.

mechanical axes are all 1.00 centimeter long, whose electrical axes are very thin, being 0.05 centimeter, and whose optical axes vary in dimension from 0.1 centimeter to 1.00 centimeter. In order to eliminate the effect of a series capacitance due to an air gap, the crystals were plated with a very thin coat of platinum. The effect of an added shunt capacitance in parallel with the crystal, due to the electrode capacitances, was practically eliminated by running the crystal electrodes in an outer grounded conductor as shown in Fig. 24, which shows the

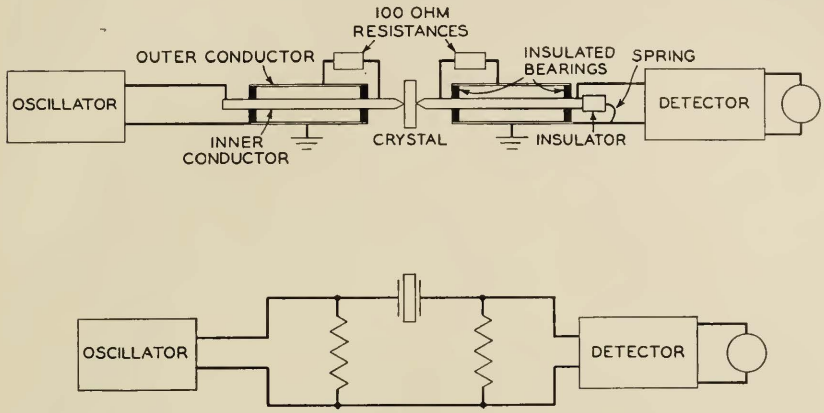


Fig. 24—Measuring circuit used to measure the resonances of a crystal.

measuring circuit. Contact to the crystal plating is made by means of small electrode points placed at the center of the crystal and kept in place by a small pressure. An increase in pressure over a moderate range was found to have no effect on the frequency of the crystal. The lowering of frequency due to plating was evaluated by depositing several films of known weight on the crystal and plotting its resonant frequency as a function of film weight. The intercept of this curve for a zero plating was taken as the frequency of the unplated crystal.

When the frequency of the oscillator was varied, the current in the detector showed frequencies of maximum and minimum current output which are respectively the frequencies of resonance and anti-resonance of the crystal. In order to locate accurately the frequencies of anti-resonance, it was found necessary to insert a stage of tuning in the detector, in order to discriminate against the harmonics of the oscillator. For a given crystal the frequency of the oscillator was varied over a wide range and the resonant and anti-resonant frequencies of the crystal were measured. The results of these measurements are shown by Fig. 25. In this curve the bottom part of the line repre-

sents the actual measured frequency, while the width of the line is proportional to the frequency difference between resonance and anti-resonance. In order to make this quantity observable, the frequency difference between resonance and anti-resonance is multiplied by a factor 6.

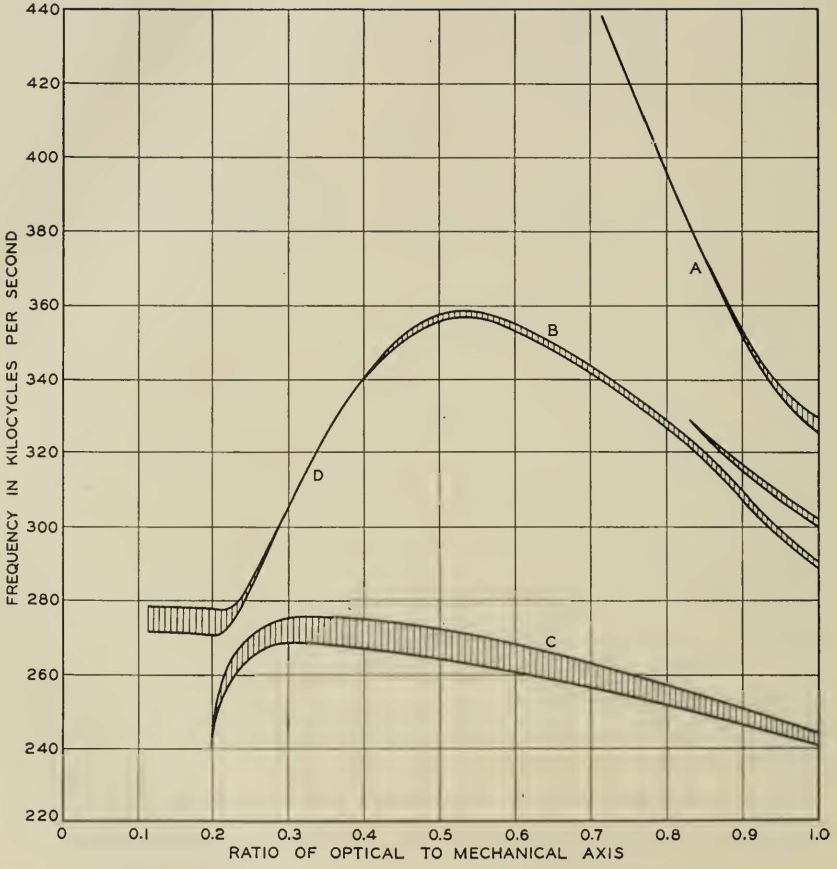


Fig. 25—Measured resonances of a perpendicularly cut crystal.

As long as the ratio of the optical to the mechanical axis is less than 0.2, the assumption of plane wave motion agrees well with experiment since there is only one resonance and its frequency does not depend to any great extent on the optical axis. However, above this point two frequencies make their appearance and react on each other to produce the coupled circuit curve shown. Finally when the ratio of optical to mechanical axes becomes larger a total of four resonant frequenices appear. Since a large number of crystals are used whose ratios of

optical to mechanical axes are greater than 0.2, it becomes a matter of some importance to investigate the causes of the additional resonances.

Interpretation of the Measured Resonance Frequency Curves of a Perpendicularly Cut Crystal

The plane wave assumption is valid for crystals whose width is less than 1/5 of their length, but it fails for wider crystals. It fails to represent a rectangular crystal because it does not allow for a wave motion in any other direction. That such a motion will occur is readily found by inspecting the stress-strain equations of a quartz crystal, given by equation (7).

$$\begin{aligned} -x_x &= s_{11}X_x + s_{12}Y_y + s_{13}Z_z + s_{14}Y_z, \\ -y_y &= s_{12}X_x + s_{11}Y_y + s_{13}Z_z - s_{14}Y_z, \\ -z_z &= s_{13}X_x + s_{13}Y_y + s_{33}Z_z, \\ -y_z &= s_{14}X_x - s_{14}Y_y + s_{44}Y_z, \\ -z_z &= s_{44}Z_x + s_{14}X_y, \\ -x_y &= s_{14}Z_x + \tfrac{1}{2}(s_{11} - s_{12})X_y, \end{aligned} \tag{7}$$

where x_x, y_y, z_z are the three components of extensional strain, and y_z, z_x, x_y the three components of shearing strains. X_x, Y_y, Z_z, Y_z, Z_x, and X_y are the applied stresses and s_{11}, etc. are the six elastic compliances of the crystal. Their values are not determined accurately but the best known values are given in equation (42). In this equation the X axis coincides with the electrical axis of the crystal, the Y axis with the mechanical axis, and the Z axis with the optical axis.

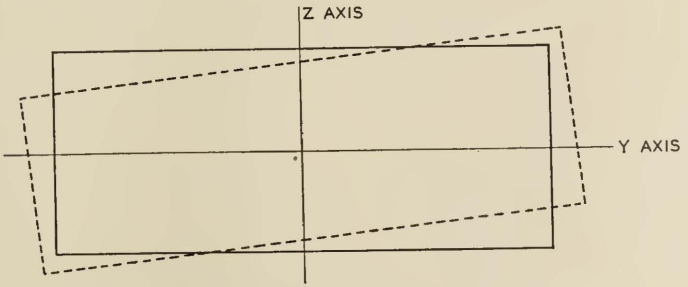


Fig. 26—Form of crystal distorted by an applied Y_y force.

Limiting ourselves now to an X or perpendicularly cut crystal the only stresses applied by the piezo-electric effect are an X_x, a Y_y, and a Y_z stress. Hence for such a crystal only four of the six possible types of motion are excited, three extensional motions x_x, y_y, z_z and one shear

motion y_z. Under static conditions, then, the motion at any point in the crystal is given as the sum of four elementary motions, three extensional motions and one shear motion. Moreover, these motions are coupled [13] as is shown by the fact that a force along one mode produces displacements in other modes of motion. Figure 26 shows how a perpendicularly cut crystal will be distorted for an applied Y_y force.

Suppose now that an alternating force is applied to the crystal. The simplest assumption that we can make regarding the motion is that the motion of any point is composed of four separate plane wave motions of the four types of vibration and that these react on each other in the way coupled vibrations are known to act in other mechanical [14] or electrical circuits. For the present purpose we can neglect motion along the X or electrical axis since this axis has been assumed small. The three remaining motions if existing alone will have resonances as shown by the solid lines of Fig. 27. That along the mechanical axis will have a constant frequency, since the mechanical axis is assumed constant, and is shown by the line C. The extensional motion along the optical axis will have a frequency inversely proportional to the length of the optical axis and will be represented by the line A of the figure. The shear vibration y_z, as shown by the section on the resonance frequency of a crystal vibrating in a shear mode, will have a frequency varying with dimension as shown by the line B.

In view of the coupling between the motions, the actual measured frequencies will be as shown by the dotted lines in agreement with well known coupled theory results.

If we compare these hypothetical curves with the actual measured values some degree of agreement is apparent. The main resonant frequency except in the region $0.2 < l_0/l_m < 0.3$ follows the dotted curve drawn. Also, the extensional motion along the optical axis has a frequency agreeing with that of Fig. 25. The shear vibration, however, has an entirely different curve from that conjectured. What is happen-

[13] The idea of elementary motions in the crystal being coupled together appears to have been first suggested in a paper by Lack "Observations on Modes of Vibration and Temperature Coefficients of Quartz Crystal Plates," *B. S. T. J.*, July, 1929, and was used by him to explain the effect of one mode of motion on the temperature coefficient of another mode and vice versa. The idea of associating this coupling with the elastic constants of the crystal occurred to the writer in 1930 but was not published at that time. It is, however, incorporated in a patent applied for some time ago on the advantages of crystals cut at certain orientations. More recently the same idea is given in a paper by E. Giebe and E. Bleckschmidt, *Annalen der Physik*, Oct. 16, 1933, Vol. 18, No. 4. They have extended their numerical calculations to include three modes of motion.

[14] This coupling is shown clearly for a mechanical system by one of the few rigorously solved cases of mechanical motion for two degrees of freedom—the vibration of a thin cylindrical shell—given by Love in "The Mathematical Theory of Elasticity," Fourth Edition, page 546.

ing there is I think evident from a consideration of Fig. 28. Here in solid lines are drawn two frequency curves one of which, *B*, is the shear frequency curve of Fig. 27. The other curve, *D*, has a rising frequency with an increase in the optical axis dimension. Assuming these vibra-

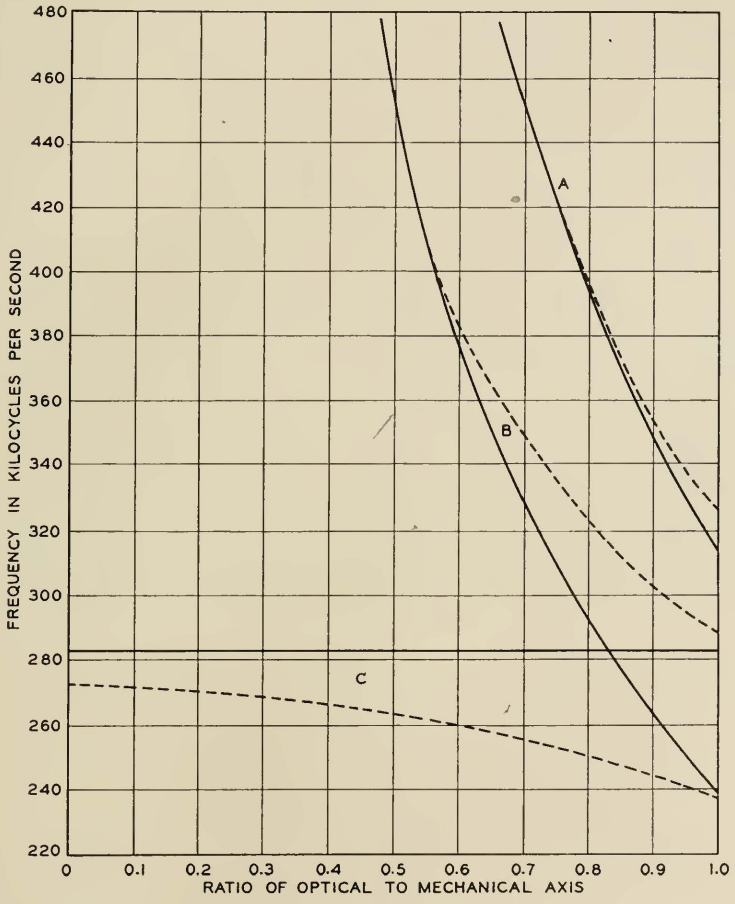


Fig. 27—Theoretical resonances of a perpendicularly cut crystal showing effect of coupling.

tions coupled a resonance frequency curve shown by the dotted line will be obtained. If this curve is substituted for the shear curve of Fig. 27 and the actual resonant frequency raised to take account of the effect of coupling with the longitudinal motion along the mechanical axis, a curve very similar to the measured curve of Fig. 25 is obtained.

The type of motion coupled to the shear motion is easily found. Its

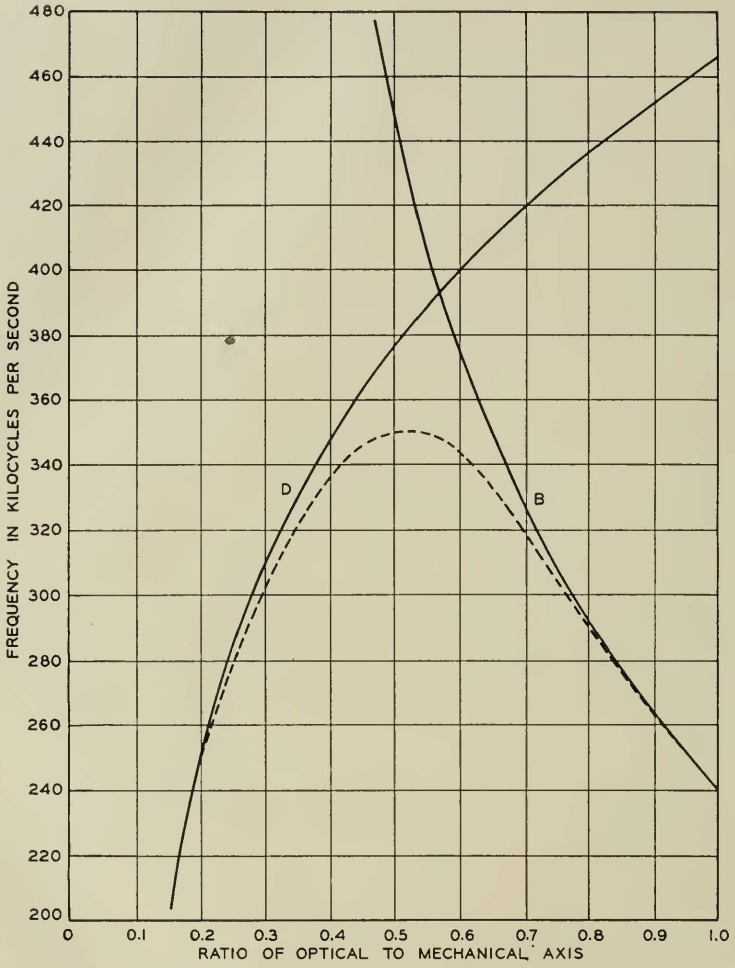


Fig. 28—Coupled frequency curve for shear and flexure vibrations.

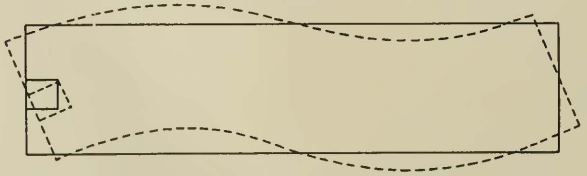


Fig. 29—Bar bent in its second flexural mode of vibration.

frequency increases as the optical axis dimension is increased and about the only type of motion which does this is a flexural motion as shown by Fig. 29. This figure shows the second type of motion possible to a bar in flexure rather than the first for experiments by Harrison [15] show that the frequency for the first type of motion is too low to account for this vibration. Harrison has also measured the frequencies of a bar in its second flexural mode and the solid line, D, of Fig. 28 is an actual plot of these measured frequencies up to a ratio of $l_0/l_m = 0.25$, which is as far as Harrison carries his measurements. The rest of the curve is obtained by extrapolation. There is no doubt then that a flexural motion is involved in this coupling. The mechanism by which the bar is driven in flexure will be evident if we observe what happens to a square on the crystal in the unstrained state. As shown by Fig. 29, its deformation is similar to that of a shear deformation. The amount of shear depends on the distance from the nodes of the crystal. Some of the shear is in one direction and some in the other but the two amounts are not balanced and hence a pure shear in one direction can excite a flexural motion of the crystal.

The strength of the coupling from the mechanical axis motion y_y to the shear motion y_z and the extensional motion along the optical axis z_z are indicated by the coupling compliances $s_{24}/\sqrt{s_{22}s_{44}}$ and $s_{23}/\sqrt{s_{22}s_{33}}$, respectively. From the values of these constants we find that the shear motion is more closely coupled than the z extensional motion, and this is indicated experimentally by the greater width of the shear line.

Effect of a Rotation of the Longest Axis with Respect to the Electrical Axis on the Resonances of a Crystal

From the qualitative explanation of the secondary resonances given above, it is possible to predict how these resonances will be affected by any change in the crystal which changes the constants determining the three modes of motion and their coupling coefficients. One method for varying these constants is to change the direction for cutting the crystal slab from the natural crystal. In the present paper consideration is limited to those crystals which have their major faces perpendicular to an electrical axis, i.e., a perpendicularly cut crystal with its longest direction rotated by an angle θ from the direction of the mechanical axis. The convention is adopted that a positive angle is a clockwise rotation of the principal axis for a right handed crystal, when the electrically positive face (determined by a squeeze) is up.

[15] "Piezo-Electric Resonance and Oscillatory Phenomena with Flexural Vibration in Quartz Plates," J. R. Harrison, *I. R. E.*, December, 1927.

For a left-handed crystal a positive angle is in a counter-clockwise direction.

In the section dealing with elastic and piezo-electric constants for rotated crystals (page 449) is given a method for determining the elastic constants of a rotated crystal and curves are given for the ten elastic constants. These have been worked out by Mr. R. A. Sykes of the Laboratories. The method of designation is the following: The X axis remains fixed and is designated by 1′. The axis of greatest length is designated by 2′, since in the unrotated crystal the mechanical axis, corresponding to the Y direction, is the axis of greatest length. Extensional motion perpendicular to the 2′ axis is designanted by 3′, and shear motion in the plane determined by the 2′, 3′ axes is designated by 4′. The ten resulting constants s_{11}', s_{22}', s_{33}', s_{44}', s_{12}', s_{13}', s_{14}' s_{23}', s_{24}', s_{34}' are shown evaluated in terms of the angle θ on Fig. 30. Since

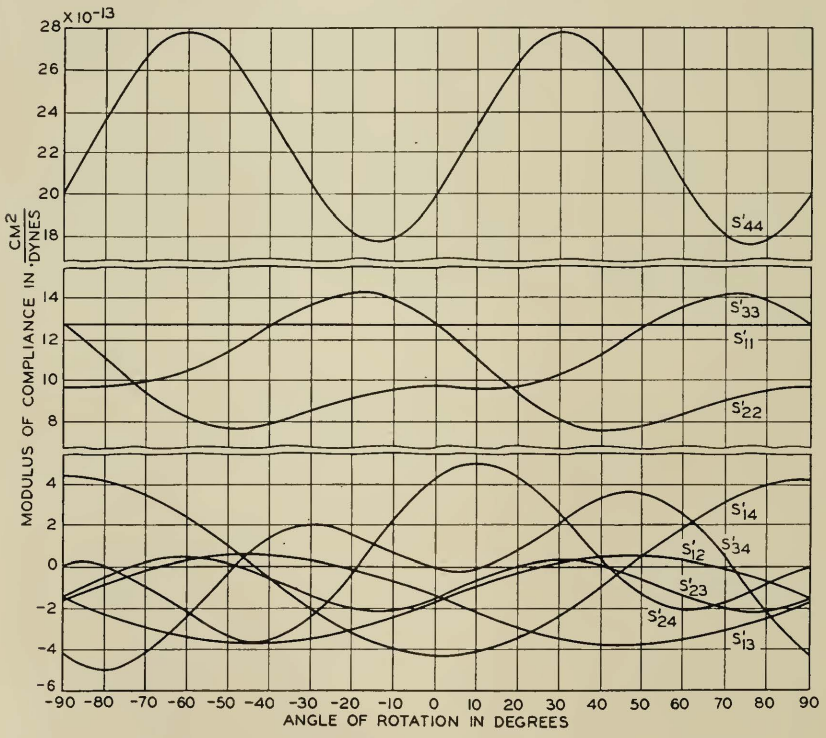


Fig. 30—Elastic compliances of a perpendicularly cut crystal as a function of the angle of rotation.

motion and coupling to motion along the X axis can be neglected, the constants of interest are s_{22}', s_{33}', s_{44}', s_{23}', s_{24}', s_{34}'. Since the 2′ or y'

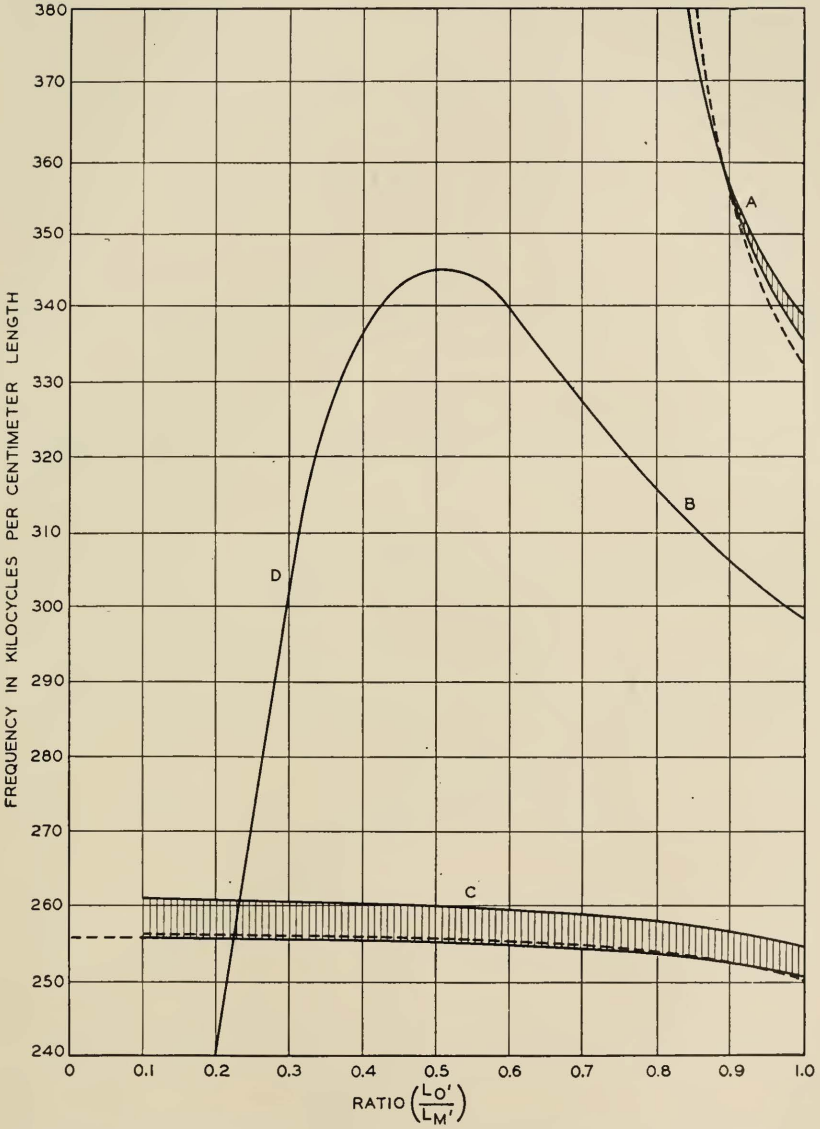


Fig. 31—Measured resonances of a $\theta = -18.5°$ perpendicularly cut crystal.

axis is the principal axis of motion, the mutual compliances of principal interest are s_{23}', determining the coupling between the Y' extensional motion and the Z' extensional motion, and s_{24}' determining the coupling between the Y' extensional motion and the Y_z' shear motion. It is the shear motion which is most objectionable, because it is more highly coupled than the Z' extensional motion, because it is lower in frequency, and because it is coupled to a flexural mode. Hence, if this motion can be eliminated or made very small, a much better crystal for most purposes is obtained. We note that if θ is $-18.5°$ or if $\theta = 41.5°$ the shear coupling coefficient vanishes and hence a force in the Y_y direction produces no y_z shear or vice versa. Of these the $-18.5°$ crystal is driven more strongly by the piezo-electric effect and hence has a more prominent resonance.

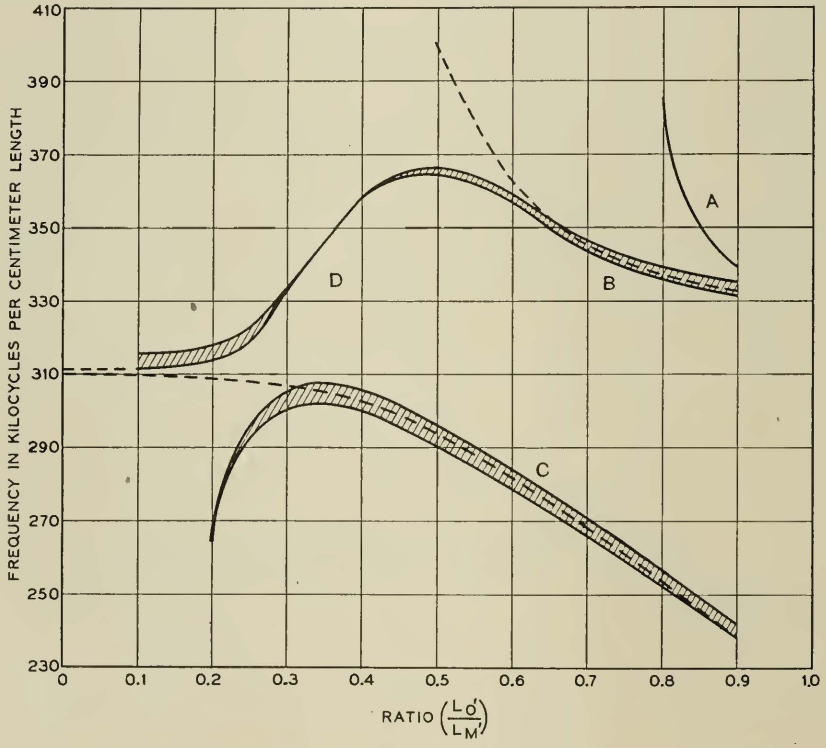


Fig. 32—Measured resonances of a $\theta = +18.5°$ perpendicularly cut crystal.

Accordingly the resonances of a $\theta = -18.5°$ cut crystal have been measured in a similar way to the $\theta = 0°$ cut crystal shown in Fig. 25. The result is shown on Fig. 31. As will be seen from the figure, the

shear resonance indicated by B is barely noticeable, while the z extensional mode indicated by A is somewhat stronger although higher in frequency. The frequency of the principal mode is not greatly affected by an increase in the z' axis until the ratio of axes is greater than .6. Another angle of some interest is $\theta = +18.5°$ since there the z' extensional coupling disappears. The resulting resonances are shown on Fig. 32. It will be noted that the z' extensional resonance curve A is very weak, while the shear curve B is quite pronounced.

An Equivalent Electrical Circuit for a Crystal Possessing Two Degrees of Motion

The above explanation accounts qualitatively for all the resonances observed in the crystal and how they are varied by a rotation of the crystal. It is desirable, however, to see if a quantitative check can be obtained from the known elastic constants of the crystal. To obtain a complete check would require a system capable of five degrees of motion. However, if we take the simplest case, the -18.5 degree cut crystal, only two modes of motion have to be considered, and even for the zero cut crystal, a good agreement is obtained by lumping the shear and extensional mode as one mode of motion and considering its reaction on the fundamental mode. Hence consideration is limited in this paper to a circuit having two modes of motion.

The properties of a single mode of motion can be represented for frequencies which do not exceed the first resonant frequency of the crystal, by the simple electrical circuit of Fig. 33*A*. Here the capaci-

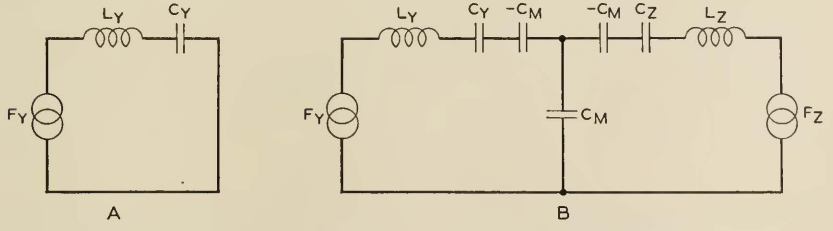


Fig. 33—Equivalent electrical circuit of a crystal having two modes of motion.

tance represents the mechanical compliance of the bar, the charge on the condenser represents a displacement per unit length of the bar, while the current flowing through the circuit represents the velocity of a point on the bar. The inductance represents the mass reaction of the crystal. The representation of the motion of a bar by a simple lumped circuit assumes that the bar moves as a whole, that is, if a force is applied to the body it contracts or expands equally at all parts of the

bar. This is contrary to actual conditions, since expansions or contractions proceed in the form of a wave from the ends of the bar toward the center. However, if consideration is limited to low frequencies, i.e. frequencies which do not exceed by much the first resonance of the bar, the approximation is good and a considerable simplification in the analysis is made. To take account of wave motion, the representation has to be an electric line as was pointed out in connection with acoustic filters.[16]

To represent two separate modes of motion and their coupling, the circuit shown by Fig. 33*B* is employed. A little consideration shows that the type of coupling existing in a crystal is capacitative since an extension along the mechanical axis produces a contraction along the optical axis, and vice versa. Since strains in mechanical terms are equivalent to charges in electrical terms, this type of coupling can be represented only by a capacitative network. This representation is entirely analogous to the T network representation for a transformer.[17] The constants of the network can be evaluated in terms of the elastic constants of the crystal as follows: For a -18.5 degree cut crystal, we can write the stress strain equation (7) as

$$\begin{aligned} y_y &= s_{22}' Y_y + s_{23}' Z_z, \\ z_z &= s_{23}' Y_y + s_{33}' Z_z, \end{aligned} \tag{8}$$

since we are neglecting motion along the X axis and since s_{24}', the coupling coefficient of the shear to the Y' axis is zero. No Y_z force is assumed acting. If we work out the equation for the charges on the condensers of the equivalent representation shown in Fig. 33*B* we have, with the charges and voltages directed as shown

$$\begin{aligned} Q_1 &= \frac{e_y C_y}{1 - K^2} + e_z \frac{\sqrt{C_y C_z} K}{1 - K^2}, \\ Q_2 &= \frac{e_y \sqrt{C_y C_z} K}{1 - K^2} + \frac{e_z C_z}{1 - K^2}, \end{aligned} \tag{9}$$

where K the coupling factor between the two modes of motion, is defined by the relation

$$K = \frac{\sqrt{C_y C_z}}{C_m}. \tag{10}$$

Associating Q_1 with y_y, the displacement per unit length, Q_2 with

[16] See "Regular Combination of Acoustic Elements," W. P. Mason, *B. S. T. J.*, April, 1927, p. 258.

[17] See, for example, p. 281 in the book "Transmission Circuits for Telephonic Communication" by K. S. Johnson.

z_z, e_y with Y_y and e_z with Z_z, we have on comparing (9) with (8)

$$s_{22}' = \frac{C_y}{1 - K^2}\,;\; s_{23}' = \frac{\sqrt{C_y C_z}K}{1 - K^2}\,;\; s_{33}' = \frac{C_z}{1 - K^2} \tag{11}$$

or inversely

$$C_y = s_{22}'\left[1 - \frac{s_{23}'^2}{s_{22}' s_{33}'}\right];\; C_z = s_{33}'\left[1 - \frac{s_{23}'^2}{s_{22}' s_{33}'}\right];\; K = \frac{s_{23}'}{\sqrt{s_{22}' s_{33}'}}\cdot \tag{12}$$

If, now, alternating forces are applied to the crystal, another reaction to the applied force enters, namely the mass reaction of the crystal due to the inertia of the different parts. To take account of this reaction, the inductances are added to the two meshes representing mass reaction for the two modes of motion. To determine the value of the inductance, consider first the representation for one mode of motion shown by Fig. 33*A*. The resonant frequency of the system is given by

$$f_r = \frac{1}{2\pi\sqrt{LC}}\cdot \tag{13}$$

On the other hand, the resonant frequency of a bar is given by

$$f_r = \frac{1}{2l\sqrt{\rho s}}, \tag{14}$$

where l is the length of the bar, s its compliance, and ρ its density. But in the above representation the capacitance C is the compliance constant s so that, on comparing (13) and (14) we find

$$L = \frac{l^2\rho}{\pi^2}\cdot \tag{15}$$

In a similar manner for the coupled circuit, Fig. 33*B*, there results

$$L_y = \frac{l_y^2\rho}{\pi^2}\,;\; L_z = \frac{l_z^2\rho}{\pi^2}, \tag{16}$$

where l_y is the length of the crystal in centimeters along the y axis, and l_z the length of the crystal in centimeters along the z axis. Hence all of the constants of Fig. 33*B*, which represents the crystal for mechanical vibrations subject to the restrictions noted above, are determined and we should be able to predict all of the quantities which depend only on the mechanical constants of the crystal.

Of these the most important are the resonance frequencies of the crystal and their dependence on dimension, temperature coefficient and the like. To determine the natural mechanical resonance of a crystal,

we solve the network of Fig. 33*B* to find the frequencies of zero impedance for either an applied Y_y force or an applied Z_z force. The result is two frequencies f_1 and f_2 given by the coupled circuit equations

$$\begin{aligned} f_1^2 &= \tfrac{1}{2}[(f_A^2 + f_B^2) - \sqrt{(f_B^2 - f_A^2)^2 + 4K^2 f_A^2 f_B^2}], \\ f_2^2 &= \tfrac{1}{2}[(f_A^2 + f_B^2) + \sqrt{(f_B^2 - f_A^2)^2 + 4K^2 f_A^2 f_B^2}], \end{aligned} \tag{17}$$

where

$$f_A = \frac{1}{2\pi\sqrt{L_y C_y}} \text{ and } f_B = \frac{1}{2\pi\sqrt{L_z C_z}}. \tag{18}$$

Then f_A and f_B represent the natural frequencies along the Y and Z axis respectively when these two motions are not coupled together.

Two limiting cases of interest are obtainable from these relations. If f_B is much larger than f_A, the equations reduce to

$$\begin{aligned} f_1 &= f_A\sqrt{1 - K^2} = \frac{1}{2l_y\sqrt{\rho s_{22}'}}, \\ f_2 &= f_B = \frac{1}{2l_z\sqrt{\rho s_{33}'[1 - s_{23}'^2/s_{22}' s_{33}']}} \end{aligned} \tag{19}$$

upon substituting the value of the constants given before. The first equation shows that for a long thin rod the frequency depends on the elastic constant s_{22}', which is the inverse of Young's modulus. For the frequency f_2, which corresponds to that of a thin plate, a different elastic constant appears. Upon evaluating the expression $s_{33}'[1 - s_{23}'^2/s_{22}' s_{33}']$ in terms of the elastic constants which express the forces in terms of the strains—see equation (25)—we find that $s_{33}'(1 - s_{23}'^2/s_{22}' s_{33}') = 1/c_{33}$. c_{33} measures the ratio of force to strain when all the other coupling coefficients are set equal to zero, and corresponds to the frequency of one mode vibrating by itself without coupling to other modes. Hence the frequency of a thin plate should be

$$f = \frac{1}{2t}\sqrt{\frac{c_{nn}}{\rho}}, \tag{20}$$

where c_{nn} represents the elastic coefficient for the mode of motion considered, and t is the thickness of the plate. This deduction has been verified by experimental tests on thin plates.

Let us consider now the curves for the -18.5 degree cut crystal shown by Fig. 31. The values of the elastic constants for this case are

$$s_{22}' = 144 \times 10^{-14} \text{ cm.}^2/\text{dynes}; \; s_{23}' = -21.0 \times 10^{-14}; \; s_{33}' = 92.5 \times 10^{-14}. \tag{21}$$

Hence from equations (17), (18) and (21) one should be able to check the measured frequency curves of Fig. 31. The result is shown on the dotted lines of these curves. The agreement is quite good although a slightly better agreement would be obtained if s_{23} had a smaller value. Since these constants have never been measured with great accuracy, it is possible that they deviate somewhat from the curves of Fig. 30.

This theory can be applied also to a $\theta = +18.5$ degree cut crystal since the extensional coupling coefficient vanishes for this angle. The agreement is quite good if the frequency for the uncoupled mode given by the section on vibration in shear mode (page 446) is used in place of equation (14). The resonances for the $\theta = 0°$ cut crystal shown by Fig. 25 cannot be accounted for quantitatively by the simple theory given here since there are three modes of motion operating. The shear mode of motion is more closely coupled to the principal mode than is the Z_z extensional mode and hence a fair approximation is obtained by considering only the shear mode. However, for complete agreement the theory should be extended to a triply coupled circuit and that is not done in this paper.

Another phenomenon of interest which can be accounted for by the circuit of Fig. 33*B* is the temperature coefficient of the crystal and its variation with different ratios of axes and different angles of rotation. To obtain the relation, we assume that each of the vibrations may have a temperature coefficient of its own as may also the coefficient of coupling K. If a small change of temperature occurs, f_1 will change to $f_1(1 + T\Delta T)$, f_A to $f_A(1 + T_A\Delta T)$, f_B to $f_B(1 + T_B\Delta T)$ and K to $K(1 + T_K\Delta T)$. Assuming ΔT small so that its squares and higher powers can be neglected, we find from equation (17) that

$$T = \frac{1}{2f_1^2}\left[T_A f_A^2 \left[1 + \frac{f_B^2(1 - 2K^2) - f_A^2}{\sqrt{(f_B^2 - f_A^2)^2 + 4K^2 f_A^2 f_B^2}} \right] + T_B f_B^2 \times \left[1 - \frac{f_B^2(1 + 2K^2) - f_A^2}{\sqrt{(f_B^2 - f_A^2)^2 + 4K^2 f_A^2 f_B^2}} \right] - \frac{2T_K f_A^2 f_B^2}{\sqrt{(f_B^2 - f_A^2)^2 + 4K^2 f_A^2 f_B^2}} \right]. \quad (22)$$

The temperature coefficients of the six elastic constants have been measured at the Laboratories [18] by measuring the frequency temperature coefficients of variously oriented crystals. The temperature coefficients of the six elastic constants can be calculated from these

[18] These coefficients have been evaluated in cooperation with Messrs. F. R. Lack, G. W. Willard and I. E. Fair and their work is discussed in detail in their companion paper in this issue of the *B. S. T. J.*

measurements and have been found to be, in parts per million per degree centigrade:

$$T_{s_{11}} = +13;\ T_{s_{12}} = -1230;\ T_{s_{13}} = -347; \qquad (23)$$
$$T_{s_{14}} = +130;\ T_{s_{33}} = +213;\ T_{s_{44}} = +172.$$

Using these values and neglecting the extensional motion Z_z, the temperature coefficients calculated from equation (22) for a 0 degree cut crystal are shown on the dotted line of Fig. 5 and agree quite well with the measured values.

The Resonance Frequencies of a Crystal Vibrating in a Shear Mode

The equations of motion for any aelotropic body are

$$\rho\frac{\partial^2 u}{\partial t^2} = \frac{\partial X_x}{\partial x} + \frac{\partial X_y}{\partial y} + \frac{\partial X_z}{\partial z},$$
$$\rho\frac{\partial^2 v}{\partial t^2} = \frac{\partial Y_x}{\partial x} + \frac{\partial Y_y}{\partial y} + \frac{\partial Y_z}{\partial z}, \qquad (24)$$
$$\rho\frac{\partial^2 w}{\partial t^2} = \frac{\partial Z_x}{\partial x} + \frac{\partial Z_y}{\partial y} + \frac{\partial Z_z}{\partial z},$$

where u, v, w are the displacements of any point in the crystal along the x, y, z axes respectively and X_x, etc. are the six applied stresses. The strains have been expressed in terms of the stresses by equation (7). It is more advantageous for the present purpose to express the stresses in terms of the strains, which can be done by the following equations:

$$X_x = c_{11}x_x + c_{12}y_y + c_{13}z_z + c_{14}y_z,$$
$$Y_y = c_{12}x_x + c_{22}y_y + c_{23}z_z + c_{24}y_z,$$
$$Z_z = c_{13}x_x + c_{23}y_y + c_{33}z_z, \qquad (25)$$
$$Y_z = c_{14}x_x + c_{24}y_y + c_{44}y_z,$$
$$Z_x = c_{44}z_x + c_{14}x_y,$$
$$X_y = c_{14}z_x + \tfrac{1}{2}(c_{11} - c_{12})x_y,$$

where the c's are the elastic constants and the strains x_x, etc., are given in terms of the displacements u, v, w by the equations

$$x_x = \frac{\partial u}{\partial x};\ y_y = \frac{\partial v}{\partial y};\ z_z = \frac{\partial w}{\partial z};\ y_z = \left(\frac{\partial v}{\partial z} + \frac{\partial w}{\partial y}\right);$$
$$z_x = \left(\frac{\partial w}{\partial x} + \frac{\partial u}{\partial z}\right);\ x_y = \left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right). \qquad (26)$$

In equation (24) there exist the reciprocal relations

$$X_y = Y_x;\ X_z = Z_x;\ Y_z = Z_y. \qquad (27)$$

For a free edge, i.e. no resulting forces being applied to the crystal, the conditions existing for every point of the boundaries are

$$\begin{aligned} X_\nu &= X_x \cos(\nu,x) + X_y \cos(\nu,y) + X_z \cos(\nu,z) = 0, \\ Y_\nu &= Y_x \cos(\nu,x) + Y_y \cos(\nu,y) + Y_z \cos(\nu,z) = 0, \\ Z_\nu &= Z_x \cos(\nu,x) + Z_y \cos(\nu,y) + Z_z \cos(\nu,z) = 0, \end{aligned} \tag{28}$$

where ν is the normal to the boundary under consideration.

If these equations are combined and completely solved, the motion of a quartz crystal is completely determined. The results which were obtained above in an approximate manner could be rigorously solved. However, on account of the difficulty [19] of the solution, this is not attempted here. In the present section it is simply the purpose to find out what resonances a crystal will have if it is vibrating in a shear mode only. To avoid setting up motion in the other modes of vibration, the coupling elasticities c_{14}, c_{24}, c_{34} are assumed zero. Similarly if c_{12}, c_{13}, c_{23} were set equal to zero we should have the possibility of three extensional modes and one shear mode vibrating simultaneously with no reaction on one another, and the equation of motion would be

$$\begin{aligned} \rho \frac{\partial^2 u}{\partial t^2} &= \frac{\partial}{\partial x}(c_{11}x_x), \\ \rho \frac{\partial^2 v}{\partial t^2} &= \frac{\partial}{\partial y}(c_{22}y_y) + \frac{\partial}{\partial_z}[c_{44}y_z], \\ \rho \frac{\partial^2 w}{\partial t^2} &= \frac{\partial}{\partial y}(c_{44}y_z) + \frac{\partial}{\partial z}[c_{33}z_z]. \end{aligned} \tag{29}$$

The displacements u, v, and w would be the sum of the displacements caused by the four motions. To find the displacements and resonances caused by the shear mode y_z, we neglect the other modes and have the equations

$$\begin{aligned} \rho \frac{\partial^2 v}{\partial t^2} &= c_{44} \frac{\partial}{\partial z}(y_z), \\ \rho \frac{\partial^2 w}{\partial t^2} &= c_{44} \frac{\partial}{\partial y}(y_z). \end{aligned} \tag{30}$$

Differentiating the first of equations (30) by $\frac{\partial}{\partial z}$, and the second by $\frac{\partial}{\partial y}$, and adding, there results,

$$\rho \frac{\partial^2}{\partial t^2}\left(\frac{\partial v}{\partial z} + \frac{\partial w}{\partial y}\right) = c_{44}\left[\frac{\partial^2 y_2}{\partial y^2} + \frac{\partial^2 y_2}{\partial z^2}\right]. \tag{31}$$

[19] For example if motion is limited to the y and z directions, and the coefficient c_{14} is set equal to zero, the equations reduce approximately to those for a plate bent in flexure, and this case has never been solved for the boundary condition of interest here, namely all four edges being free to move—see Rayleigh "Theory of Sound," page 372, Vol. I, 1923 edition.

Since $\frac{\partial v}{\partial z} + \frac{\partial w}{\partial y} = y_z$, this reduces to

$$\frac{\partial^2 y_z}{\partial t^2} = c^2 \left[\frac{\partial^2 y_z}{\partial y^2} + \frac{\partial^2 y_z}{\partial z^2} \right], \tag{32}$$

where $c^2 = c_{44}/\rho$.

For a simple harmonic vibration, of frequency f, the equation reduces to

$$\left[\frac{\partial^2 y_z}{\partial y^2} + \frac{\partial^2 y_z}{\partial z^2} \right] + \frac{p^2}{c^2} y_z = 0, \tag{33}$$

where $p = 2\pi f$. The solution of this equation consistent with the boundary conditions (28) is

$$y_z = \Sigma\Sigma A_{mn} \left[\sin \frac{m\pi y}{a} \sin \frac{n\pi z}{b} \right] \cos pt, \tag{34}$$

where a is the length of the crystal in the y direction, b the length of the crystal in the z direction, and m and n are integers. Substituting this equation in the equation (32), we find that it is a solution provided

$$\left[\frac{m^2\pi^2}{a^2} + \frac{n^2\pi^2}{b^2} \right] = \frac{p^2}{c^2} = \frac{(2\pi f)^2}{c^2}. \tag{35}$$

Hence the resonant frequencies of the crystal in shear vibration are

$$f = \frac{c}{2} \sqrt{\frac{m^2}{a^2} + \frac{n^2}{b^2}}. \tag{36}$$

To find the shape of the deformed crystal, we have from (30) for simple harmonic vibration

$$v = -\frac{c^2}{p^2} \frac{\partial y_z}{\partial z} = \frac{-a^2b^2}{m^2\pi^2b^2 + n^2\pi^2a^2} \frac{\partial y_z}{\partial z}, \tag{37}$$

$$w = -\frac{c^2}{p^2} \frac{\partial y_z}{\partial y} = \frac{-a^2b^2}{m^2\pi^2b^2 + n^2\pi^2a^2} \frac{\partial y_z}{\partial y}. \tag{38}$$

The cases $m = 0$, $n = 1$ and $m = 1$, $n = 0$ require a stress known as a simple shear to excite them, whereas the stress applied by the piezo-electric effect is a pure shear. Hence the case $m = 1$, $n = 1$ provides the lowest frequency solution. The displacements v and w for this case are by equations (34), (37) and (38)

$$\begin{aligned} v &= \frac{-a^2b\pi}{\pi^2a^2 + \pi^2b^2} \left(\sin \frac{\pi y}{a} \cos \frac{\pi z}{b} \right); \\ w &= -\frac{ab^2\pi}{\pi^2a^2 + \pi^2b^2} \left(\cos \frac{\pi y}{a} \sin \frac{\pi z}{b} \right). \end{aligned} \tag{39}$$

The resulting distortion of the crystal is shown by Fig. 34.

We can conclude, therefore, that the solution for a shear vibration in a quartz crystal will be given by equation (34). It is obvious from Fig. 34 that the shear vibration will have a strong coupling to a bar

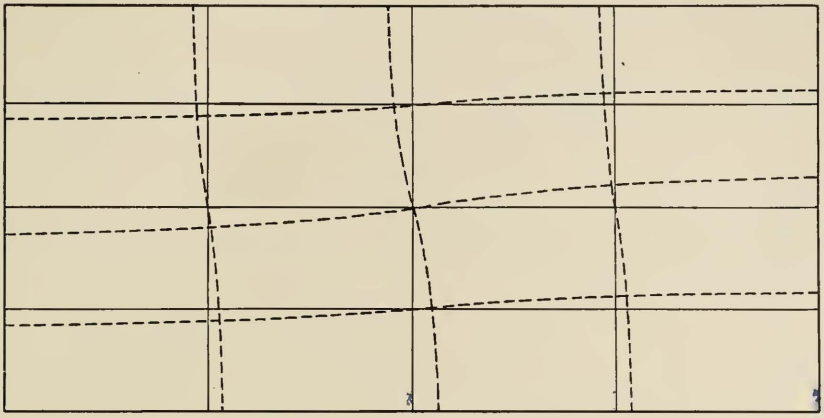


Fig. 34—Form of crystal in shear vibration.

bent in its second mode of flexure, since the form of the bent bar, as shown by Fig. 29, is very closely the same as a given displacement line in the crystal vibrating in shear. Little coupling should exist between the shear mode and a bar in its first flexure mode, since this mode of flexure requires a displacement which is symmetrical on both sides of the central line whereas the bar vibrating in shear has a motion in which the displacement on one side of the center line is the opposite of the displacement on the other side of the center line.

The Elastic and Piezo-Electric Constants of Quartz for Rotated Crystals [20]

W. Voigt [21] gives for the stress strain and piezo-electric relation in a quartz crystal, for the three extensions and one shear found above,

$$\begin{aligned} -x_x &= s_{11}X_x + s_{12}Y_y + s_{13}Z_z + s_{14}Y_z, \\ -y_y &= s_{12}X_x + s_{22}Y_y + s_{23}Z_z + s_{24}Y_z, \\ -z_z &= s_{13}X_x + s_{23}Y_y + s_{33}Z_z + s_{34}Y_z, \\ -y_z &= s_{14}X_x + s_{24}Y_y + s_{34}Z_z + s_{44}Y_z, \end{aligned} \tag{40}$$

$$-P_x = d_{11}X_x + d_{12}Y_y + d_{13}Z_z + d_{14}Y_z, \tag{41}$$

[20] The material of this section was first derived by Mr. R. A. Sykes of the Bell Telephone Laboratories.

[21] W. Voigt, Lehrbach Der Kristallphysik.

where

x_x, y_y, z_z = extensional strains = elongation per unit length,
X_x, Y_y, Z_z = extensional stresses = force per unit area,
y_z = shearing strain = cos of an angle,
Y_z = shearing stress = force per unit area,
s_{ij} = elastic compliances = displacement per dyne,
d_{ij} = piezo-electric constants = e.s.u. charge per dyne,
P_x = piezo-electric polarization = charge per unit area.

The best measured values for these constants when the X axis coincides with the electric axis of the crystal, the Y axis with the mechanical axis and the Z axis with the optical axis, are

$$\begin{aligned}
s_{11} &= s_{22} = 127.2 \times 10^{-14} \text{ cm.}^2/\text{dyne}, \\
s_{12} &= -16.6 \times 10^{-14} \text{ cm.}^2/\text{dyne}, \\
s_{13} &= s_{23} = -15.2 \times 10^{-14} \text{ cm.}^2/\text{dyne}, \\
s_{24} &= -s_{14} = 43.1 \times 10^{-14} \text{ cm.}^2/\text{dyne}, \\
s_{33} &= 97.2 \times 10^{-14} \text{ cm.}^2/\text{dyne}, \\
s_{34} &= 0, \\
s_{44} &= 200.5 \times 10^{-14} \text{ cm.}^2/\text{dyne}, \\
d_{11} &= -d_{12} = -6.36 \times 10^{-8} \frac{\text{e.s.u. charge}}{\text{dyne}}, \\
d_{13} &= 0 \\
d_{14} &= 1.69 \times 10^{-8} \frac{\text{e.s.u. charge}}{\text{dyne}}.
\end{aligned} \tag{42}$$

If, now, we maintain the direction of the electrical axis but rotate the direction of the principal axis by some angle θ, the resulting constants of equations (40) and (41) undergo a change.

Let the direction cosines for the new axes be given by

	x	y	z
x'	l_1	m_1	n_1
y'	l_2	m_2	n_2
z'	l_3	m_3	n_3

(43)

The convention is adopted that a positive angle θ is a clockwise rotation of the principal axis of the crystal, when the electrically positive face (determined by a squeeze) is up. For a left-handed crystal a positive angle is in a counter clockwise direction. θ is the angle between the previously unprimed and the primed axes.

If we transform only the y and z axes, there results

$$\begin{aligned} l_2 &= l_3 = m_1 = n_1 = 0, \\ l_1 &= 1, \\ m_2 &= n_3 = \cos\theta, \\ -n_2 &= m_3 = \sin\theta. \end{aligned} \tag{44}$$

Love [22] gives the transformation for the stress and the strain functions as

$$\begin{aligned} x_x' &= x_x l_1^2 + y_y m_1^2 + z_z n_1^2 + y_z m_1 n_1, \\ y_y' &= x_x l_2^2 + y_y m_2^2 + z_z n_2^2 + y_z m_2 n_2, \\ z_x' &= x_x l_3^2 + y_y m_3^2 + z_z n_3^2 + y_z m_3 n_3, \\ y_z' &= 2x_x l_2 l_3 + 2y_y m_2 m_3 + 2z_z n_2 n_3 + y_z(m_2 n_3 + m_3 n_2), \end{aligned} \tag{45}$$

$$\begin{aligned} X_x' &= X_x l_1^2 + Y_y m_1^2 + Z_z n_1^2 + Y_z 2m_1 n_1, \\ Y_y' &= X_x l_2^2 + Y_y m_2^2 + Z_z n_2^2 + Y_z 2m_2 n_2, \\ Z_z' &= X_x l_3^2 + Y_y m_3^2 + Z_z n_3^2 + Y_z 2m_3 n_3, \\ Y_z' &= X_x l_2 l_3 + Y_y m_2 m_3 + Z_z n_2 n_3 + Y_z(m_2 n_3 + m_3 n_2). \end{aligned} \tag{46}$$

Substituting (44) in (45) and (46) and then expressing x_x, y_y . . . X_x, Y_y . . ., etc., in terms of x_x', y_y' . . . X_x', Y_y' . . ., etc., we may substitute these values in equations (40) and (41) to give the stress-strain and polarization in a crystal whose rectangular axes do not coincide with the real optical and mechanical axis. Performing the above operations, a new set of constants s_{ij}', are obtained which are functions of θ, namely:

$$s_{11}' = s_{11},$$

$$s_{12}' = \tfrac{1}{2}[s_{12} + s_{13} + (s_{12} - s_{13})\cos 2\theta - s_{14}\sin 2\theta],$$

$$s_{13}' = \tfrac{1}{2}[s_{13} + s_{12} + (s_{13} - s_{12})\cos 2\theta + s_{14}\sin 2\theta],$$

$$s_{14}' = s_{14}\cos 2\theta + (s_{12} - s_{13})\sin 2\theta,$$

$$s_{22}' = s_{11}\cos^4\theta + s_{33}\sin^4\theta + 2s_{14}\cos^3\theta\sin\theta + (2s_{13} + s_{44})\sin^2\theta\cos^2\theta,$$

$$s_{23}' = s_{13}(\cos^4\theta + \sin^4\theta) + s_{14}(\sin^3\theta\cos\theta - \cos^3\theta\sin\theta) + (s_{11} + s_{33} - s_{44})\sin^2\theta\cos^2\theta,$$

$$s_{24}' = -s_{14}(\cos^4\theta - 3\sin^2\theta\cos^2\theta) + (2s_{11} - 2s_{13} - s_{44})\cos^3\theta\sin\theta + (2s_{13} - 2s_{33} + s_{44})\sin^3\theta\cos\theta,$$

$$s_{33}' = s_{33}\cos^4\theta + s_{11}\sin^4\theta - 2s_{14}\sin^3\theta\cos\theta + (2s_{13} + s_{44})\sin^2\theta\cos^2\theta,$$

[22] "The Mathematical Theory of Elasticity," Cambridge University Press, pp. 42 and 78.

$$s_{34}' = s_{14}(\sin^4\theta - 3\sin^2\theta\cos^2\theta) + (2s_{11} - 2s_{13} - s_{44})\sin^3\theta\cos\theta + (2s_{13} - 2s_{33} + s_{44})\cos^3\theta\sin\theta,$$

$$s_{44}' = (4s_{33} + 4s_{11} - 8s_{13} - 2s_{44})\sin^2\theta\cos^2\theta + 4s_{14} \times (\sin^3\theta\cos\theta - \sin\theta\cos^3\theta) + s_{44}(\sin^4\theta + \cos^4\theta)$$

and

$$d_{11}' = d_{11},$$
$$d_{12}' = -\tfrac{1}{2}[d_{11}(1 + \cos 2\theta) + d_{14}\sin 2\theta],$$
$$d_{13}' = -\tfrac{1}{2}[d_{11}(1 - \cos 2\theta) - d_{14}\sin 2\theta],$$
$$d_{14}' = d_{14}\cos 2\theta - d_{11}\sin 2\theta.$$

The curves representing the s' values for varying angles of orientation are plotted on Fig. 30 while the values of d' are plotted on Fig. 35.

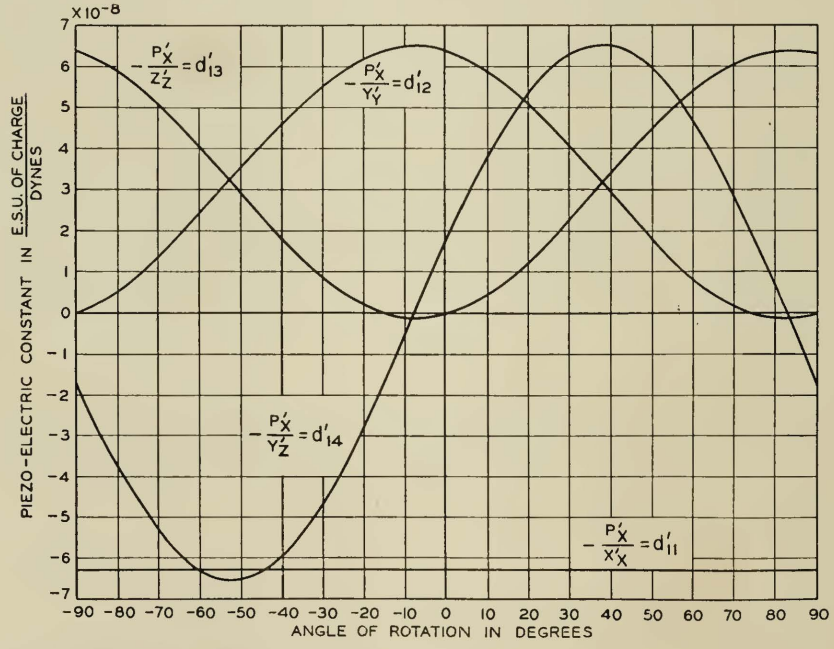


Fig. 35—Piezo-electric constants of a perpendicularly cut crystal as a function of the angle of rotation.

Some Improvements in Quartz Crystal Circuit Elements

By F. R. LACK, G. W. WILLARD, and I. E. FAIR

The characteristics of the *Y*-cut quartz crystal plate are discussed. It is shown that by rotating a plate about the *X* axis special orientations are found for which the frequency spectrum is simplified, the temperature coefficient of frequency is reduced practically to zero and the amount of power that can be controlled without fracture of the crystal is increased. These improvements are obtained without sacrificing the advantages of the Y cut plate, i.e., activity and the possibility of rigid clamping in the holder.

THERE are at the present time two types of crystal quartz plates in general use as circuit elements for frequency stabilization at radio frequencies, namely, the *X*-cut and *Y*-cut.[1] This paper is concerned with the improved characteristics of plates having radically new orientations.

In its usual form the *Y*-cut plate is cut from the mother crystal, as shown in Fig. 1. The electric field is applied along the *Y* direction

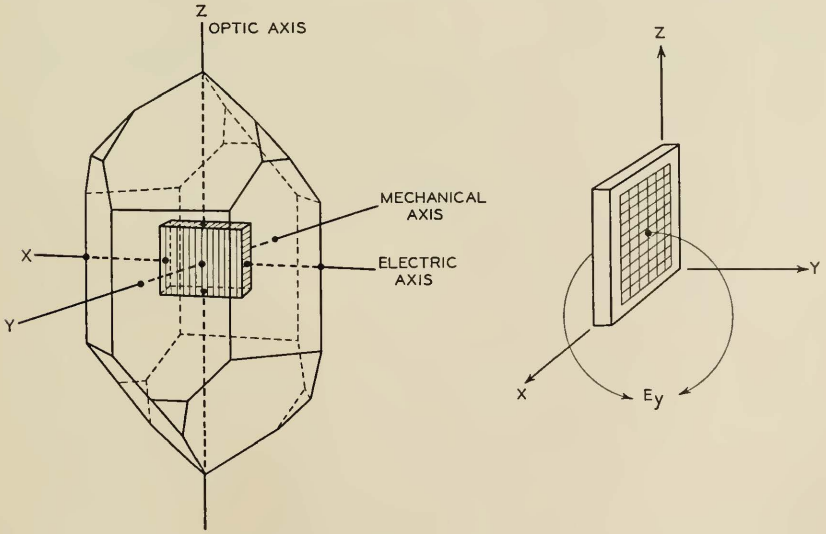


Fig. 1—Showing relation of Y-cut quartz crystal to the crystallographic axes.

and for high frequencies an x_y shear vibration is utilized.

The frequency of such a vibration is given by the expression:

$$f = \frac{1}{2l}\sqrt{\frac{c_{66}}{\rho}}, \qquad (1)$$

[1] "Piezo-Electric Terminology," W. G. Cady, *Proc. I. R. E.*, 1930, p. 2136.

where c_{66} = the elastic constant for quartz connecting the X_y stress with an x_y strain = 39.1×10^{10} dynes per cm.[2]

ρ = the density of quartz = 2.65 gms. per cm.[3]

l = the thickness in cm.

On substituting the numerical values in equation (1), a frequency-thickness constant of 192 kc. per cm. is obtained which checks within 3 per cent the value of this constant found by experiment.

This x_y shear vibration is not appreciably affected when the plate is rigidly clamped, the clamping being applied either around the periphery if the plate is circular, or at the corners if square. Hence a mechanically rigid holder arrangement is possible which is particularly suitable for mobile radio applications.[2]

The temperature coefficient of frequency of this vibration is approximately + 85 parts/million/C.°, which means that for most applications it must be used in a thermostatically controlled oven. In operation, this comparatively large temperature coefficient is responsible for a major part of any frequency deviations from the assigned value.

Another important characteristic of the *Y*-cut crystal is the secondary frequency spectrum of the plate. This secondary spectrum consists of overtones of low frequency vibrations which are mechanically coupled to the desired vibration and cause discontinuities in the characteristic frequency-temperature and frequency-thickness curves of the crystal. In some instances these coupled secondary vibrations can be utilized to produce a low temperature coefficient over a limited temperature range.[3] But in general, at the higher frequencies (above one megacycle) this secondary spectrum is a source of considerable annoyance, not only in the initial preparation of a plate for a given frequency but in the added necessity for some form of temperature control. In practice, these plates are so adjusted that there are no discontinuities in the frequency-temperature characteristic in the region where they are expected to operate, but at high frequencies it is difficult to eliminate all of these discontinuities over a wide temperature range. If, then, for any reason the crystal must be operated without the temperature control, a frequency discontinuity with temperature may cause a large frequency shift greatly in excess of that to be expected from the normal temperature coefficient.

From the above considerations it may be concluded that the standard *Y*-cut plate has two distinct disadvantages: namely, a

[2] U. S. Patent No. 1883111, G. M. Thurston, Oct. 18, 1932. "Application of Quartz Plates to Radio Transmitters," O. M. Hovgaard, *Proc. I. R. E.*, 1932, p. 767.

[3] "Observations on Modes of Vibrations and Temperature Coefficients of Quartz Plates," F. R. Lack, *Proc. I. R. E.*, 1929, p. 1123; *Bell Sys. Tech. Jour.*, July, 1929.

temperature coefficient requiring close temperature regulation and a troublesome secondary frequency spectrum. Assuming that the temperature coefficient of the desired frequency could be materially reduced, the effect of any secondary spectrum must also be minimized before temperature regulation can be abandoned. In fact, from the standpoint of satisfactory production and operation of these crystal plates, it is perhaps more important that the secondary spectrum be eliminated than that the temperature coefficient be reduced.

The Secondary Spectrum

The secondary spectrum of these plates, as has been indicated above, is caused by vibrations of the same or of other types than the wanted vibration taking place in other directions of the plate and coupled to the wanted vibration mechanically. This condition of affairs exists in all mechanical vibrating systems but is complicated in the case of quartz by the complex nature of the elastic system involved.

Considering specifically the case of the Y-cut plate the desired vibration is set up through the medium of an x_y strain. Hence any coupled secondary vibrations must be set in motion through this x_y strain. Referring to the following elastic equations for quartz (in these equations X, Y and Z are directions coincident with the crystallographic axes; see Fig. 1 and Appendix),

$$\left.\begin{aligned}
-X_x &= c_{11}x_x + c_{12}y_y + c_{13}z_z + c_{14}y_z \\
-Y_y &= c_{12}x_x + c_{22}y_y + c_{23}z_z + c_{24}y_z \\
-Z_z &= c_{13}x_x + c_{23}y_y + c_{33}z_z \\
-Y_z &= c_{14}x_x + c_{24}y_y + c_{44}y_z \\
-Z_x &= + c_{55}z_x + c_{56}x_y \\
-X_y &= + c_{56}z_x + c_{66}x_y
\end{aligned}\right\} \qquad (2)$$

it will be seen that by reason of the constant c_{56} an x_y strain will set up a stress in the Z_x plane which in turn will produce a z_x strain. Hence the x_y and z_x strains are coupled together mechanically, the value of the constant c_{56} being a measure of that coupling.

High order overtones of vibrations resulting from this z_x strain constitute the major part of the secondary frequency spectrum of these plates.

The technique for dealing with this secondary spectrum in the past has been the proper choice of dimensions. At high frequencies these overtones occur very close together and when one set is moved out of the range by grinding a given dimension another set will appear. Some benefit is obtained with the clamped holder, which tends to inhibit certain types of transverse vibrations; but as indicated above,

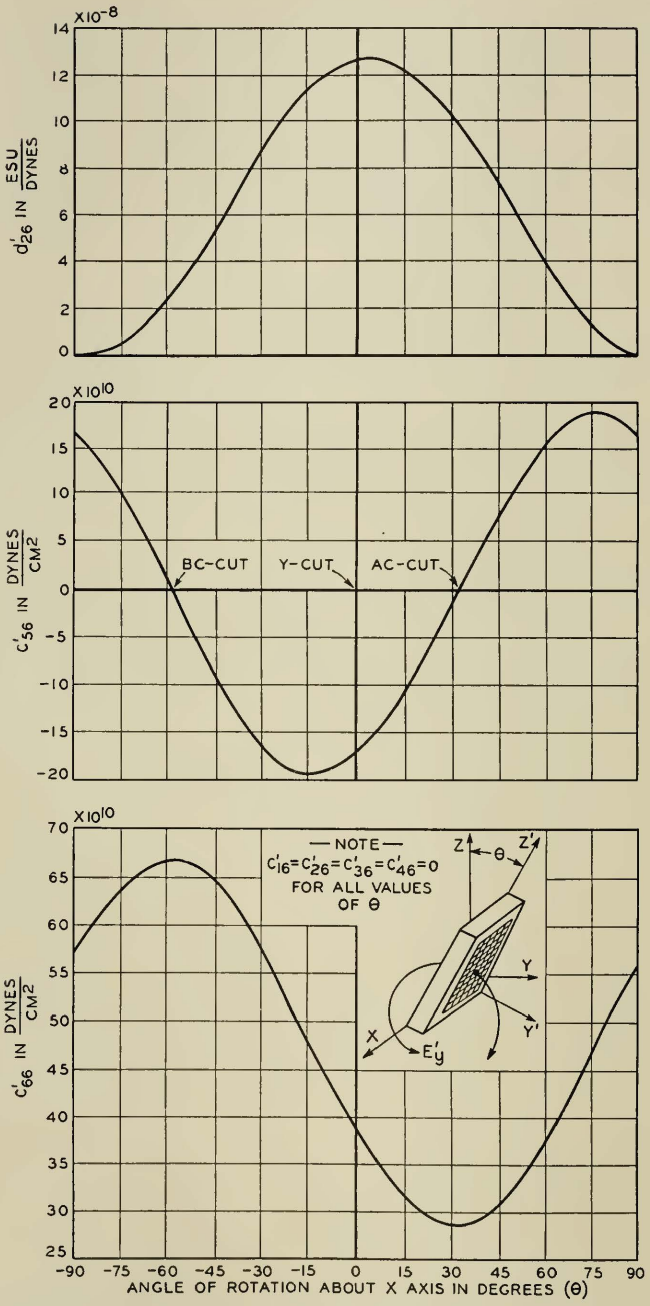


Fig. 2—c_{66}', c_{56}', and d_{26}' as a function of rotation about the X axis.

the elimination of the effects of the coupled secondary frequency spectrum over a wide temperature range is a difficult matter.

Another method has been developed recently for dealing with these coupled vibrations.[4] This consists of reducing, by a change in orientation, the magnitude of the elastic constant responsible for the coupling. If the orientation of the crystal plate be shifted with respect to the crystallographic axes then, in general, the elastic constants with reference to the axes of the plate will vary. The direct constants (c_{11}', c_{22}', $\cdots$) which represent the longitudinal and shear moduli will of course vary in magnitude only, while the cross constants (c_{ij}', $\cdots$) will vary both as to magnitude and sign. There is a possibility therefore that the proper choice of orientation of the plate will reduce c_{56}' to zero without at the same time introducing other couplings.

Figure 2 shows graphically the variation of c_{66}' and c_{56}' as a function of rotation about the X axis. These have been calculated by means of the equations given in the appendix. It will be seen that at approximately $+ 31°$ and $- 60°$ c_{56}' becomes zero. Here then are two orientations for which the coupling between the x_y and z_x strains should be zero.

In shifting the orientation of the plate the necessity for exciting the wanted vibration piezo-electrically must not be lost sight of. Hence in addition to computing the values of the elastic constants the variation of the piezo-electric moduli as a function of orientation must also be examined. Figure 2 also shows the effect of rotation about the X axis on d_{26}' (the constant connecting the E_y' electric field with the x_y' strain). It will be seen that at both $+ 31°$ and $- 60°$ the x_y' vibration can be excited piezo-electrically but it is to be expected that a plate cut at $- 60°$ will be relatively inactive,[5] for d_{26}' at this point is only 20 per cent of its value for the Y-cut plate. On the other hand at $+ 31°$ a plate would be practically equivalent to the Y-cut as far as activity is concerned. The frequency of the x_y' vibration for these special orientations can be calculated by means of equation (1) substituting for c_{66} the value of c_{66}' for the given angle as read from the curve of Fig. 2.

[4] The expression of the coupling between two modes of vibration in quartz in terms of the elastic constants was first suggested in 1930 by Mr. W. P. Mason of the Bell Telephone Laboratories.

[5] The word "activity" is a rather loose term used by experimenters in this field to describe the ease with which a given vibration can be excited in a particular circuit. It is often spoken of in terms of the grid current that is obtained in that circuit or the amount of feedback necessary to produce oscillation. It can better be expressed quantitatively as the coupling between the electrical and mechanical systems (not to be confused with the mechanical coupling between different vibrations described above) which is a simple function of the piezo-electric and elastic moduli of the vibration involved and the dielectric constant of the crystal plate.

For the purpose of identification the plate cut at + 31° has been designated as the AC-cut and the plate cut at − 60° the BC-cut. Crystal plates having these orientations have been made up and tested. It is evident from the frequency-temperature and frequency-thickness characteristics of both cuts that a simplification of the frequency spectrum results from the reduction in coupling to secondary

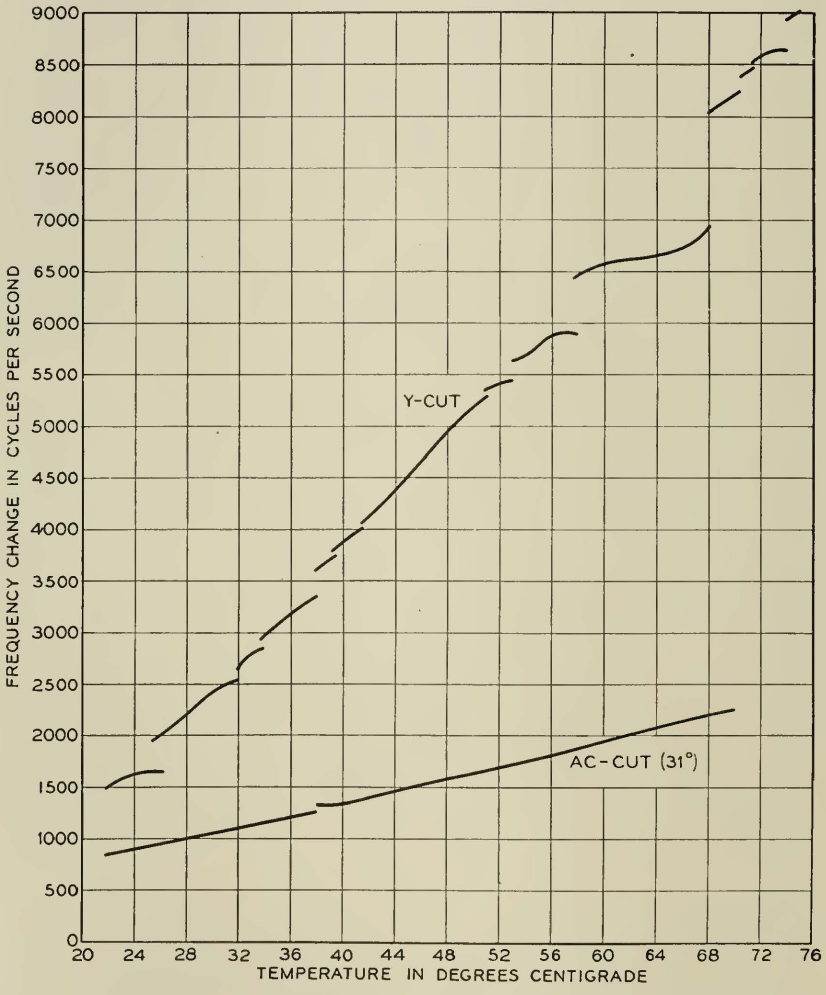


Fig. 3—Frequency-temperature characteristics of AC-cut and Y-cut plates of same frequency and area.

Frequency 1600 KC.
Dimensions:

Y-cut	y = 1.22 mm.	x = 38 mm.	z = 38 mm.
AC-cut (31°)	y' = 1.00 mm.	x = 38 mm.	z' = 38 mm.

vibrations. Frequency discontinuities of the order of a kilocycle or more which are a common occurrence with the Y-cut plate have disappeared and frequency-temperature curves that are linear over a considerable temperature range can be obtained without much difficulty. This is illustrated by Fig. 3 which shows frequency-temperature characteristics for both AC-cut and Y-cut plates of the same frequency and area. The AC-cut plate can be clamped to the same extent as the Y-cut plate.

. There is still some coupling remaining to certain secondary frequencies. These frequencies are difficult to identify but are thought to be caused by overtones of flexural vibrations set up by the x_y' shear itself and hence would be unaffected by the reduction of c_{56}'. These remaining frequencies do not cause much difficulty above 500 kc. For the AC-cut ($+ 31°$) plate the temperature coefficient of frequency is $+ 20$ cycles/million/C.°, while for the BC-cut ($- 60°$) plate it is $- 20$ cycles/million/C.°.

In addition to these calculations for the x_y' vibration in plates rotated about the X axis, a detailed study has been made of other types of vibration and rotation about the other axes. For high frequencies nothing has been found to compare with the reduction in complexity of frequency spectrum obtainable with these two orientations.

Temperature Coefficients

This study has produced in the AC-cut a new type of plate which has superior characteristics to the standard Y-cut: i.e., a simplified frequency spectrum and a lower temperature coefficient. The values of the temperature coefficients obtained for these new orientations are significant and suggest that perhaps other orientations can be chosen for which the temperature coefficient will be zero. With the measured values of the temperature coefficients for the different orientations and the c_{66}' equation (Appendix) it is possible to compute the temperature coefficient for any angle. Figure 4 shows graphically the results of such a computation for an x_y' vibration as a function of rotation about the X axis. It will be seen that at approximately $+ 35°$ and $- 49°$ the x_y' vibration will have a zero temperature coefficient of frequency.

This curve has been checked experimentally, the check points being indicated on the curve. Concentrating on a plate cut at $+ 35°$, which has been designated the AT-cut, it will be seen that this type of plate offers considerable possibilities. Figure 5 shows the frequency-temperature curves of a 2-megacycle AT-cut plate and a

standard Y-cut of the same frequency and area. These curves not only illustrate the reduction in temperature coefficient but also show that in the AT-cut plate the secondary frequency spectrum has been eliminated over the temperature range of the test. This is to be expected, for 35° is close to the 31° zero coupling point; hence such coupling as does exist is small in magnitude.

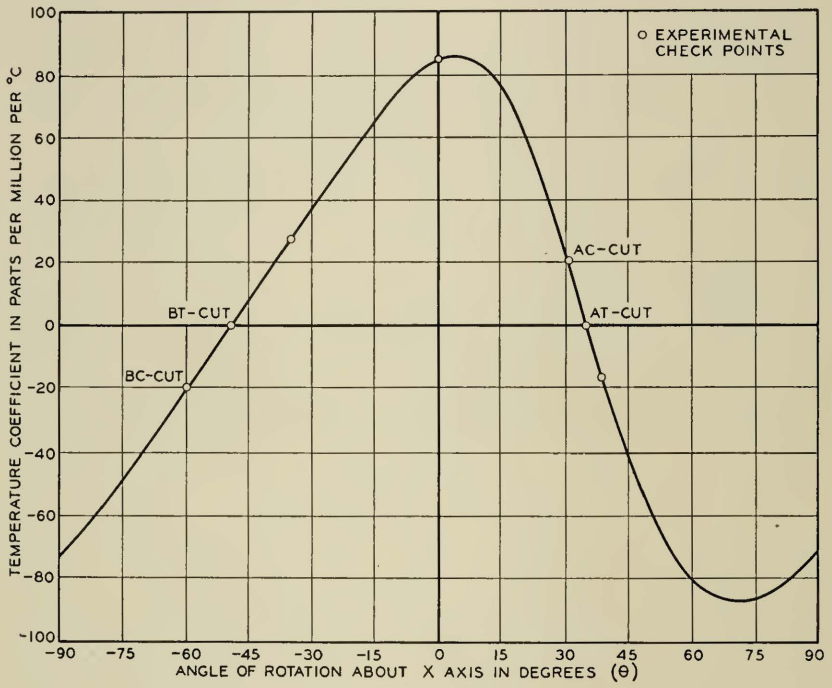


Fig. 4—Temperature coefficient of frequency of the vibration depending upon c_{66}' as a function of rotation about the X axis.

These AT-cut plates can be produced with a sufficiently low temperature coefficient so that for most applications the temperature regulating system can be discarded, and in addition a simplification of the secondary frequency spectrum is obtained. Furthermore, the advantages of the Y-cut plate, i.e., clamping and activity, have not been sacrificed.

Additional tests on AT-cut plates indicate that it will be possible to use them to control reasonable amounts of power without danger of fracture. At 2 megacycles, 50-watt crystal oscillators would appear to be practical and in some experimental circuits the power output has been run up to 200 watts without fracturing the crystal. The

explanation for this lies in the fact that the reduction in magnitude of the coupling to transverse vibrations has reduced the transverse stresses which in the *Y*-cut plate are responsible for the fractures.

Experimental crystals of this type have been produced in the frequency range from 500 kc. to 20 megacycles. The possibility of high frequencies, together with the elimination of the temperature

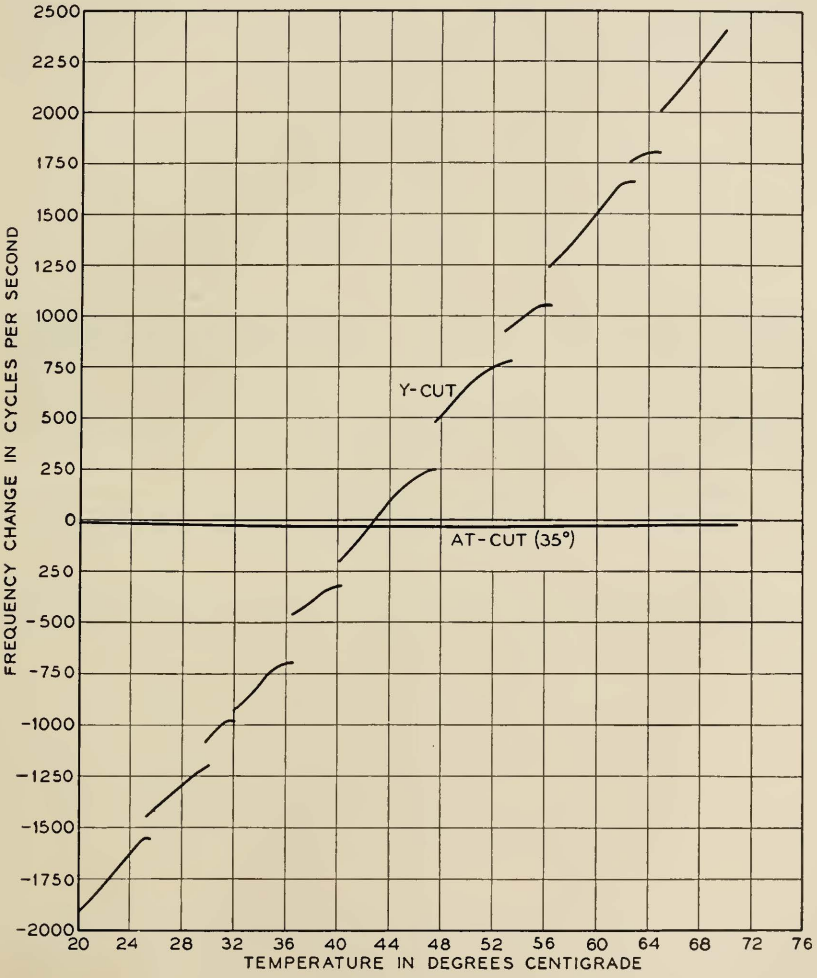


Fig. 5—Frequency-temperature characteristics of AT-cut and Y-cut plates of same frequency and area.

Frequency 1000 KC.
Dimensions:

Y-cut	y = 1.970 mm.	x = 38 mm.	z = 38 mm.
AT-cut (35°)	y' = 1.675 mm.	x = 38 mm.	z' = 38 mm.

control and the increase in the amount of power that may be controlled, should result in a considerable simplification of future short wave radio equipment.

APPENDIX

Elastic Equations

The general elastic equations for any crystal are given below, X', Y' and Z' representing any orthogonal set of axes.

$$\left.\begin{aligned}
-X_x' &= c_{11}'x_x' + c_{12}'y_y' + c_{13}'z_z' + c_{14}'y_z' + c_{15}'z_x' + c_{16}'x_y' \\
-Y_y' &= c_{12}'x_x' + c_{22}'y_y' + c_{23}'z_z' + c_{24}'y_z' + c_{25}'z_x' + c_{26}'x_y' \\
-Z_z' &= c_{13}'x_x' + c_{23}'y_y' + c_{33}'z_z' + c_{34}'y_z' + c_{35}'z_x' + c_{36}'x_y' \\
-Y_z' &= c_{14}'x_x' + c_{24}'y_y' + c_{34}'z_z' + c_{44}'y_z' + c_{45}'z_x' + c_{46}'x_y' \\
-Z_x' &= c_{15}'x_x' + c_{25}'y_y' + c_{35}'z_z' + c_{45}'y_z' + c_{55}'z_x' + c_{56}'x_y' \\
-X_y' &= c_{16}'x_x' + c_{26}'y_y' + c_{36}'z_z' + c_{46}'y_z' + c_{56}'z_x' + c_{66}'x_y'
\end{aligned}\right\} \quad (3)$$

When in quartz X', Y' and Z' coincide with the crystallographic axes of the material (X the electric axis, Y the mechanical axis, and Z the optic axis), equation (3) reduces to equation (2) of the text. In addition the following relations exist between the constants of equation (2) because of conditions of symmetry

$$c_{11} = c_{22}, \qquad c_{44} = c_{55}, \qquad c_{66} = (c_{11} - c_{12})/2, \qquad c_{13} = c_{23}$$
$$c_{14} = -c_{23} = c_{56}.$$

The numerical values of these constants have been determined experimentally by Voigt [6] and others.

$$\begin{aligned}
c_{11} &= 85.1 \times 10^{10} \frac{\text{dy.}}{\text{cm.}^2} & c_{12} &= 6.95 \times 10^{10} \frac{\text{dy.}}{\text{cm.}^2}, \\
c_{33} &= 105.3 \times 10^{10} \frac{\text{dy.}}{\text{cm.}^2} & c_{13} &= 14.1 \times 10^{10} \frac{\text{dy.}}{\text{cm.}^2}, \\
c_{44} &= 57.1 \times 10^{10} \frac{\text{dy.}}{\text{cm.}^2} & c_{14} &= 16.8 \times 10^{10} \frac{\text{dy.}}{\text{cm.}^2}, \\
c_{66} &= 39.1.
\end{aligned}$$

Using these constants it is possible to calculate the c_{ij}' for any orientation by means of transformation equations.[7] The expressions giving c_{16}', c_{26}', $\cdots$ c_{66}' (the constants relating to the x_y' strain) in terms of the c_{ij} for rotation about the X axis, are given below, θ being the

[6] W. Voigt, "Lehrbuch der Kristallphysik," 1928, p. 754.
[7] A. E. H. Love, "Mathematical Theory of Elasticity," 4th ed., p. 43.

angle between the Z' and Z axis (Fig. 2).

$$\begin{aligned} c_{16}' &= c_{26}' = c_{36}' = c_{46}' = 0, \\ c_{56}' &= c_{14}(\cos^2\theta - \sin^2\theta) + (c_{66} - c_{44}) \sin\theta\cos\theta, \\ c_{66}' &= c_{44}\sin^2\theta + c_{66}\cos^2\theta - 2c_{14}\sin\theta\cos\theta. \end{aligned} \tag{4}$$

Piezo-Electric Equations

The inverse piezo-electric relations for the X', Y', Z' system of axes can be expressed by the following equations:

$$\left.\begin{aligned} x_x' &= d_{11}'E_x' + d_{21}'E_y' + d_{31}'E_z' \\ y_y' &= d_{12}'E_x' + d_{22}'E_y' + d_{32}'E_z' \\ z_z' &= d_{13}'E_x' + d_{23}'E_y' + d_{33}'E_z' \\ y_z' &= d_{14}'E_x' + d_{24}'E_y' + d_{34}'E_z' \\ z_x' &= d_{15}'E_x' + d_{25}'E_y' + d_{35}'E_z' \\ x_y' &= d_{16}'E_x' + d_{26}'E_y' + d_{36}'E_z' \end{aligned}\right\} \tag{5}$$

When in quartz X', Y', Z' coincide with the crystallographic axes, eq. 5 reduces to the following:

$$\left.\begin{aligned} x_x &= d_{11}E_x \\ y_y &= -d_{11}E_x \\ z_z &= 0 \\ y_z &= d_{14}E_x \\ z_x &= -d_{14}E_y \\ x_y &= -2d_{11}E_y \end{aligned}\right\} \tag{6}$$

where

$$d_{11} = -6.36 \times 10^{-8}\frac{\text{esu}}{\text{dyne}},$$

$$d_{14} = 1.69 \times 10^{-8}\frac{\text{esu}}{\text{dyne}}.$$

For rotation about the X axis,

$$d_{26}' = (d_{14}\sin\theta - 2d_{11}\cos\theta)\cos\theta. \tag{7}$$

A Theory of Scanning and Its Relation to the Characteristics of the Transmitted Signal in Telephotography and Television

By PIERRE MERTZ and FRANK GRAY

By the use of a two-dimensional Fourier analysis of the transmitted picture a theory of scanning is developed and the scanning system related to the signal used for the transmission. On the basis of this theory a number of conclusions can be drawn:

1. The result of the complete process of transmission may be divided into two parts, (*a*) a reproduction of the original picture with a blurring similar to that caused in general by an optical system of only finite perfection, and (*b*) the superposition on it of an extraneous pattern not present in the original, but which is a function of both the original and the scanning system.

2. Roughly half the frequency range occupied by the transmitted signal is idle. Its frequency spectrum consists of alternating strong bands and regions of weak energy. In the latter the signal energy reproducing the original is at its weakest, and gives rise to the strongest part of the extraneous pattern. In a television system these idle regions are several hundred to several thousand cycles wide and have actually been used experimentally as the transmission path for independent signaling channels, without any visible effect on the received picture.

3. With respect to the blurring of the original all reasonable shapes of aperture give about the same result when of equivalent size. The sizes (along a given dimension) are determined as equivalent when the apertures have the same radius of gyration (about a perpendicular axis in the plane of the aperture).

4. With respect to extraneous patterns certain shapes of aperture are better than others, but all apertures can be made to suppress them at the expense of blurring. An aperture arrangement is presented which almost completely eliminates extraneous pattern while about doubling the blurring across the direction of scanning as compared with the usual square aperture. From this and other examples the degradation caused by the extraneous patterns is estimated.

IN the usual telephotographic or television systems the image field is scanned by moving a spot or elementary area along some recurring geometrical path over this field. In the more common arrangement this path consists simply of a series of successive parallel strips. Imagining the path developed or straightened out (or in the more common case, the strips joined end to end), this method of scanning is equivalent to transmitting the image in the form of a long narrow strip.

The theoretical treatment of such transmission has usually been developed by completely ignoring variations in brightness across the image strip, assuming the brightness to have a uniform distribution across this strip. This permits the image to be analyzed as an ordinary one-dimensional or single Fourier series (or integral) along the length of the strip; and the theory is then developed in terms of the

one-dimensional steady state Fourier components. Such a method of treatment naturally gives no information in regard to the reproduction or distortion of the detail in the original image across the direction of scanning, nor, as will appear below, does it give any detailed information in regard to the fine-structure distribution of energy over the frequency range occupied by the signal.

The need of a more detailed theoretical treatment originally arose in connection with studies of the reproduction of detail in telephotographic systems, especially in comparisons of distortion occurring along the direction of scanning with that across this direction. Later, this same need was strikingly shown by the discovery that a television signal leaves certain parts of the frequency range relatively empty of current components. Certain considerations indicated that a large part of the energy of a signal might be located in bands at multiples of the frequency of line scanning. Actual frequency analyses more than confirmed this suspicion. The energy was found to be so closely confined to such bands as to leave the regions between relatively empty of signal energy.

Such bands and intervening empty regions are illustrated by the examples of current-frequency curves in Fig. 1. These curves were taken with the various subjects as indicated, and the television current was generated by an apparatus scanning a field of view in 50 lines at a rate of about 940 lines per second. The energy is grouped in bands at multiples of 940 cycles and the regions between are substantially devoid of current components. In addition to the bands shown by the curves, it is known that similar bands occur up to about 18,000 cycles and that there is also a band of energy extending up from about 20 cycles.

Certain of the relatively empty frequency regions were also investigated by including a narrow band elimination filter in a television circuit. The filter eliminated a band about 250 cycles wide and was variable so that the band of elimination could be shifted along the frequency scale at will. By shifting the region of elimination along in this manner it was found that a band about 500 or 600 cycles wide could be removed from a television channel between any two of the current components without producing any detectable effect on the reproduced image.

At a later date a 1500-cycle current suitable for synchronization was introduced into a relatively empty frequency region, transmitted over the same channel with a television current, and filtered out—all without visibly affecting the image.

These results indicated quite clearly the need of a more complete

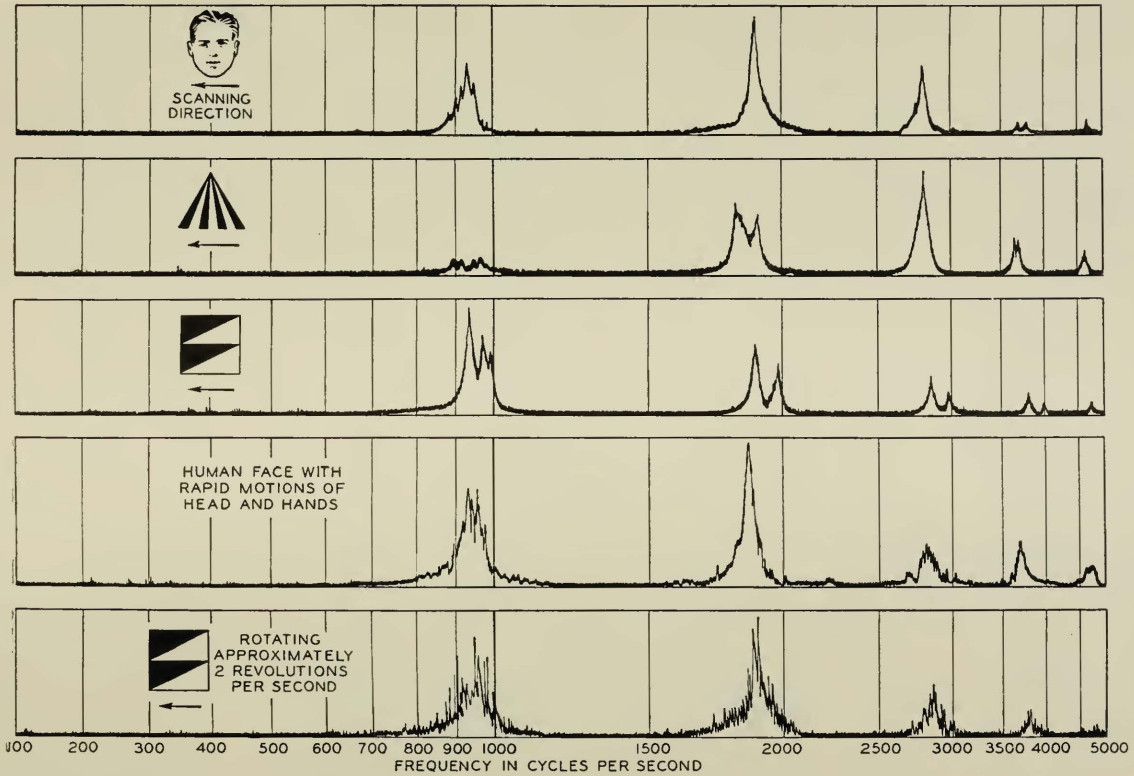


Fig. 1—Frequency analyses of television currents.

theory of the scanning processes used in telephotography and television and led to the study outlined in the following pages. Since this study will be confined to characteristics of the scanning processes all other processes in the system, wherever used, will be assumed to be perfect and cause no distortion.

The general trend of this more complete theory can be foreseen when it is considered that to obtain an adequate reproduction of the original it is necessary to scan with a large number of lines as compared with the general pictorial complexity of this original. This means that for any original presenting a large scale pattern (as distinguished from a random granular background) the signal pattern along successive scanning lines will, in general, differ by only small amounts. Thus, the signal wave throughout a considerable number of scanning lines may be represented to within a small error by a function periodic in the scanning frequency. Since such a function, developed in a Fourier series, is equal to the sum of sine waves having frequencies which are harmonics of the scanning line frequency, it will be natural to expect the total signal wave to have a large portion of its energy concentrated in the regions of these harmonics.

Furthermore, the existence of signal energy at odd multiples of half the scanning frequency will indicate the existence of a characteristic in the picture which repeats itself in alternate scanning lines. It is to be expected that such detail in a picture cannot be transmitted without accurate registry between it and the scanning lines and that when the detail spacing or direction or both differ somewhat from the scanning line spacing and direction, beat patterns between the two will be produced in the received picture which may be strong enough to alter considerably the reproduction of the original.

These phenomena are exactly what is observed, and will be treated in more quantitative fashion in the discussion below.[1]

An Image Field as a Double Fourier Series

Let us first consider the usual expression of the image field as a single Fourier series. The picture will be considered as a "still" so that entire successive scannings are identical. Then if the long strip corresponding to one scanning extends from $-L$ to $+L$, the illumina-

[1] In the following treatment an effort has been made to confine the necessary mathematical demonstrations almost exclusively to two sections entitled, respectively, "Effect of a Finite Aperture at the Transmitting Station," and "Reconstruction of the Image at the Receiving Station." Even in these sections a number of conclusions are explained in text which do not require reading the mathematics if the demonstrations are taken for granted. The occasional mathematical expressions occurring in the earlier sections are very largely for the purpose of introducing notation.

tion E as a function of the distance x along the strip may be expressed as the sum of an infinite number of Fourier components, thus:

$$E(x) = \sum_{n=0}^{\infty} a_n \cos\left(\frac{n\pi x}{L} + \varphi_n\right). \tag{1}$$

In this summation a_n represents the intensity of the nth component and φ_n its phase angle. The complete array of these for all components will vary if the picture is changed.

The cosine series above is very convenient for physical interpretation. It will be simple, however, for some of the later mathematical work to use the corresponding exponential series. The cosine series can be returned to, each time, as physical interpretation is required. That is, since

$$2a \cos\left(\frac{\pi x}{L} + \varphi\right) = (ae^{i\varphi})e^{(i\pi x/L)} + (ae^{-i\varphi})e^{(-i\pi x/L)} \tag{2}$$

the series in equation (1) can be written

$$E(x) = \sum_{n=-\infty}^{+\infty} A_n \exp i\pi(nx/L) \tag{3}$$

if we make

$$A_n = (1/2)a_n \exp(i\varphi_n)$$

and

$$A_{-n} = (1/2)a_n \exp(-i\varphi_n) \tag{4}$$

and if we use the notation $\exp\theta = e^{\theta}$.

In this new summation the complex amplitude A_n represents both the absolute intensity and the phase angle of the nth component. The complex amplitude of the corresponding component with a negative subscript is merely the conjugate of this.

As has already been noted, however, and as might readily be expected, the single Fourier series in equations (1) or (3) above do not always represent a two-dimensional picture with sufficient completeness. In order to consider the two-dimensional field more in detail, let us assume that Fig. 2 represents such an image field of dimensions $2a$ and $2b$, and take axes of reference x and y as indicated. The brightness or illumination of the field is a function $E(x, y)$ of both x and y. Along any horizontal line (i.e., in the x direction, constantly keeping $y = y_1$) the illumination may be expressed as a single Fourier series

$$E(x, y_1) = \sum_{m=-\infty}^{+\infty} A_m \exp i(mx/a). \tag{5}$$

Along any other line in the x direction a similar series holds with different coefficients, that is, the A's are functions of y. They may

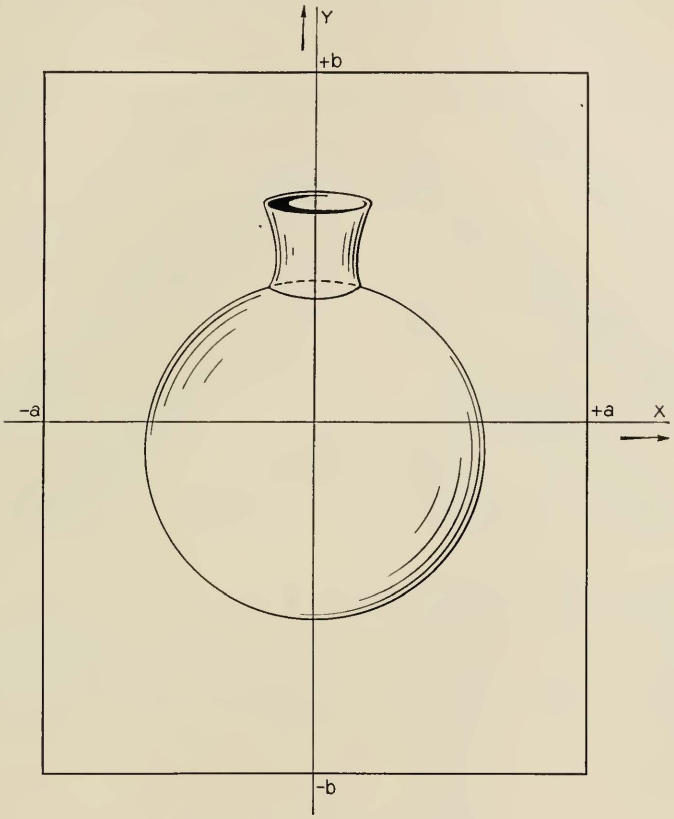


Fig. 2—Scanned field and image.

therefore each be written as a Fourier series along y

$$A_m = \sum_{n=-\infty}^{+\infty} A_{mn} \exp i\pi(ny/b). \tag{6}$$

Substitution in equation (5) gives the double Fourier series,

$$E(x, y) = \sum_{m=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} A_{mn} \exp i\pi \left(\frac{mx}{a} + \frac{ny}{b} \right). \tag{7}$$

For purposes of physical interpretation, as in the case of the simple Fourier series, it is desirable to combine the $+m$, $+n$ term with the $-m$, $-n$ term (giving the single $(m, +n)$th component) and similarly

the $+m$, $-n$ with the $-m$, $+n$ terms (giving the single $(m, -n)$th component). This brings equation (7) back to a cosine series,

$$E(x, y) = \sum_{m=0}^{\infty} \sum_{n=-\infty}^{+\infty} a_{mn} \cos \left[\pi \left(\frac{mx}{a} + \frac{ny}{b} \right) + \varphi_{mn} \right] \tag{8}$$

when

$$A_{mn} = (1/2) a_{mn} \exp (i\varphi_{mn})$$

and

$$A_{-m-n} = (1/2) a_{mn} \exp (-i\varphi_{mn})$$

and where a_{mn} is always a real quantity. Each term of this series represents a real, two-dimensional, sinusoidal variation in brightness extending across the image field. The image is built up of a superposition of a series of such waves extending across the field in various directions and having various wave lengths.

Imagining brightness as a third dimension, we may, as an aid in visualizing the components of an image field, draw separate examples of various components as shown in Fig. 3. It will be noted that any given component (m, n) passes through m periods along any horizontal line in the image field, and through n periods along any vertical line. The slope of the striations with respect to the x-axis is therefore $-mb/na$ (the negative reciprocal of the slope of the line of fastest variation in brightness). For the same values of m and of n, the m, $+n$ component and the m, $-n$ component have equal wave lengths but are sloped in opposite directions to the x-axis. If m is zero the crests are parallel to the x-axis; if n is zero they are parallel to the y-axis. The component with both m and n zero is a uniform distribution of brightness covering the entire image field. The wave length of a component is

$$\lambda_{mn} = 1 \Big/ \sqrt{\left(\frac{m}{2a} \right)^2 + \left(\frac{n}{2b} \right)^2}.$$

A complete array of the components, up to m and n equal to 4, is illustrated in Fig. 4.

As of course is characteristic of the harmonic analysis, the wave lengths and orientations of the components are seen to vary only with the shape and size of the rectangular field, and to be independent of the particular subject in the field. A change of subject, or motion of the subject, merely alters the amplitudes of the components and shifts their phase; but their wave length and inclination with respect to the x-axis remain unchanged. Consequently, for the same rectangular field all subjects appearing in it may be considered as built up from the same set of components. For a "still" subject, the amplitudes and

phase angles of the cosine components, or the complex amplitudes of the exponential components, remain constant with time. For a moving subject these complex amplitudes may be considered modulated as functions of time.

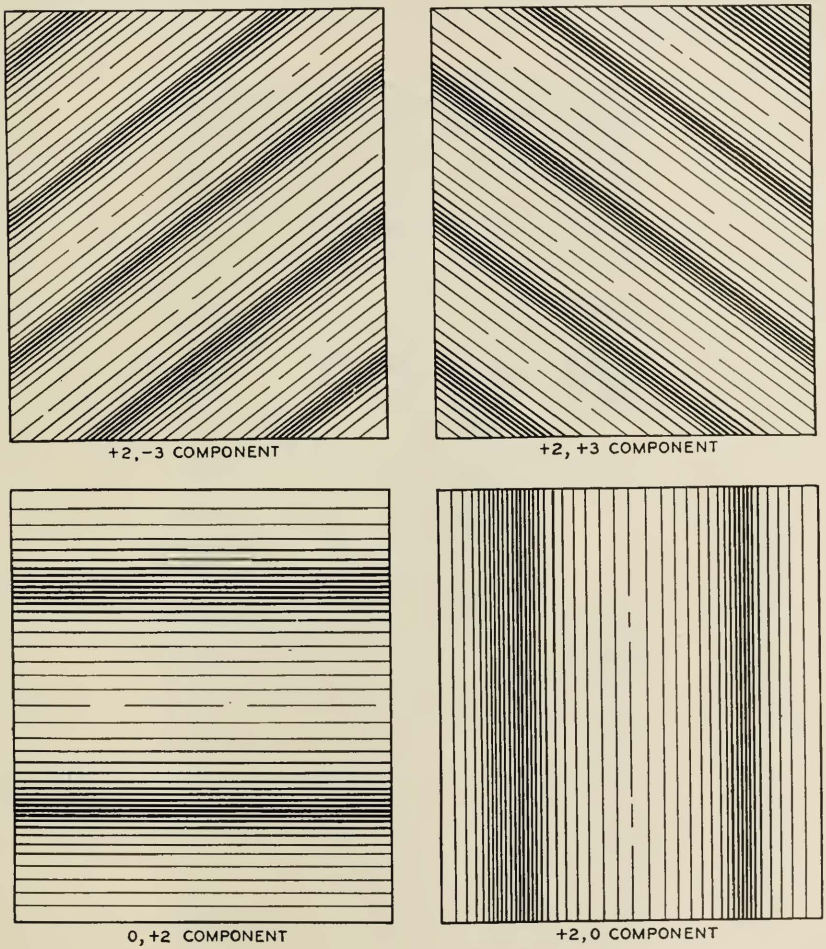


Fig. 3—Examples of field components.

The real amplitudes a_{mn} for a circular area of uniform brightness on a black background are relatively easily calculated, and this subject is also a good one to study as a picture from some points of view because it has a simple sharp border sloping in various directions. The amplitudes

for a circle of unit illumination of radius R are

$$a_{mn} = \frac{\pi R^2}{2ab}\left(\frac{\lambda_{mn}}{2\pi R}\right) J_1\left(\frac{2\pi R}{\lambda_{mn}}\right), \tag{9}$$

where J_1 is the first order Bessel function. In this particular subject all components of a given wave length have equal amplitudes; and the

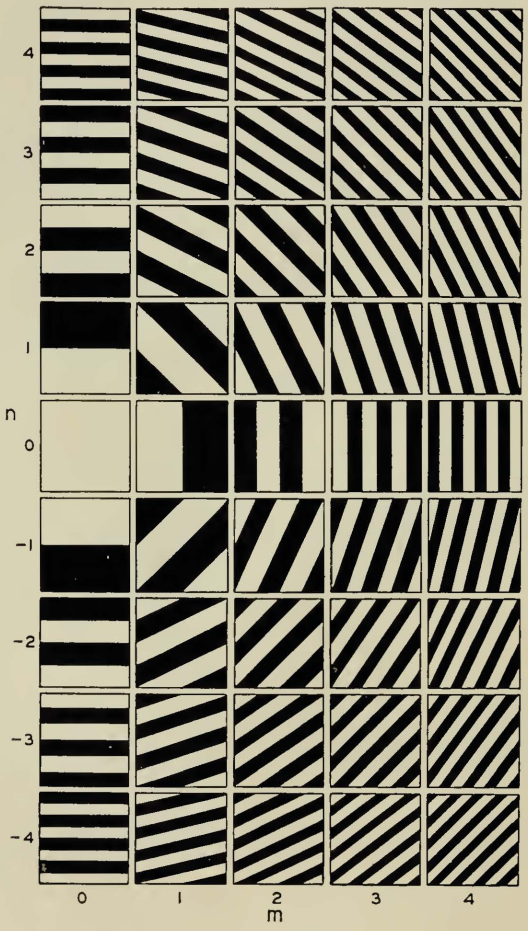


Fig. 4—Array of field components.

amplitudes may therefore be plotted as a function of wave length alone, as in Fig. 5. The curve illustrates the rapidity with which the amplitudes fall off for the higher order components in a subject of this nature.

The Frequency Spectrum of the Signal

When an image field is scanned by a point aperture tracing across it, each portion of the picture traversed causes variations in the light reaching the light sensitive cell and is thus translated into a corre-

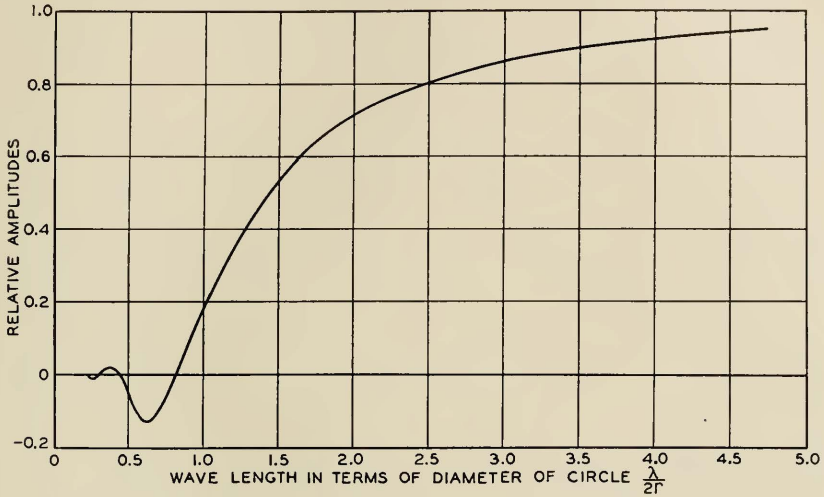


Fig. 5—Amplitudes of field components for a circular area of brightness.

sponding signal. Further, each Fourier component in the field is translated into a corresponding Fourier component in the signal. An equivalent translation occurs when a pencil of light traces over a photographic film in telephotography, or when a subject is scanned by a beam of light in television, whether or not a simple flat two-dimensional image is ever physically formed at the transmitting station. For clarity and simplicity, the discussion will be confined to the case in which a point aperture traces across a plane image field.

In most systems the aperture traces a line across the field and then there is a sudden jump back to the beginning of the next succeeding line. This discontinuous motion is naturally not easily subjected to mathematical treatment. It is much simpler to deal with the equivalent result that would be obtained if the scanning point, instead of tracing successive parallel paths across the same field, moved continuously across a series of identical fields. Such an equivalent scanning motion can fortunately easily be used because a double Fourier series represents not only a single field, but a whole succession of identical image fields covering the entire xy plane, and repeated periodically in both the x and y direction as illustrated in Fig. 6.

The equivalent of scanning a single field in parallel lines is obtained by assuming that the scanning point moves across the repeated fields along a sloping path as indicated. Let u be the velocity parallel to

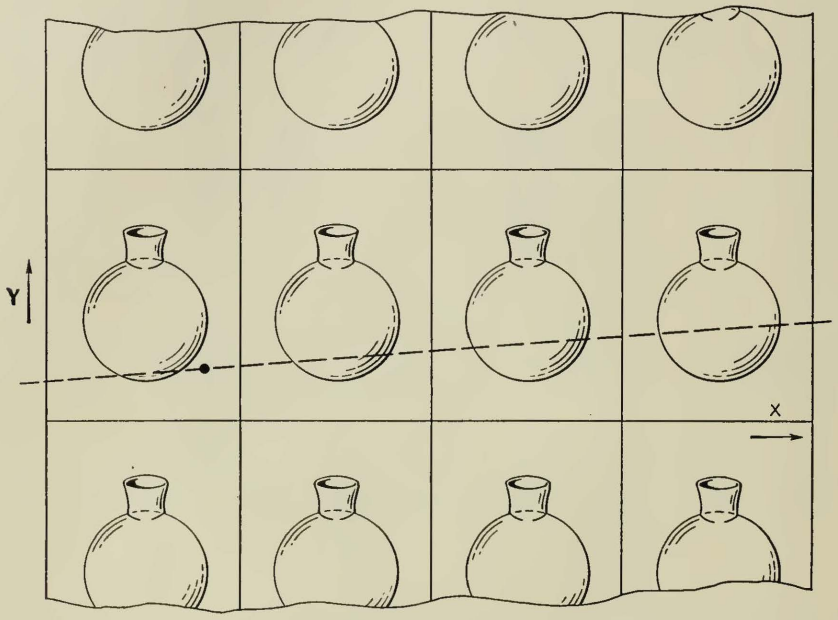


Fig. 6—Array of periodically recurring scanned fields.

the x axis and v the velocity parallel to the y axis. Then the picture illumination at the scanning point at any instant, and consequently the signal current, may be obtained by substituting

$$x = ut, \qquad y = vt$$

in the double Fourier series representing the image field, equation (8). Of course the entire expression must be multiplied by a factor K which is the constant ratio between the signal current and the picture illumination. This gives for the real signal as a function of time [2]

[2] It will be noted that this process does not explore the picture completely, inasmuch as, no matter how fine the scanning, there will always be unexplored regions between scanning lines. In this respect the process is quite analogous to that followed in analyzing a function of a single variable into a simple Fourier series when the values of the function are given only at discrete (even though closely spaced) values of the variable. The complete exact theory, which necessarily depends upon the size and shape of the finite scanning spot or aperture, will be given further below.

$$I(t) = K \sum_{m=0}^{\infty} \sum_{n=-\infty}^{+\infty} a_{mn} \cos \left[\pi \left(\frac{mu}{a} + \frac{nv}{b} \right) t + \varphi_{mn} \right]. \qquad (10)$$

Thus if u and v are constants, each wave of the image field gives rise to a corresponding Fourier component of the signal. The frequencies of the signal components are

$$f = \frac{mu}{2a} + \frac{nv}{2b}. \qquad (11)$$

The frequency spectrum of the signal is thus made up of a series of possible discrete lines, the position of which in that spectrum is determined by u and v, that is, by the particular scanning motion employed. We shall designate these lines by the indices $m, +n$ and $m, -n$, as they are correlated with the particular components of the image field that generated them.

A different choice of values for u and v (so long as these, once having been chosen, remain constant) changes the location of the lines in the frequency spectrum, but their amplitudes, depending only on the corresponding components of the image field, remain unchanged. In other words, the lines in the frequency spectrum of the signal are characteristic of the image field, and the scanning motion merely determines where they will appear in the frequency spectrum. Thus, if for a given subject the distribution of energy over the frequency scale is known for one method of scanning, it can be predicted for a great many other methods.

To scan a field in lines approximately parallel to the x axis, the velocity v must be made small compared to u. Under such conditions, $u/(2a)$ of equation (11) is the line scanning frequency and $v/(2b)$ is the frequency of image repetitions (or "frame frequency"). The frequency spectrum of the signal for a "still" picture thus consists of certain fundamental components at multiples of the line scanning frequency $u/(2a)$, each of which is accompanied by a series of lines spaced at equal successive intervals to either side of it. The spacing between these satellites is the image repetition or frame frequency $v/(2b)$.

If the picture changes with time the amplitudes of these fundamental lines and their satellites are modulated, also with respect to time. In other words they each develop sidebands or become diffuse. The diffuseness will not overlap from satellite to satellite unless the frequency of modulation becomes as great as half the frame frequency.

Thus for motions in the picture which are not too fast to be expected to be reproduced with reasonable fidelity, this diffuseness of the fundamental lines and their satellites will not obliterate their identity.

A diagrammatic arrangement of some of the possible lines in a frequency spectrum, with their corresponding m and n indices, is shown in Fig. 7.

It is important to note that the correlation between the wave lengths of the field components and the frequencies of the current components is not the one that is naturally assumed on first consideration. We

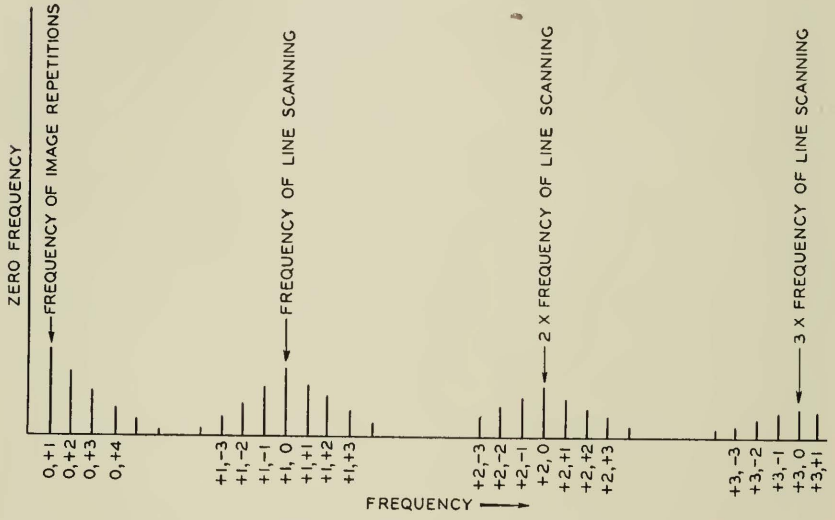


Fig. 7—Diagram of signal frequency spectrum.

are quite likely to make the erroneous assumption that high frequencies correspond to all sharp changes in brightness and that low frequencies correspond only to slow changes. The error in this assumption is readily realized by noting that sharp changes in brightness may generate very low frequencies if the scanning point passes over them in a sloping direction. An actual correlation is shown schematically in Fig. 8. It is seen that the same general type of correlation is repeated periodically over the frequency scale at multiples of the line scanning frequency. There are evidently numerous regions of the spectrum in which short image waves, or fine grained details of the image field, may appear in the signal. They are not confined to the high-frequency region alone.

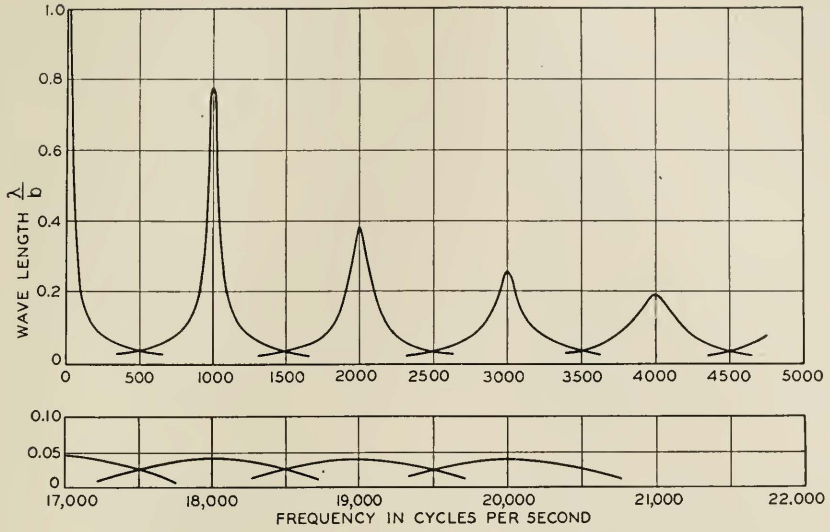


Fig. 8—Correlation between wavelength and frequency of signal components.

In telephotography the frequency of line scanning is usually low and the groups of lines in the frequency spectrum are so closely spaced that such fine grained details of the signal are of little practical importance as far as the electrical parts of the system are concerned. In television, however, these bands are widely spaced, of the order of 1000 cycles or so apart, and such details of the signal are quite important.

As a specific example, it is interesting to plot the frequency spectrum of the television signal that results from scanning a circular area of uniform brightness on a black background. So far as the present theory extends, this may be done by converting the field components of equation (9) into current components with the aid of equations (10) and (11). Taking $b/a = 1.28$, the radius of the circle as $b/3$, and assuming that the field is scanned in 50 lines 20 times per second, we obtain the amplitude-frequency spectrum shown in Fig. 9. Since it is not convenient to show the individual current components—only 20 cycles apart—the curve shows simply the envelope of the peaks of these components. At low frequencies, the energy is largely confined to bands at multiples of the line scanning frequency, 1000 cycles, and to an additional band extending up from zero frequency. In the retions between the bands, the signal components are so small that they do not show when plotted to the same scale. At higher frequencies the signal energy as thus far computed is not confined to such bands. It

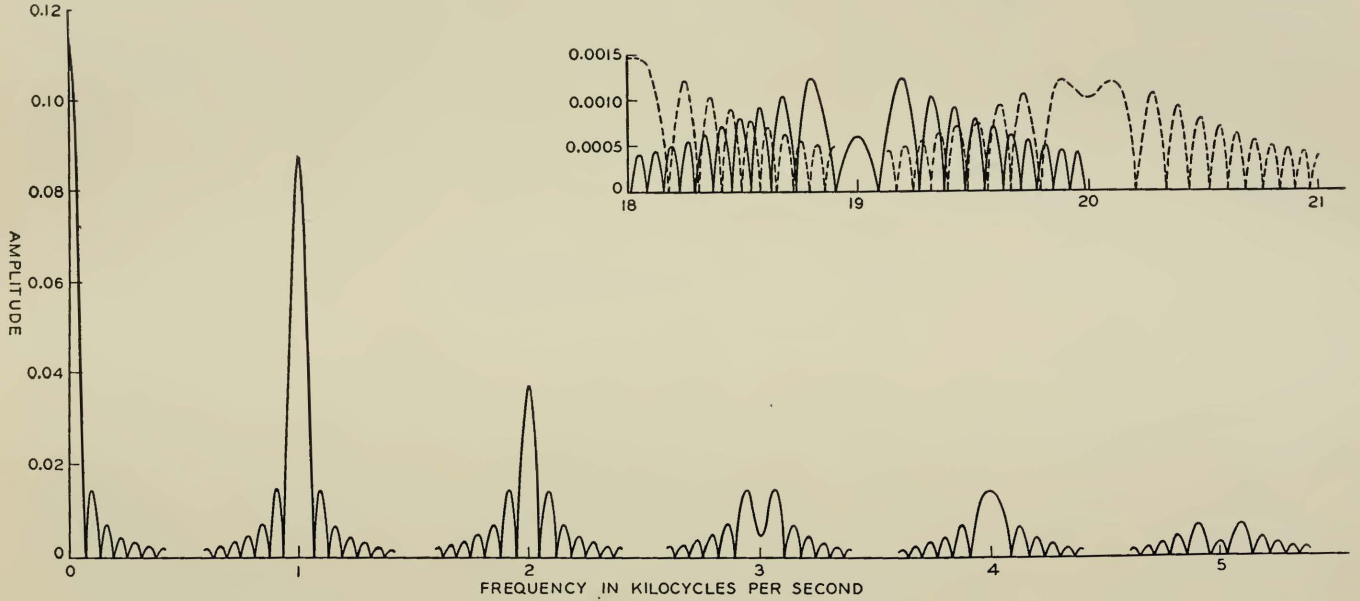


Fig. 9—Amplitudes of signal components for a circular area of brightness (diameter, 1/3 of field length).

will be shown farther on, however, that the effect of the use of a finite aperture for scanning is to confine the signal energy more rigorously to such bands throughout the frequency range.

The theoretical energy distribution for the circular area is in excellent agreement with actual frequency analyses of television currents, which show the energy confined to bands at multiples of the line scanning frequency with apparently empty regions between. It is evident from the theory so far, however, that these regions are not really empty but are filled with weak signal components representing fine details of the subjects; and subjects of greater pictorial complexity than a simple circular area may be devised to give large signal components in such regions. We must therefore look for other factors to explain why these frequency regions do not transmit any appreciable details of an image.

Confusion in the Signal

With the usual method of scanning, one such factor is the confusion of components in the signal. This confusion arises from the fact that two or more image components sloping across the field in different directions may intercept the line of scanning with their crests spaced exactly the same distance apart along this line of scanning. As the scanning point passes over them they thus give rise to signal current components of exactly the same frequency. Consequently the two image components are represented by a single, confused, signal current component that can transmit no information whatever in regard to their relative amplitudes and phases. This confusion evidently depends on the scanning path.

If the image field is scanned in N lines, the velocity v of the scanning point parallel to the y axis is

$$v = \frac{b}{Na} u \tag{12}$$

and the signal frequencies from equation (11) are

$$f = \frac{u}{2a}\left(m + \frac{n}{N}\right). \tag{13}$$

Field components with indices m, n and m', n' such that

$$m + \frac{n}{N} = m' + \frac{n'}{N} \tag{14}$$

give rise to current components of the same frequency.

In other words, the bands of components in the frequency spectrum really overlap. Consequently the components of one band may coincide in frequency with the components of adjacent bands. Such coinciding components are illustrated schematically in Fig. 10.

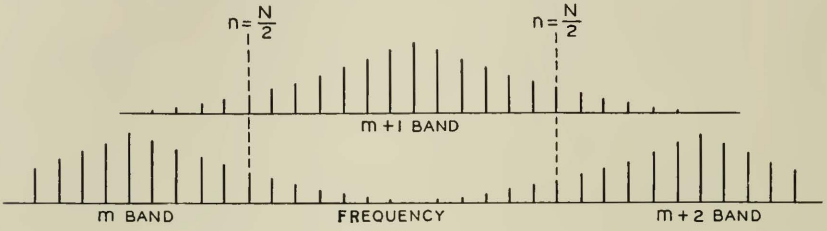


Fig. 10—Coinciding lines of confused bands.

It is obvious that a single a-c. component cannot transmit the separate amplitudes and phases of two or more image components. Consequently the receiving apparatus has no information to judge how the components in the original image are supposed to be distributed in the reproduction.

The situation is most serious where the intensities of coinciding components have the same order of magnitude, that is, at the centers of the frequency regions intermediate to the strong bands. The confusion in these regions is the most important factor that renders them incapable of transmitting any appreciably useful image detail.

On first consideration it would appear that the overlapping of bands in the signal might result in a hopeless confusion. The situation is saved, however, by the fact that components with large n numbers will tend to be weak due to the convergence of the Fourier series, and are further reduced, as will be shown later, by the effects of a scanning aperture of finite size. They therefore do not usually seriously interfere with the stronger components. The interference usually manifests itself in the form of serrations on diagonal lines and occasional moiré effects in the received picture.

Confusion in the signal may be practically eliminated by using an aperture of such a nature that it cuts off all components with n numbers greater than $N/2$, that is, cuts off each band before it reaches the center of the intervening frequency regions so that adjacent bands do not overlap. The practical possibilities of this arrangement will be discussed further below.

The mere elimination of confusion in the signal itself does not necessarily prevent the appearance of extraneous components in the reproduced image. The receiving apparatus itself must be so designed that

when it reproduces all the image components represented by a given signal component, it suitably suppresses all those but the dominant one desired.

Effect of a Finite Aperture at the Transmitting Station

In the preceding pages the scanning aperture has been assumed as infinitesimal in size, or merely a point. In any actual scanning system the necessary finite size of the aperture introduces effects which will now be considered for the transmitting end.

Let us first review briefly the usual theory of this effect when the picture is analyzed simply as a one-dimensional Fourier series. According to equation (3) above, this series is

$$E_1(x) = \sum_{n=-\infty}^{+\infty} A_n \exp i\pi(nx/L).$$

Let ξ be a coordinate fixed with respect to the scanning aperture as shown in Fig. 11 and let the optical transmission of the aperture for

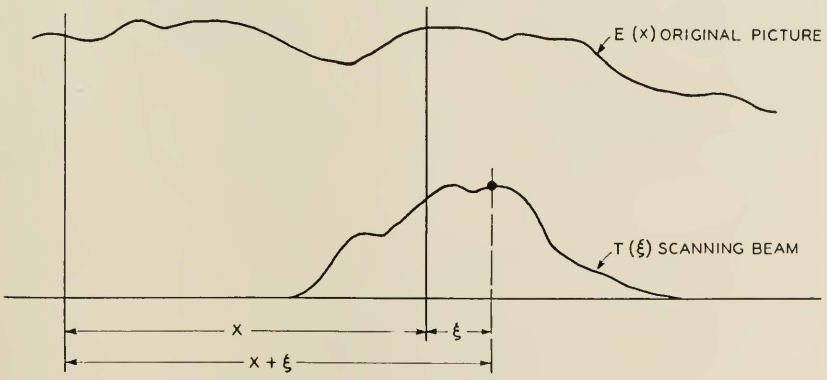


Fig. 11—Analysis of one-dimensional scanning operation.

any value of ξ be $T(\xi)$.[3] Then if x is taken as a coordinate of the origin of ξ the illumination at any point ξ of the aperture is

$$E_1(x + \xi) = \sum_{n=-\infty}^{+\infty} A_n \exp i\pi(n[x + \xi]/L), \qquad (15)$$

[3] This optical transmission may represent either the transparency of an aperture of constant width or the width of an aperture which is a shaped hole in an opaque screen.

so that the total flow of light through the aperture at any position x is

$$F_1(x) = \int_{\text{aperture}} T(\xi)E_1(x + \xi)d\xi. \tag{16}$$

Since x is a constant with respect to the integration the exponential term may be factored and the part involving x only may be brought outside the integral sign. This gives

$$F_1(x) = \sum_{n=-\infty}^{+\infty} Y(n)A_n \exp i\pi(nx/L), \tag{17}$$

where

$$Y(n) = \int_{\text{aperture}} T(\xi) \exp i\pi(n\xi/L)d\xi. \tag{17'}$$

For a symmetrical aperture (that is, about the origin of ξ)

$$Y(n) = \int_{\text{aperture}} T(\xi) \cos (\pi n\xi/L)d\xi \tag{17''}$$

and $Y(n)$ in this case is, therefore, a pure real quantity.

The important conclusion to be drawn from equation (17) as to the effect of a finite transmitting aperture is that it multiplies the complex amplitude A_n of each original image component by a quantity $Y(n)$ which is independent of the picture being scanned. This is entirely similar to the effect of a linear electrical network in a circuit, and the quantity $Y(n)$ is quite analogous to the transfer admittance of that network.

The quantity $Y(n)$ has been plotted for variously shaped apertures in Fig. 12. For convenience in comparison, the ordinates of each curve have been multiplied by a numerical factor to make $Y(0) = 1$. The curves show the characteristics that are by this time familiar, which are that the effect of the finite size of the scanning aperture in the transmitter is similar to that of introducing a low-pass filter in the circuit, namely, cutting down the amplitudes of the signal components for which n is numerically high, i.e., the high-frequency components.

The curves are remarkable, however, in that in the useful frequency band (i.e. from $n = 0$ to something like half of the first root of $Y(n) = 0$) all the distributions considered give practically the same transfer admittance if the dimensions of the beam along the direction of scanning are suitably chosen, as has been done in the figure. This results from

[4] The integral is mathematically taken from $-\infty$ to $+\infty$ but the regions outside the aperture give no contribution since the integrand is there equal to zero.

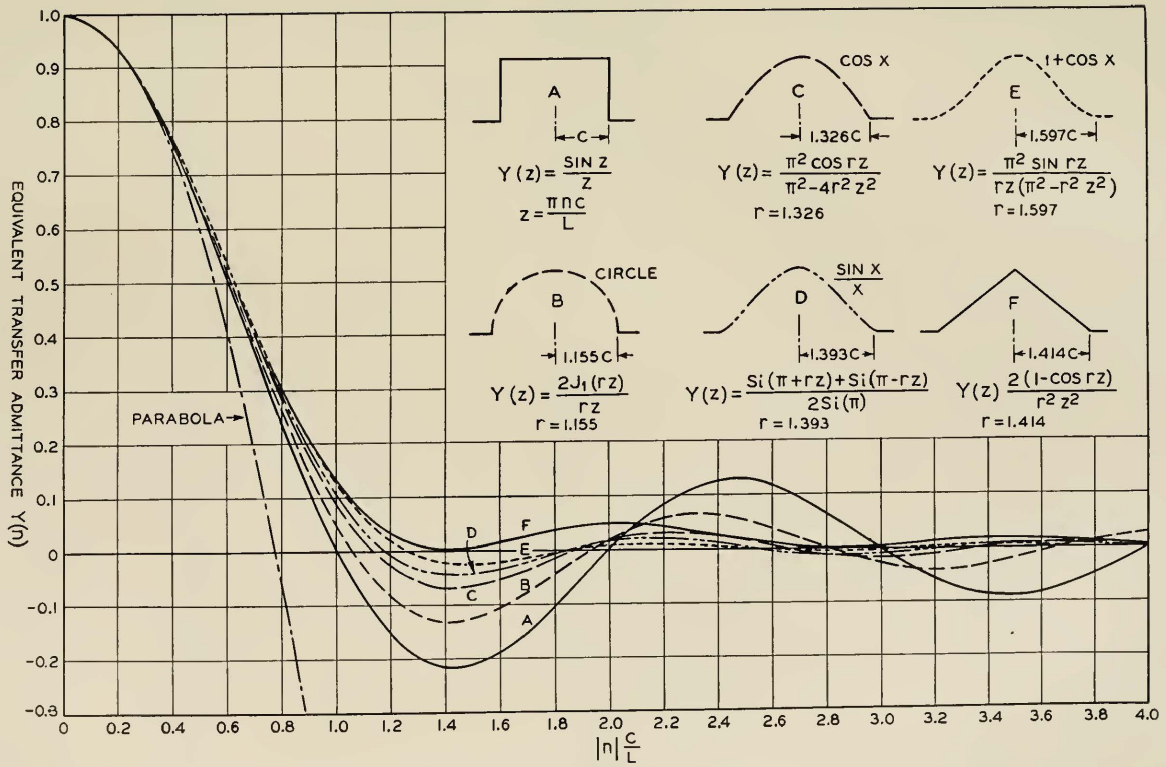


Fig. 12—Equivalent transfer admittances for various distributions of light in the scanning beam.

the physical limitation that the illumination in any part of the beam must be positive, that is, the illumination from one part of the beam must always add to that from another part and cannot subtract from it.[5] This observation enables one to define the resolution of two apertures of different shapes as being equal along a certain direction when their transfer admittances in the useful frequency range show the same filtering effect, if that direction is used as the direction of scanning. This will occur when the radii of gyration (about a normal axis in the plane of each aperture) are equal. According to this definition all the apertures illustrated in Fig. 12 have the same resolution along a horizontal direction.

When the picture is analyzed as a two-dimensional Fourier series the equations which have been given above become

$$E_1(x, y) = \sum_{m=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} A_{mn} \exp i\pi \left(\frac{mx}{a} + \frac{ny}{b} \right),$$

$$E_1(x + \xi, y + \eta) = \sum_{m=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} A_{mn} \exp i\pi \left(\frac{m[x + \xi]}{a} + \frac{n[y + \eta]}{b} \right), \tag{18}$$

$$F_1(x, y) = \int\int_{\text{aperture}} T_1(\xi, \eta) E_1(x + \xi, y + \eta) d\xi d\eta, \tag{19}$$

$$F_1(x, y) = \sum_{m=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} Y_1(m, n) A_{mn} \exp i\pi \left(\frac{mx}{a} + \frac{ny}{b} \right), \tag{20}$$

where

$$Y_1(m, n) = \int\int_{\text{aperture}} T_1(\xi, \eta) \exp i\pi \left(\frac{m\xi}{a} + \frac{n\eta}{b} \right) d\xi d\eta. \tag{20'}$$

For an aperture symmetrical about both ξ and η axes

$$Y_1(m, n) = \int\int_{\text{aperture}} T_1(\xi, \eta) \cos \pi \left(\frac{m\xi}{a} + \frac{n\eta}{b} \right) d\xi d\eta. \tag{20''}$$

[5] The shape of the transfer admittance curve near $n = 0$ depends upon the power of n in the first variable term of the Taylor expansion for $Y(n)$ about $n = 0$, and upon the sign of this term. Assuming a symmetrical aperture, the expansion from equation (17″) is

$$Y(n) = \int T d\xi - \frac{\pi^2 n^2}{2!\, L^2} \int \xi^2 T d\xi + \frac{\pi^4 n^4}{4!\, L^4} \int \xi^4 T d\xi - \cdots.$$

Since T is everywhere positive the first variable term is always in n^2 and negative. The shape of the curve near $n = 0$ is, therefore, always a parabola (indicated in Fig. 12), which can be made the same parabola by suitably choosing the two disposable constants in the aperture. Even after departing from this common parabola, the curves maintain the same general shape over a substantial range; for the next variable term is in n^4 and positive, and has the same order of magnitude for all usual types of apertures. Consequently, the curves for these apertures have approximately the same shape over a wide range extending up from $n = 0$. The results are the same for an unsymmetrical aperture, but the reasoning is more involved.

In the two-dimensional case $T(\xi, \eta)$ is defined, for a hole in an opaque screen, as unity throughout the area of the hole, and zero for the screen. Where the aperture is covered with a non-uniform screen T may take on intermediate values.

The transfer admittances have been calculated for a variety of shapes of aperture in Appendix I. It will be noted that for those types of aperture for which T can be separated into two factors, one a function of ξ only and the other a function of η only, namely, for which

$$T_1(\xi, \eta) = T_\xi(\xi) \cdot T_\eta(\eta), \tag{21}$$

then equation (20′) becomes

$$Y_1(m, n) = \int_{\text{aperture}} T_\xi(\xi) \exp(i\pi m\xi/a)d\xi \int_{\text{aperture}} T_\eta(\eta) \exp(i\pi n\eta/a)d\eta$$
$$= Y_\xi(m) \cdot Y_\eta(n) \tag{22}$$

and Y_ξ and Y_η are each one-dimensional integrals of the type illustrated in Fig. 12.

The rectangular aperture is a simple case of this type. Assume the field to be scanned in N lines and take the dimensions of the aperture, $2c$ and $2d$ parallel to the x and y axes, respectively, as

$$\frac{c}{a} = \frac{d}{b} = \frac{1}{N}. \tag{23}$$

Then

$$Y_\xi(m) = \frac{\sin \pi mc/a}{\pi mc/a}$$

and

$$Y_\eta(n) = \frac{\sin \pi nd/b}{\pi nd/b}$$

and the frequency corresponding to a given signal component mn is, from equation (11)

$$f = \frac{u}{2a}\left(m + \frac{n}{N}\right).$$

Thus, $Y_1(m, n)$ considered as a function of the signaling frequency corresponding to each component of indices mn, consists of a succession of similar curves $u/2a$ cycles apart, corresponding to the successive integral values of m (these curves are themselves really not continuous but consist of a succession of points $u/(2aN)$ cycles apart. For convenience, however, the drawings will always show the curves as con-

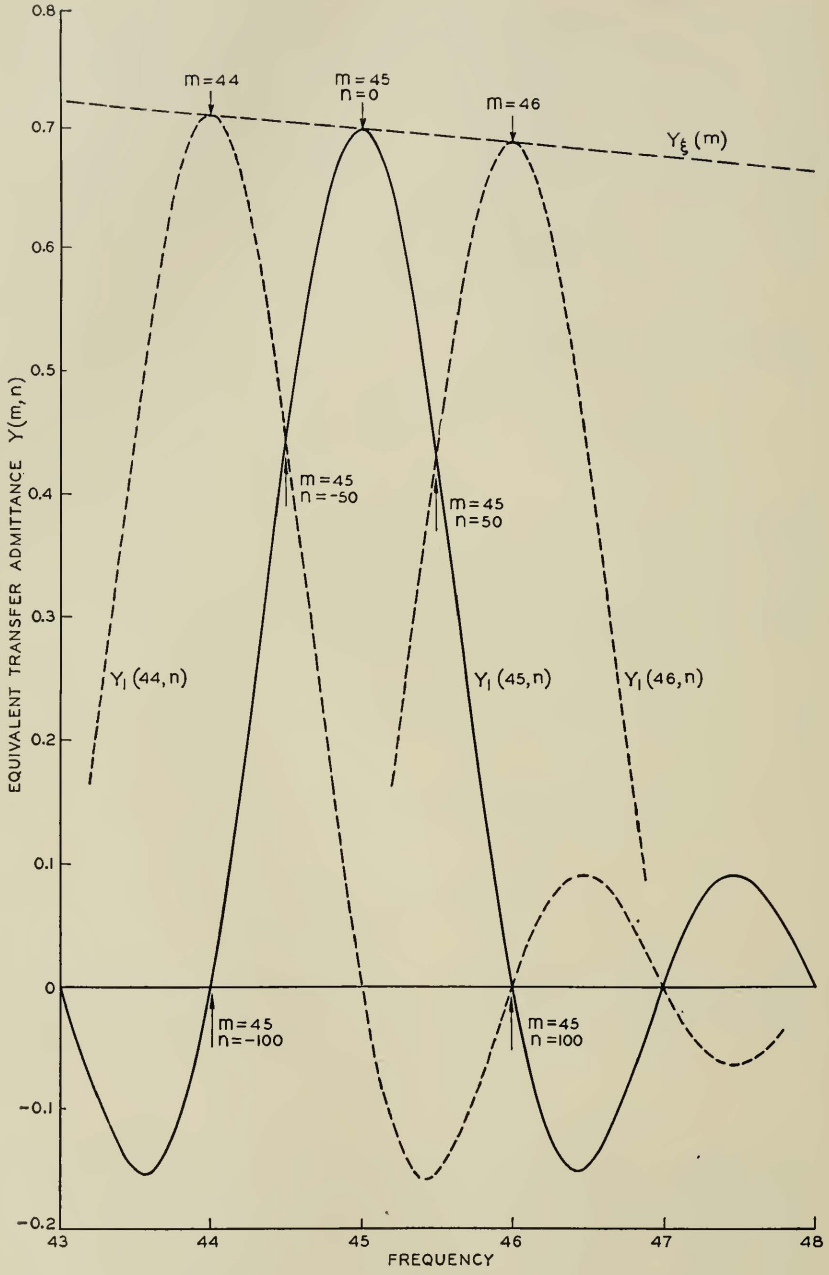


Fig. 13a—Detail of equivalent transfer admittance of aperture for two-dimensional scanning.

tinuous). Each of the curves is of the equation

$$Y_1(m, n) = Y_\xi(m) \frac{\sin \frac{2\pi Nc}{u}\left(f - \frac{mu}{2a}\right)}{\frac{2\pi Nc}{u}\left(f - \frac{mu}{2a}\right)}$$

and therefore has a peak of the value $Y_\xi(m)$ at the point where $n = 0$ or $f = mu/2a$, and trails off from the peak in each direction according to a curve of the same shape as curve "A" in Fig. 12. The successive curves are all of identical shape, but each one is to a reduced scale of ordinates as compared with the preceding (in the useful frequency range) as imposed by the factor $Y_\xi(m)$.

The peaks, it will be noted, occur at the frequencies occupied by what have been called the fundamental components (as distinguished from the satellite lines) in the discussion above on the frequency spectrum of the signal.

Assuming N to be 100 and for simplicity taking the factor $u/2a$ as equal to 1, a plot is shown in Fig. 13a of $Y_1(m, n)$ over a very limited region near the upper end of the useful frequency range. The curve shown in a solid line represents $Y_1(m, n)$ for $m = 45$, and the dotted curves on either side represent the function for $m = 44$ and 46, respectively.

The function has been redrawn for the complete useful range of frequencies and a little beyond, in Fig. 13b, with the frequencies to a logarithmic scale. This logarithmic plot opens out the scale at the low frequencies and enables the fine structure of the function to be indicated there, and still enables the complete range of useful frequencies to be shown without requiring a prohibitive size of drawing (it has, however, the disadvantages that the distortion in the frequency scale then masks the symmetry of the individual curves around the fundamental lines, the similarity of shape of these individual curves, and also the constant frequency separation between the successive fundamental lines).

The function $Y_1(m, n)$, as is clear from equation (22) and Figs. 13a and b, consists of a sort of envelope function $Y_\xi(m)$, "modulated" by a fine structure function $Y_\eta(n)$. The latter function has the value unity at the positions of the fundamental lines in the frequency spectrum of Fig. 7 and diminishes for the satellite lines in the same way that the envelope function diminishes for the fundamental lines away from zero frequency. It will be seen that the envelope function is the only one obtained by the simple one-dimensional analysis. The

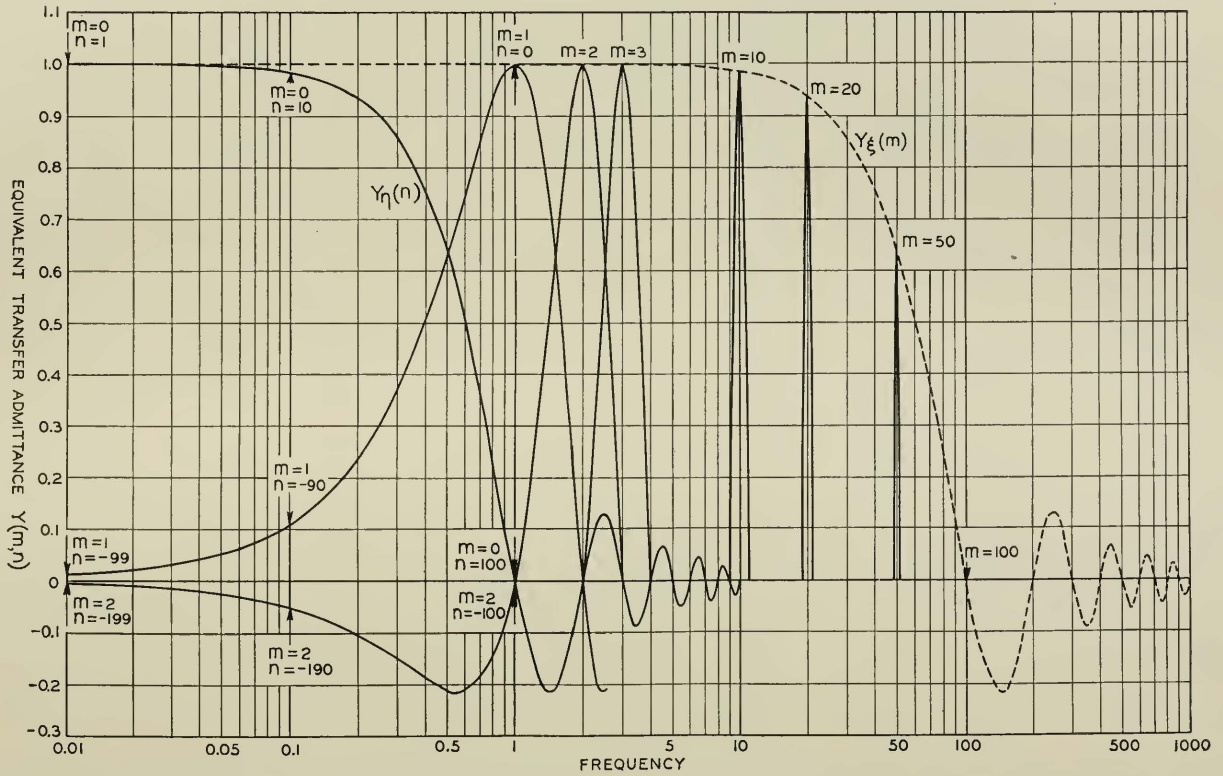


Fig. 13b—Equivalent transfer admittance of aperture for two-dimensional scanning, to a logarithmic frequency scale.

complete function shows, by the very small transfer admittance in the regions half-way between the fundamental lines, an additional reason why the signal currents in these regions will be weak and relatively incapable of transmitting appreciable image detail.

Examination of the other apertures for which computations are given in Appendix I will show that, in general, for all ordinary apertures the same broad phenomena are observed as for the rectangular aperture, although it is not always possible to express the complete function in the simple product form above, in which case the curves for the successive values of m will vary gradually in shape.

The final signal current is proportional to the light flux through the aperture, given in equation (20). Neglecting constant factors it may, therefore, be written as

$$I(t) = \sum_{m=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} Y_1(m, n) A_{mn} \exp i\pi \left(\frac{mu}{a} + \frac{nv}{b} \right) t. \qquad (24)$$

Reconstruction of the Image at the Receiving Station

At the receiving station the signal current is translated back into light to illuminate an aperture moving in synchronism with the one at the sending end. Neglecting constant factors the flow of light $F_2(t)$ to the receiving aperture is

$$F_2(t) = I(t). \qquad (25)$$

Let $E_2(x, y)$ be the resulting apparent illumination (integrated with respect to time) at a point x, y of the reproduced image, or, in telephotography, the integrated exposure of the recording film at this point. This illumination may be expressed as a double Fourier series, similar to equation (7) (but primed subscripts will be used to distinguish them from those of that equation).

$$E_2(x, y) = \sum_{m'=-\infty}^{+\infty} \sum_{n'=-\infty}^{+\infty} B_{m'n'} \exp i\pi \left(\frac{m'x}{a} + \frac{n'y}{b} \right), \qquad (26)$$

where

$$B_{m'n'} = \frac{1}{4ab} \int_{-b}^{+b} \int_{-a}^{+a} E_2(x, y) \exp -i\pi \left(\frac{m'x}{a} + \frac{n'y}{b} \right) dxdy. \qquad (27)$$

Reproduction of detail in the image may be studied by comparing these components with the corresponding ones of the original image.

The apparent illumination is the same as if the aperture traced a single strip across repeated fields in the xy plane as illustrated in Fig. 14, and all of the repeated fields included between $y = -b$ and

$y = +b$ were cut out and superposed to form the image. Let $E_3(x, y)$ be the illumination of this strip. Then, since the exponential

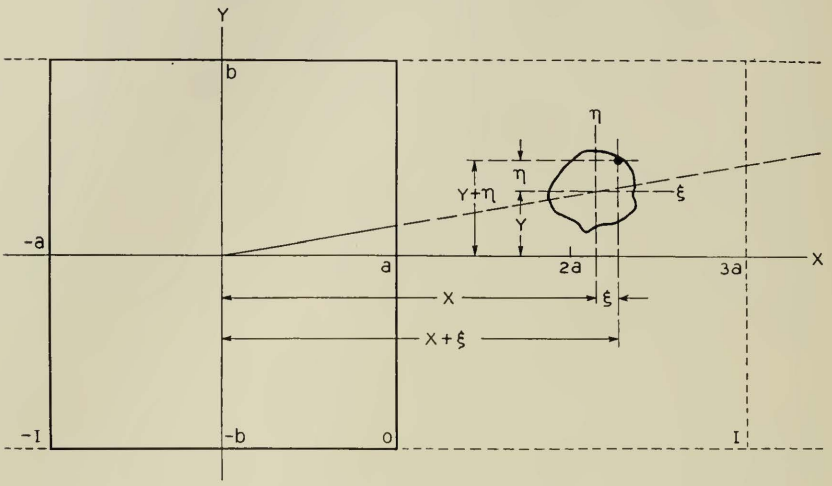


Fig. 14—Analysis of received picture.

factor of the integrand in equation (27) is periodic in x and identically reproduced in each of the fields, the integral is equal to

$$B_{m'n'} = \frac{1}{4ab}\int_{-b}^{+b}\int_{-\infty}^{+\infty} E_3(x, y) \exp -i\pi\left(\frac{m'x}{a} + \frac{n'y}{b}\right) dxdy. \quad (28)$$

The limits $-b$ to $+b$ in y and the infinite limits in x may be used because the illumination is zero everywhere outside of the strip.

Again taking a coordinate system $\xi\eta$ fixed with respect to the aperture, such that

$$x = \xi + ut, \qquad y = \eta + vt \quad (29)$$

the instantaneous illumination of any point covered by the aperture is, neglecting constant factors,

$$T_2(\xi, \eta)I(t) = T_2(x - ut, y - vt)I(t). \quad (30)$$

The total illumination of any point xy in the image strip is thus

$$E_3(x, y) = \int_{-\infty}^{+\infty} T_2(x - ut, y - vt)I(t)dt.$$

Substitution in integral (28) and a change in the order of integration

gives

$$B_{m'n'} = \frac{1}{4ab}\int_{-\infty}^{+\infty}\int_{-b}^{+b}\int_{-\infty}^{+\infty} T_2(x - ut, y - vt)I(t) \cdot \exp -i\pi\left(\frac{m'x}{a} + \frac{n'y}{b}\right) dxdydt. \quad (31)$$

Changing to the $\xi\eta$ system

$$B_{m'n'} = \frac{1}{4ab}\int_{-\infty}^{+\infty}\int_{-b-vt}^{b-vt}\int_{-\infty}^{+\infty} T_2(\xi, \eta)I(t) \exp -i\pi\left(\frac{m'u}{a} + \frac{n'v}{b}\right) t \cdot \exp -i\pi\left(\frac{m'\xi}{a} + \frac{n'\eta}{b}\right) d\xi d\eta dt. \quad (32)$$

This integral may be considered as the surface integral of a function $\varphi(\eta, t)$ taken over a strip shaped area shown in Fig. 15, in elements of

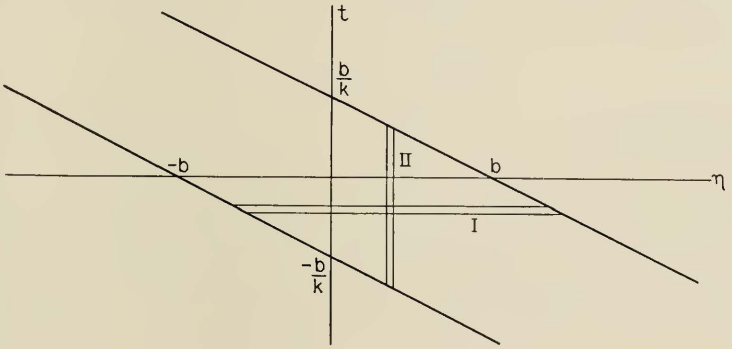


Fig. 15—Equivalent integration regions.

the type indicated as I. From this it may be seen that, when also integrated in elements of the type indicated as II,

$$\int_{-\infty}^{+\infty}\int_{-b-vt}^{b-vt}\varphi(\eta, t)d\eta dt = \int_{-\infty}^{+\infty}\int_{(-b-\eta)/v}^{(b-\eta)/v}\varphi(\eta, t)dtd\eta. \quad (33)$$

Consequently

$$B_{m'n'} = \frac{1}{4ab}\int_{-\infty}^{+\infty}\int_{(-b-\eta)/v}^{(b-\eta)/v}\int_{-\infty}^{+\infty} T_2(\xi, \eta)I(t) \exp -i\pi\left(\frac{m'u}{a} + \frac{n'v}{b}\right) t \cdot \exp -i\pi\left(\frac{m'\xi}{a} + \frac{n'\eta}{b}\right) d\xi dtd\eta. \quad (34)$$

Consider now the intensity $B_{m'n'}$ of a final reproduced picture component m', n' resulting from a single component m, n in the signal as

expressed by equation (24). The integral becomes

$$B_{m'n'} = \frac{A_{mn} Y_1(m, n)}{4ab} \int_{-\infty}^{+\infty} \int_{(-b-\eta)/v}^{(b-\eta)/v} \int_{-\infty}^{+\infty} T_2(\xi, \eta) \cdot \exp i\pi \left(\frac{m - m'}{a} u + \frac{n - n'}{b} v \right) t \cdot \exp -i\pi \left(\frac{m'\xi}{a} + \frac{n'\eta}{b} \right) d\xi dt d\eta. \quad (35)$$

It will be noted that the exponential function of t is periodic in t, one of the periods being $t_0 = 2b/v$. Furthermore, this is just the difference between the upper and lower limits in t. Hence the integral in t may be written

$$I = \int_0^{t_0} \exp i\pi \left(\frac{m - m'}{a} u + \frac{n - n'}{b} v \right) t dt.$$

This integral is zero except when

$$\frac{m - m'}{a} u + \frac{n - n'}{b} v = 0, \quad (36)$$

in which case

$$I = t_0. \quad (37)$$

The meaning of these last few equations is clear. It is, as would be expected, that a signal component m, n does not give rise to all components m', n' in the final received picture, but that these latter components are in general zero unless m' and n' satisfy a definite relationship with m and n, expressed by equation (36). A somewhat unexpected result is, however, that equation (36) allows some other m', n' components besides the normal one for which $m' = m$ and $n' = n$. That is to say, a given signal component m, n in the line will reproduce in the final picture not only a corresponding m, n component, but as has been foreshadowed in the discussion on confusion in the signal, it will also reproduce certain other components with different indices.

Let us consider first, however, the reproduction of the normal component for which $m' = m$ and $n' = n$, which is obviously allowed by equation (36). The amplitude B_{mn} is then, neglecting constant factors,[6]

$$B_{mn} = A_{mn} Y_1(m, n) Y_2(m, n), \quad (38)$$

where

$$Y_2(m, n) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} T_2(\xi, \eta) \exp -i\pi \left(\frac{m\xi}{a} + \frac{n\eta}{b} \right) d\xi d\eta. \quad (38')$$

[6] The constant factor neglected as compared with equation (35) is $t_0/(4ab)$. The t_0 is the period of image repetitions (or "frame period"). It appears here because the brightness of a single image depends on how quickly it is reproduced.

The quantity Y_2 it will be noted is almost the same, for the receiving aperture, as the Y_1 is in equation (20′) for the sending aperture. Thus, on the normally reproduced component the receiving aperture merely adds whatever filtering action it has to that which has already been caused by the sending aperture.

As noted, in addition to this normal component, the integral (35) exists in general for other values of m' and n' and thus gives rise to extraneous components in the reproduced image. If equation (36) is applied particularly to the usual system of scanning in N lines in which as in equation (12), $v = ub/(Na)$, it becomes

$$m + \frac{n}{N} = m' + \frac{n'}{N}. \tag{39}$$

For values of m' and n' satisfying equation (39), the reproduced component has the complex amplitude (neglecting constant real factors)

$$B_{m'n'} = A_{mn} Y_1(m, n) Y_2(m', n'). \tag{40}$$

Looking back at equation (14) and comparing it with equation (39) it may be seen that these components correspond in indices to the original image components that are confused in the signal to give only one signal component. The result is, therefore, after all quite reasonable from a physical point of view. For when a signal of a certain frequency is transmitted over the line the receiving apparatus has no information by which to judge which component in the original picture it is supposed to represent. So, as shown by equation (40) it impartially reproduces every one of the components it could possibly represent, each component with the intensity and phase it would have if it were really the one intended to be represented by the signal. The components are then all superimposed in the picture.

From this development it is clear that the process of scanning an image field in strips and reproducing it in a similar manner not only reproduces the components of the original image but also introduces extraneous components. The reproduced field thus consists of two superposed fields: a normal image built up from the normally reproduced components, and an additional field of extraneous components. Although not really independent, it is convenient to consider these two fields as existing separately, and thus to think of the normal image field as having an extraneous field superposed on it.

Considering the normal field alone, we may term the reproduction of its detail as the *reproduction of* NORMAL *detail.* There is a loss in such reproduction, for both the transmitting and receiving apertures intro-

duce a relative loss in the reproduction of the shorter wave components. Consequently there is a loss of definition in the finer grained details of the normal image. This type of distortion due to aperture loss may be termed *simple omission of detail.*

In addition to the simple omission of detail, the normal image is masked by the presence of the extraneous field. The more pronounced features of this field are the line structure and serrated edges that it superposes on the normal image. Its presence is not only displeasing, but it also masks the normal image components and thus results in a further loss of useful detail. This type of loss may be termed a *masking of detail* or a *masking loss.* It is true that the extraneous components may sometimes give rise to an illusory increase in resolution across the direction of scanning in special cases where they add on to the diminished normal components in just the right phase and magnitude to bring the latter back to their phases and intensities in the original image, giving no resultant distortion whatever. (In all such cases, however, to obtain this benefit it is necessary to effect a quite accurate register between the original image and the scanning lines or the distortion is very large. Such accurate registering is generally impractical and may be definitely impossible if the registry required for one portion of the image conflicts with that required in another portion. Such cases may, therefore, in general be disregarded.)

The Reproduction of Normal Detail

The preceding theory permits a numerical calculation of the reproduction of detail in the normal image. This is given directly by equation (38) above.

In order to make some of the discussion in the following pages more concrete and specific the sending and receiving apertures will be taken alike; this condition, therefore, gives $[Y(m, n)]^2$ as a measure of how well the various components are reproduced. If a picture be assumed in which all the original components have the same amplitude then $[Y(m, n)]^2$ is the amplitude of the reproduced normal components.

The relative admittance for any given pair of apertures may be calculated from equations (20′) or (38′). Such calculations have been made for various apertures and the results summarized in Appendix II.

The admittance of an aperture is not in general uniquely determined by the wave length of a component, but also depends on the orientation of the component with respect to the aperture. The admittances of reasonably shaped apertures do, however, decrease in general with increasing numerical values of the indices m and n; and the shorter wave components are, therefore, in general, less faithfully reproduced than the longer wave ones.

A circular aperture furnishes a simple example of such reproduction—because its admittance, from its symmetrical shape, is a unique function of the wave length of a component. In other words a circular aperture reproduces normal detail equally well in all directions. We may, therefore, simply plot $[Y(m, n)]^2$ as a function of the component wave length as in Fig. 16, and this single curve is a measure of how

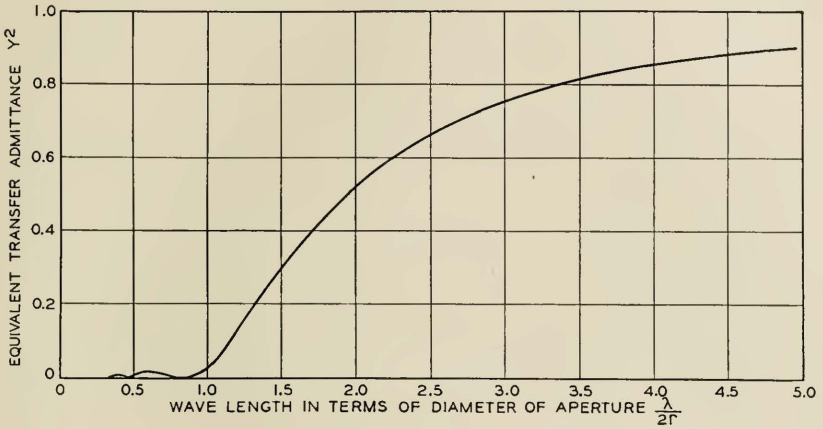


Fig. 16—Equivalent transfer admittance for circular apertures at both sending and receiving ends, *vs.* wavelength.

well the various normal components are reproduced. The shorter wave components are practically omitted in the reproduction of an image.

Other apertures do not reproduce normal image detail equally well in all directions because their admittances depend on the slope of a component. To simplify the consideration of such apertures we may resort to a practice commonly used in discussing telephotographic or television systems, and that is, we may take the resolution along the direction of scanning and across the direction of scanning separately as criteria of their performance.

Neglecting the small slope of scanning lines with respect to the x axis of the image field, the admittance of an aperture for components normal to the direction of scanning is $Y(m, 0)$. Consequently, we may take $[Y(m, 0)]^2$ as a measure of the reproduction of normal detail along the direction of scanning. In a similar manner we may take $[Y(0, n)]^2$ as a measure of the reproduction of normal detail across the direction of scanning.

It thus follows that an aperture gives the same resolution of normal detail along the direction of scanning and across the direction of scan-

ning when the two admittances $Y(m, 0)$ and $Y(0, n)$ are substantially equal for components of the same wave length over the useful range. Circular apertures, square apertures and other apertures that are suitably symmetrical fulfill this condition exactly, and consequently give equal resolution of normal detail in the two directions.

The curve $[Y(0, n)]^2$ has been plotted, by way of illustration for a rectangular aperture, in the middle line of Fig. 17.

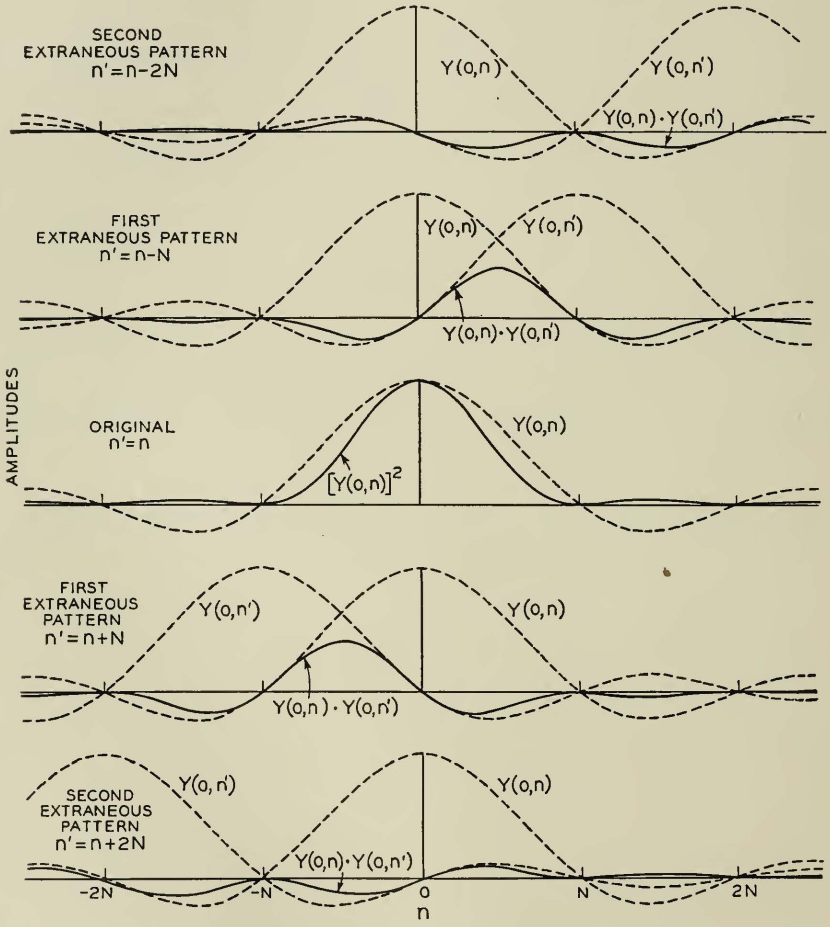


Fig. 17—Reproduction of original and extraneous patterns.

It may be noted incidentally that the simple omission of detail which occurs in the reproduction of normal components is quite similar to the loss of resolution that an image suffers when it is reproduced through

an imperfect optical system. Specifically the effect of a sending or receiving circular aperture alone, or $Y(m, n)$, is the same as that caused by an optical system which reproduces a mathematical point in the original as a circle of uniform illumination (circle of confusion) in the image of the same size (with respect to the image) as the scanning aperture. The effect of the two apertures in tandem, or $[Y(m, n)]^2$, may be very closely simulated by a circle of confusion of about twice the area of either aperture, as can be judged from the discussion which has been given above regarding the curves in Fig. 12.

The Extraneous Components

It will be clearly understood that the discussion immediately preceding has been confined entirely to the normal image components, that is, to the image that would be seen if no extraneous components were present. In particular, it should be clear that the reproduction of normal detail equally well in the direction of scanning and across the direction of scanning does not mean that the details of the total resultant image will be seen equally well in the two directions, for the extraneous components will to a certain extent mask the normal image.

In the same manner as for the normally reproduced components, the amplitudes of the extraneous components, according to the preceding theory, are given by equation (40) above, where

$$m' = m + \mu,$$
$$n' = n - \mu N,$$

where

$$\mu = \text{an integer} = m' - m = (1/N)(n - n'). \tag{41}$$

The composite transfer admittance $Y(m, n) \cdot Y(m', n')$ may therefore be taken as a measure of the extent to which the extraneous components are introduced. If a picture be assumed in which all the original components have the same amplitude then $Y(m, n) \cdot Y(m', n')$ is the amplitude of the extraneous components.

A given original component of indices m, n gives rise to a whole series of extraneous components, m', n', as μ ranges from 1 up through the positive integers and -1 down through the negative integers. As an illustration we have plotted the case of a rectangular aperture of a width just equal to the scanning pitch, in Fig. 17, which has just been referred to in considering the normal components. The two lines marked "first extraneous pattern" show the relative amplitudes for μ equal to 1 and -1, respectively, and those marked "second extran-

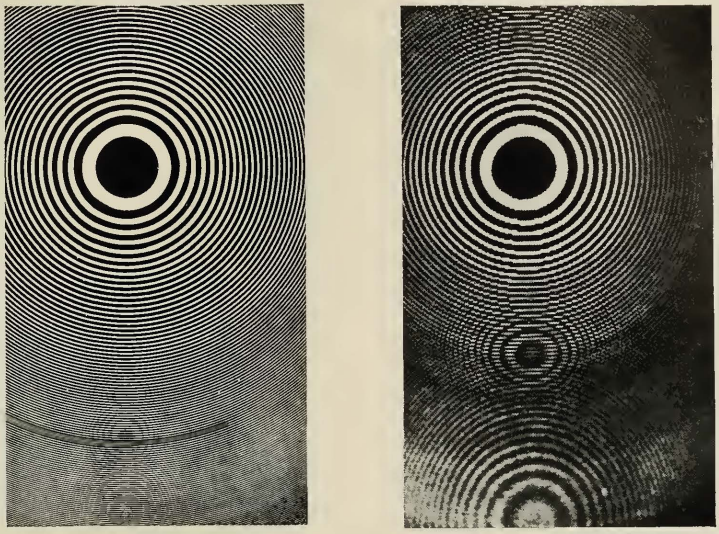
eous pattern," for μ equal to 2 and -2, respectively, for $m = 0$. (The shift from $m' = 0$ to $m' = \pm 1$ and ± 2 has been ignored since if N is at all large this has a negligible effect on $Y(m', n')$, as may be noted from Fig. 13.) An examination of Fig. 17 and a consideration of the nature of $Y(m, n)$ and $Y(m', n')$ shows that the principal interference effect will come from the pattern for which $|\mu| = 1$, and that the relative amplitudes become very small as μ increases in absolute magnitude. In general, therefore, only the first extraneous pattern may be considered as of really serious importance. Considering this pattern in Fig. 17 it will be seen that the amplitude $Y(0, n)\ Y(0, n')$ increases as $|n'|$ increases from zero, the extraneous components becoming more and more comparable to the normal components. At $N/2$, both components are of the same amplitude, and the extraneous components are therefore masking the normal components. It will be noted that the index region at $N/2$ corresponds to the centers of the relatively empty regions in the frequency spectrum of the signal. The large masking effect caused by the extraneous components explains why such small signal energy as exists in these regions is almost completely incapable of transmitting any useful image detail.

It will be noted that the components with values of $|n'|$ in the neighborhood of $N/2$ and greater are in general almost parallel to the direction of scanning. The masking loss will therefore be greatest across the direction of scanning and practically negligible along the direction of scanning. This is quite reasonable because the extraneous components constitute the line structure of the reproduced image, and should therefore cause the greatest loss of detail across the direction of scanning.

For clarity in the explanation up to this point, masking loss has been discussed as if an extraneous component could only mask the normal component with which its indices happened to coincide. In reality the masking is of a more serious nature. An extraneous component undoubtedly obscures any normal component that has about the same wave length and the same slope across the field even though it does not exactly coincide in these characteristics.

More detailed curves than Fig. 17, showing the amplitudes of the extraneous components have been prepared in Appendix II. These also show the results for other index values of m than zero, and for other than the simple rectangular aperture. The results indicate that the extraneous patterns diminish in intensity progressively as more overlap is tolerated between adjacent scanning lines, at the expense, of course, of increased aperture loss for the normal components. This point will be taken up again below.

The reality of these extraneous components is strikingly demonstrated in Fig. 18, for which we are indebted to Mr. E. F. Kingsbury.



a. Original. *b*. Transmitted.

Fig. 18—Fresnel zone plate.

This shows at (*a*) the original of a Fresnel zone plate and at (*b*) the picture after transmission through a telephotographic system. The first extraneous pattern is very prominent in the lower corner of (*b*) and a detailed study of the slope and spacing of the extraneous striations shows them to be in exact accord with the theory which has been given.[7]

The special case of the extraneous components which are formed when the original consists of a flat field is of some interest due to the high visibility of these components under such a condition. This scanning line structure is quite familiar as an imperfection in many pictures

[7] The extraneous pattern, although it is (and should be according to the theory) very nearly a transposed reproduction of the original pattern, must not be confused with a long delayed echo of that original pattern. In other words, if only the lower half instead of the whole of (*a*) had been transmitted, the lower half of (*b*) would still have been exactly as it is, the extraneous components being generated entirely irrespective of whether components representing a similar configuration exist in other portions of the original or not.

In the region about half-way between the centers of the normal picture and the first extraneous picture the resulting pattern gives very much the appearance of another set of extraneous components. It is not such, however, that successive rings are not really bright and dark, as they would be in the case of a genuine extraneous component, but alternating uniform gray and striped black and white, so that the average intensity along the circumference of a ring is independent of the diameter of the ring, except for some photographic non-linearity.

transmitted by telephotography and television. It can be removed only by insuring that $Y(m', n')$ shall vanish whenever $n' = N$, so that

$$Y(0, 0) \cdot Y(\mu, \mu N) = 0.$$

The requirement can be met for the elementary shapes of apertures A, E and F of Fig. 12, but cannot be met in the others. In these other cases the overlap between adjacent scanning lines is usually adjusted so that the requirement is met for $\mu = \pm 1$, to remove the most serious pattern. Thus, for example, for the circular aperture B this requires an overlap of around 25 per cent.

The Reproduction of Detail

In optical instruments the reproduction of detail is usually measured by what is called the "resolving power" which in turn is defined from the smallest separation between two mathematical point (or parallel line) sources of light in the original which can be distinguished as double in the reproduced image.

For the present it is perhaps simpler to consider another criterion of the resolving power, namely, the shortest element length in an image resembling a telegraph signal, used as an original, which can be recognizably reproduced with certainty in the received picture. For reasons that have already been mentioned above it is necessary to insist that the received picture be recognizable with certainty without any registry requirement between the original image and the scanning lines.

Using this criterion for the resolution along the direction of scanning and assuming the apertures at the sending and receiving ends to be rectangular and of the same length with respect to the picture size, the minimum signal element required for a recognizable picture (as set by the apertures as distinguished from the electrical transmission circuits) will be of about the length of either aperture. For other shapes of aperture the minimum element length will be very nearly the length of the equivalent rectangular aperture using the term "equivalent" in the same sense that it was used in the discussion regarding Fig. 12.

According to the same criterion, for the resolution across the direction of scanning the minimum element length required for recognizable transmission, in the case of a rectangular aperture of width equal to the scanning pitch, will be twice the scanning pitch. It will be noted that this is twice the length which would be required if only the normal image components were reproduced, and this difference may be considered as a measure of the degradation caused by the masking effect of the extraneous components for this arrangement of apertures.

This figure for the degradation must be taken with a certain reserve, partly because the exact telegraph theory for the criterion of resolution considered has really been inferred rather than presented in complete logical form, and partly because the figure may be expected to vary according to the criterion of resolution chosen. Some rough studies have indicated the degradation to be materially less if the more conventional criterion of resolution (two parallel line sources of light) were used.

This degradation may be estimated in another manner. In Appendix II the extraneous components have been computed for a variety of apertures and degrees of overlap between adjacent scanning lines. In Fig. 19 there have been plotted the maximum amplitudes of these

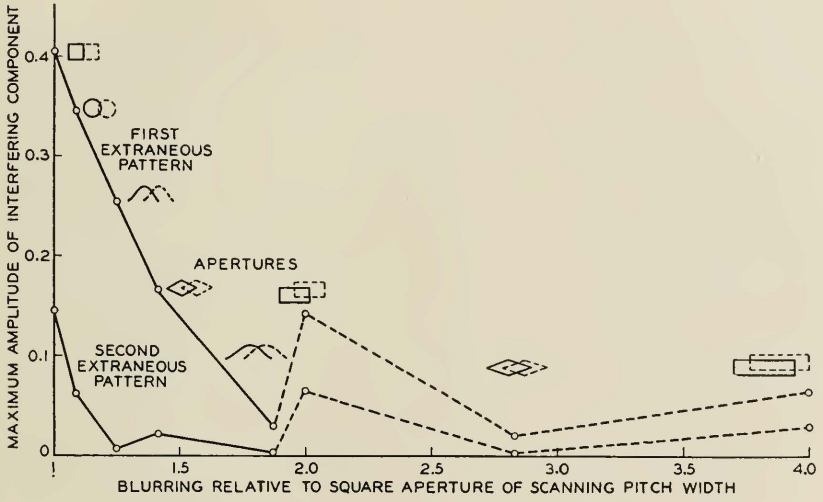


Fig. 19—Magnitude of extraneous components as a function of resolution.

extraneous components in each case (the first and second extraneous patterns being plotted separately) as a function of the relative coarseness of resolution for the normal image alone. This latter quantity is taken relative to a rectangular aperture of width equal to the scanning pitch, and, for example, for a rectangular aperture of width equal to twice the scanning pitch, is represented by the figure 2. For convenience, above the various points have been inserted small diagrammatic representations of the corresponding apertures. Also for convenience the points have been arbitrarily connected together.

From inspection of Fig. 19 several conclusions may be drawn, namely,

1. Considering apertures of a given shape, the more overlap allowed between adjacent scanning lines the weaker will be the extraneous patterns but the coarser will be the reproducible detail in the normal image.

2. Not all shapes of aperture are equally efficient in suppressing extraneous components, and at the same time retaining a given resolution of normal detail. Of the shapes considered, the rectangular aperture is least efficient in this respect, and the full-wave sinusoidal aperture (E in Fig. 12), is the most efficient.

3. Although not proved, it may be inferred from the figure that the finest resolution in the normal image that can be obtained (assuming a given scanning pitch) without showing a first order extraneous pattern on a flat field, is that obtained with the rectangular aperture of width equal to the scanning pitch.

4. With the most suitable aperture it is possible practically to suppress the extraneous components, at the expense of coarsening the normal reproducible detail to slightly under twice that given by the rectangular aperture just mentioned.

The last point in particular enables us to draw a conclusion in regard to the degradation contributed by the extraneous components. For a rectangular aperture of width equal to the scanning pitch it appears that the degradation amounts to a little less than doubling the coarseness of resolution to normal detail. This substantially checks the estimate which has already been made above. It may further be surmised for all the other shapes of aperture shown with a value of abscissa under 2 that as the degradation contributed by the extraneous components is reduced, the coarseness of resolution to normal detail is increased to just about make up for this, and that in the overall picture the minimum element length which can be recognizably reproduced remains substantially constant at about twice the scanning pitch.[8] For aperture arrangements with values of abscissa over 2, either the inefficiency in suppressing extraneous components, or the unnecessarily large overlap, tends to coarsen the overall resolution to a minimum elementary length greater than twice the scanning pitch. In this region the line connecting the points has been dotted.

[8] It may very well be that even if all these aperture arrangements transmit an about equal amount of information they do not give the same psychological satisfaction to the viewer at the receiving end. The general effect of a square aperture of scanning pitch width is to give a "snappy" appearance, disturbed, however, by the presence of the extraneous patterns. When these are removed, keeping the overall resolution about the same, the appearance becomes "woolly" or "fuzzy."

An Estimate of the Idle Frequency Regions

As mentioned at the beginning of this paper, the frequency regions between the strong bands appear to be empty when examined with a frequency analyzer of limited level range, or when a narrow band elimination filter is used in connection with visual observations of the reproduced image. These regions are not really completely empty, but do contain weak signal components as shown by the preceding theory, which are not, however, particularly useful inasmuch as, in the final result, they give rise about equally to components simulating the original picture and to masking extraneous components. The regions may, therefore, be considered as idle.

The factors determining the extent of these idle regions are too complicated to permit an exact theoretical evaluation of their width, but an estimate may be attempted from an inspection of Fig. 12 and of the curves given in Appendix II.

From Fig. 12 and the experience that along the direction of scanning the minimum recognizably transmitted elementary signal length is the length of a rectangular aperture it can be deduced that in the absence of extraneous components the useful band of an aperture extends up to the point where its relative admittance, for a single aperture, is in the neighborhood of 0.65. For two apertures in tandem the corresponding relative admittance is $0.65^2 = 0.42$.

Now in Fig. 27 of Appendix II the extraneous components are very small and may be considered negligible. According to the above criterion, therefore, the useful frequency band constitutes approximately 54 per cent of the total space. The idle frequency regions would, therefore, occupy the remainder, or 46 per cent of the total space.

Experimental examination of a television signal with a narrow band elimination filter gave the width of the idle regions as 50 to 60 per cent of the total space. This was for a field scanned with a circular aperture giving a one-quarter overlap of scanning strips. The discrepancy for a quantity so vaguely defined is not large but is probably due to incomplete utilization of even the theoretically active region by the television set because of inherent imperfections in parts of the complete system outside the scanning mechanism proper.

The width of the individual idle bands is then about half the frequency of repetition of scanning lines. For most systems of telephotography this runs in the order of magnitude of one cycle per second, making the waste regions very narrow and close together. For systems of television the waste bands come in much more significant "slices," although the same fraction of the frequency space is wasted. For ex-

ample, in a 50-line system the waste bands are each about 500 cycles wide. In a system using a single sideband of one million cycles width the waste bands are each about 3300 cycles wide.

These idle frequency regions naturally lead to the questions whether (*a*) there is any way of segregating all of the relatively useless signal components in one region of the frequency spectrum so that the useful parts of the signal may be transmitted over a channel of about half the width, or (*b*) whether it would be worth while placing other communication channels in these waste regions. It must be realized, however, that even when the complete frequency space is utilized (by any one of a number of possible schemes), the required frequency band for transmitting a picture of given detail at a given rate is still only halved as compared with the simple system considered above, which is not a change in order of magnitude. The problems of transmitting the wide band of frequencies necessary, for example, in television, while lessened, therefore still remain.

APPENDIX I

The calculation of $Y_1(m, n)$ according to equation (20′) is, for the three simple apertures here considered, a straightforward mathematical process which will therefore not be reproduced. The results are plotted in the form of charts in the conventional manner for functions of two variables, namely as a series of contours, one of the two variables being kept constant for each contour. This constant value changes progressively for each successive contour.

The variables are taken as m and n, multiplied by parameters depending on the sizes of the scanning aperture and of the picture. Because of the obvious symmetry of the function, only half of each chart has been drawn. In order to avoid confusion the contours have been dotted when $|m|$ is greater than the first root of $Y_1(m, 0) = 0$. In one case the contour is shown in a dashed line when $|m|$ is equal to this root. Constant factors in the scale of ordinates have been neglected, to make $Y_1(0, 0) = 1$.

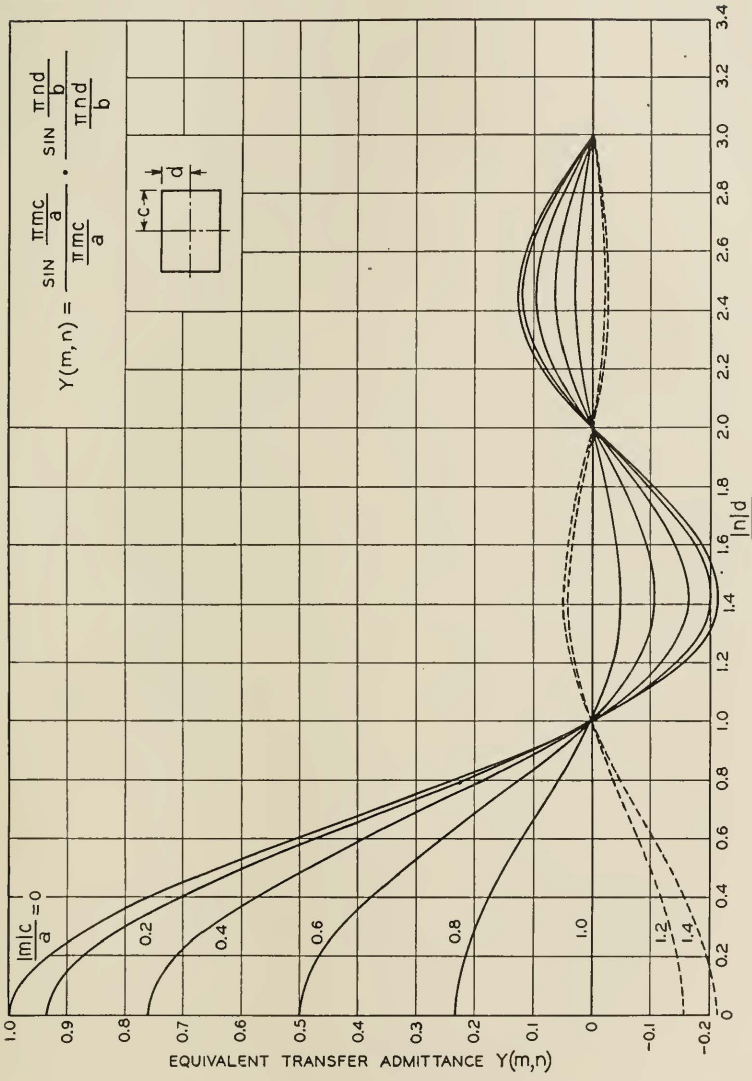


Fig. 20—Rectangular aperture.

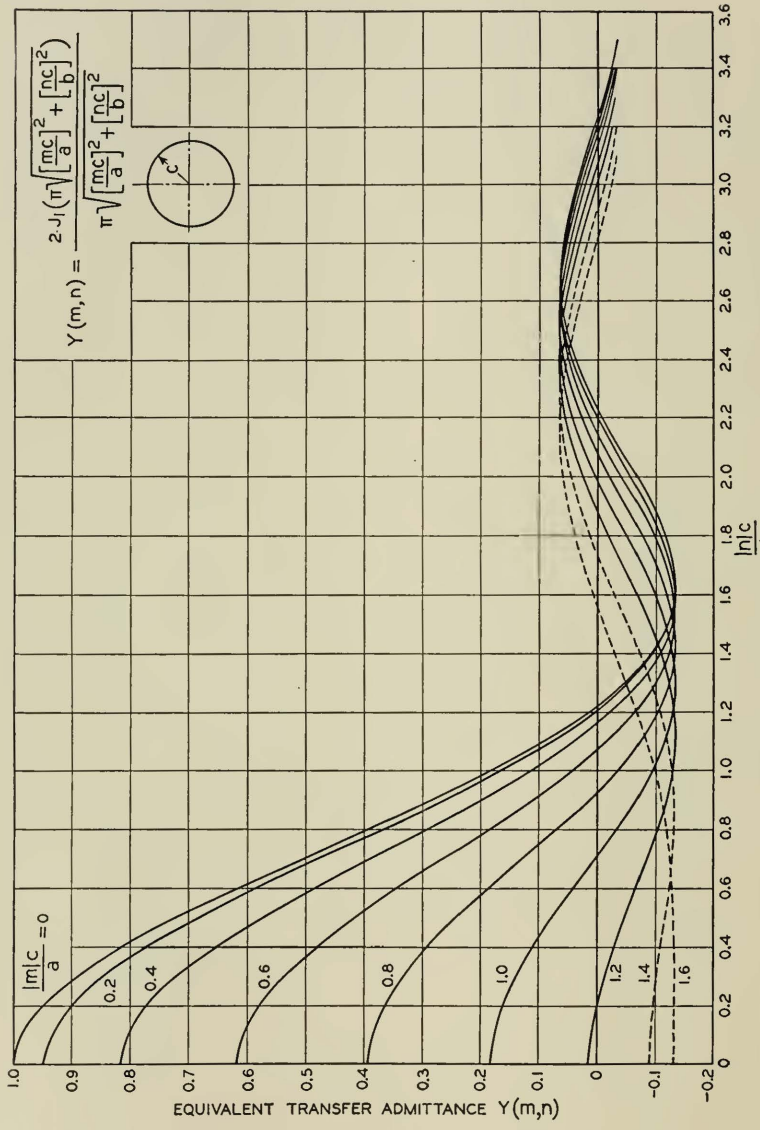


Fig. 21—Circular aperture.

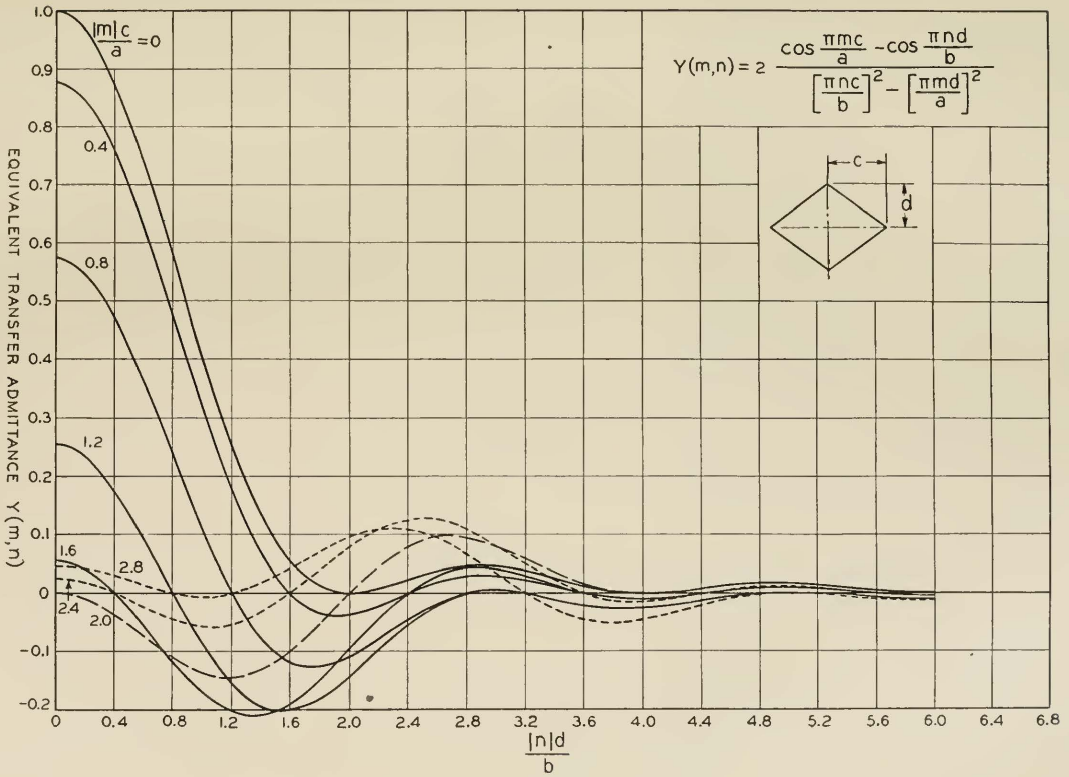


Fig. 22—Diamond shaped aperture.

APPENDIX II

As in the case of Appendix I, the calculation of $Y_1(m, n) \cdot Y_2(m', n')$ is a straightforward mathematical procedure which will not be reproduced. The results are again presented in the form of charts.

In these charts the intensities of the principal extraneous components are indicated by solid lines, while the higher order extraneous components are indicated by dotted lines. The normal components have been indicated by dashed lines.

It should be explained that what has really been plotted is $Y_1(m, n) \cdot Y_2(m, n')$ rather than $Y_1(m, n) \cdot Y_2(m', n')$. This is because the difference between m' and m, when multiplied by the parameters chosen for the charts, varies with the proportions of the scanning system used. As discussed in the text with regard to Fig. 17, however, this difference between m' and m has a negligible effect for any system employing a useful number of scanning lines.

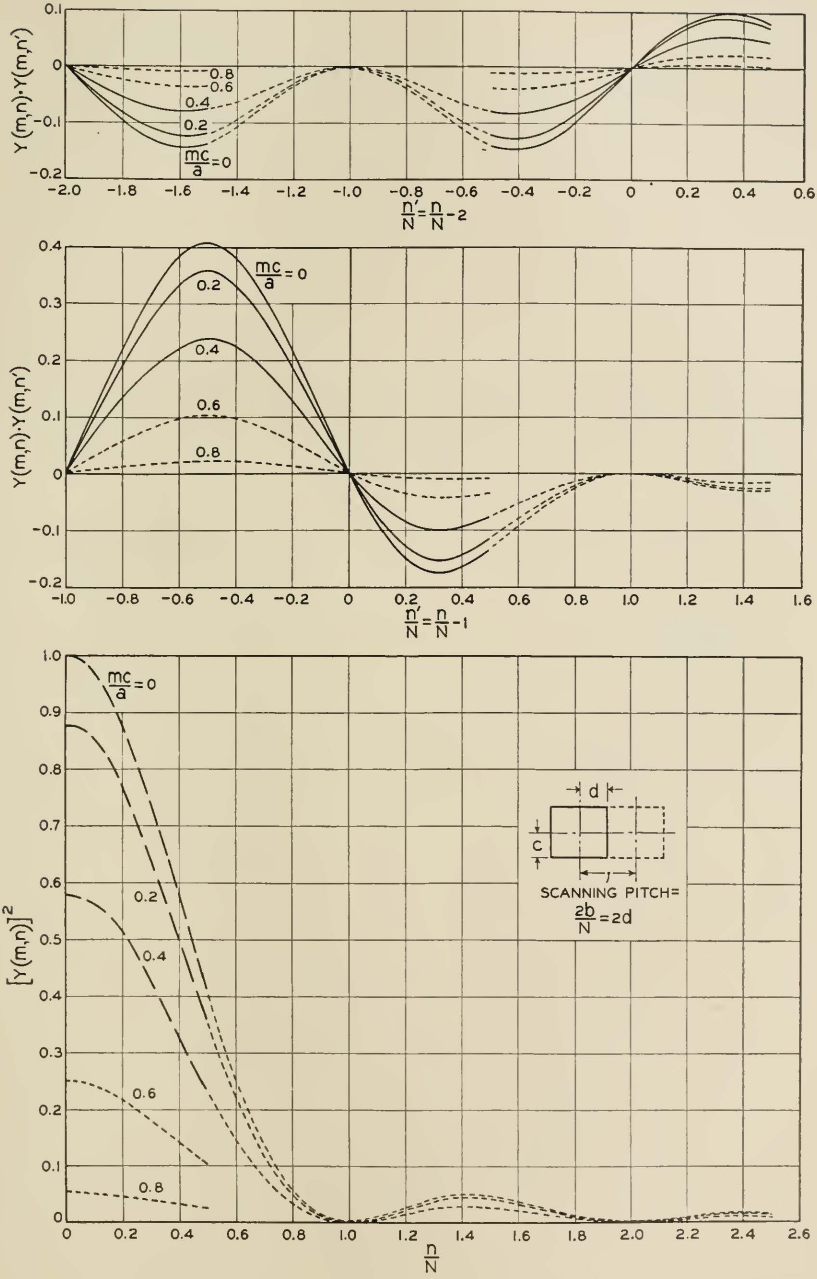


Fig. 23—Rectangular aperture with no overlap.

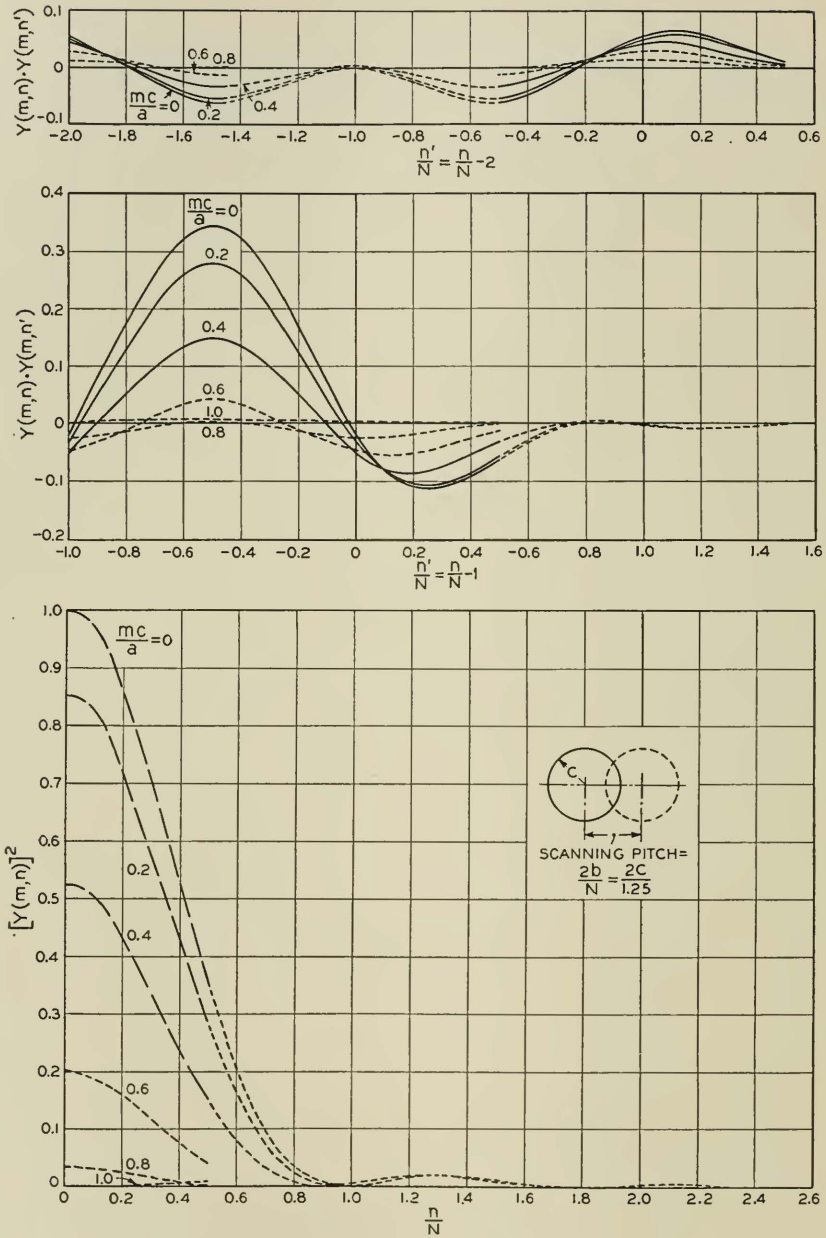


Fig. 24—Circular aperture with 25 per cent overlap.

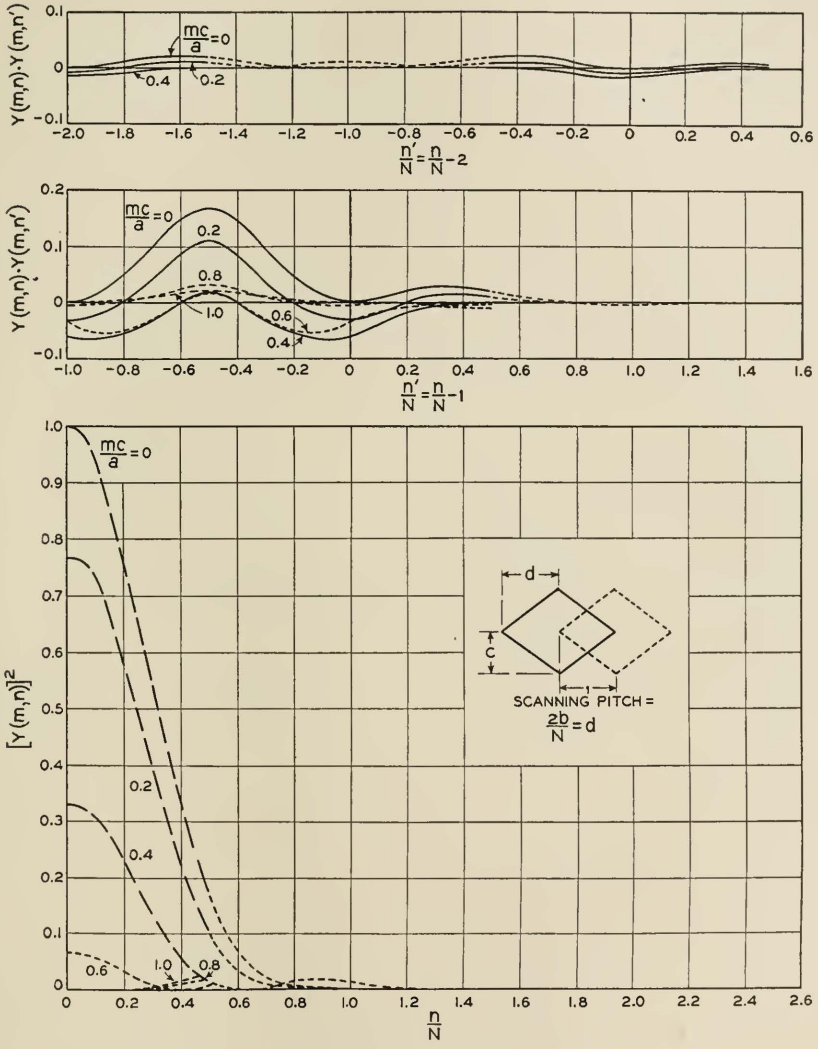


Fig. 25—Diamond shaped aperture with half diagonal overlap.

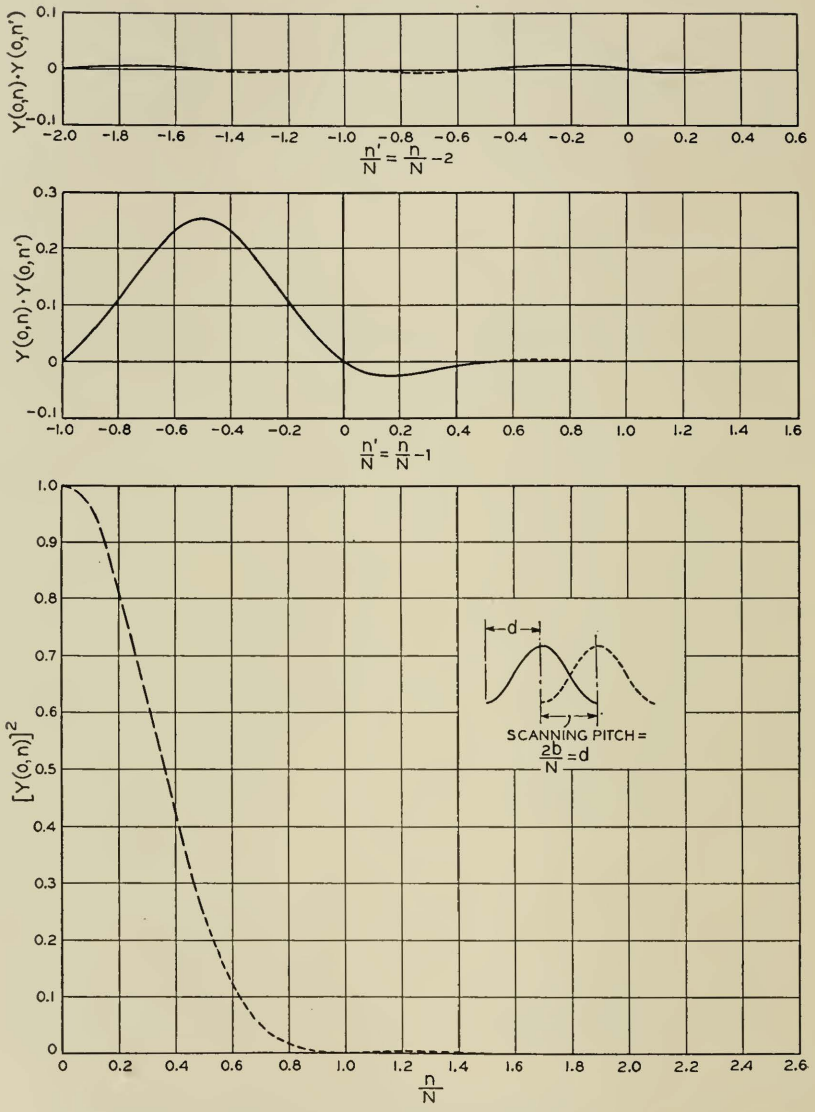


Fig. 26—Sinusoidal aperture with half wavelength overlap.

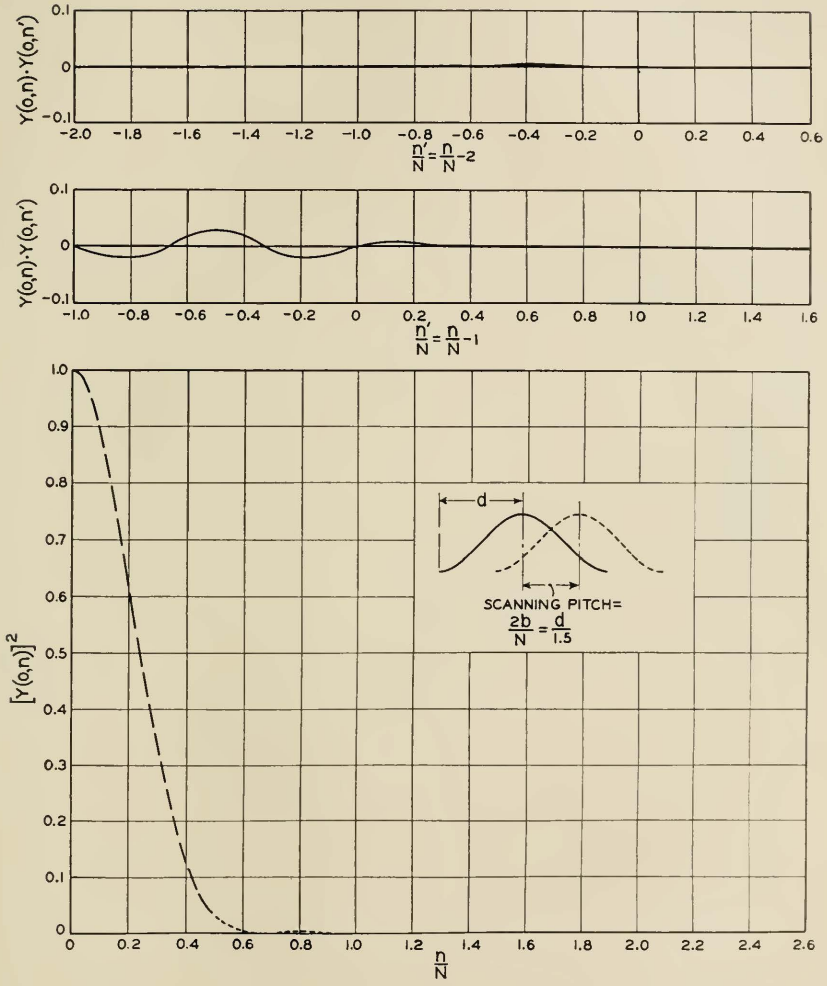


Fig. 27—Sinusoidal aperture with two-thirds wavelength overlap.

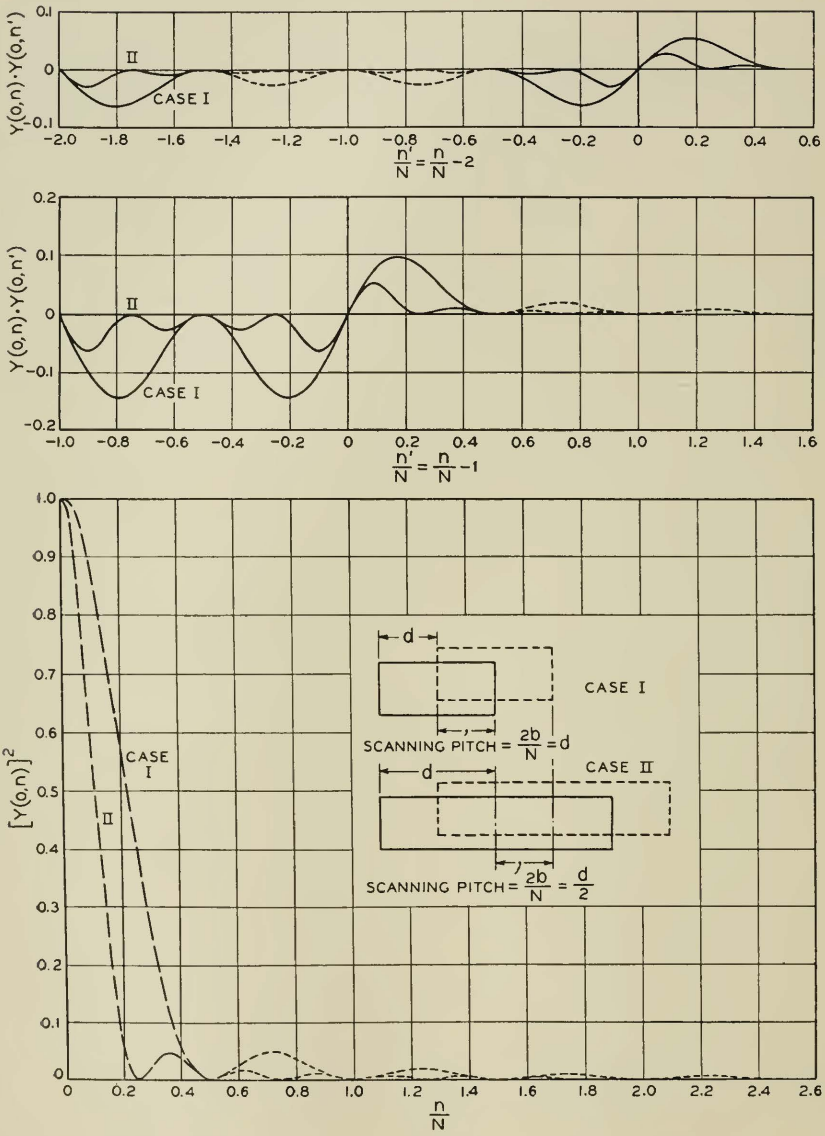


Fig. 28—Rectangular aperture with different degrees of overlap.

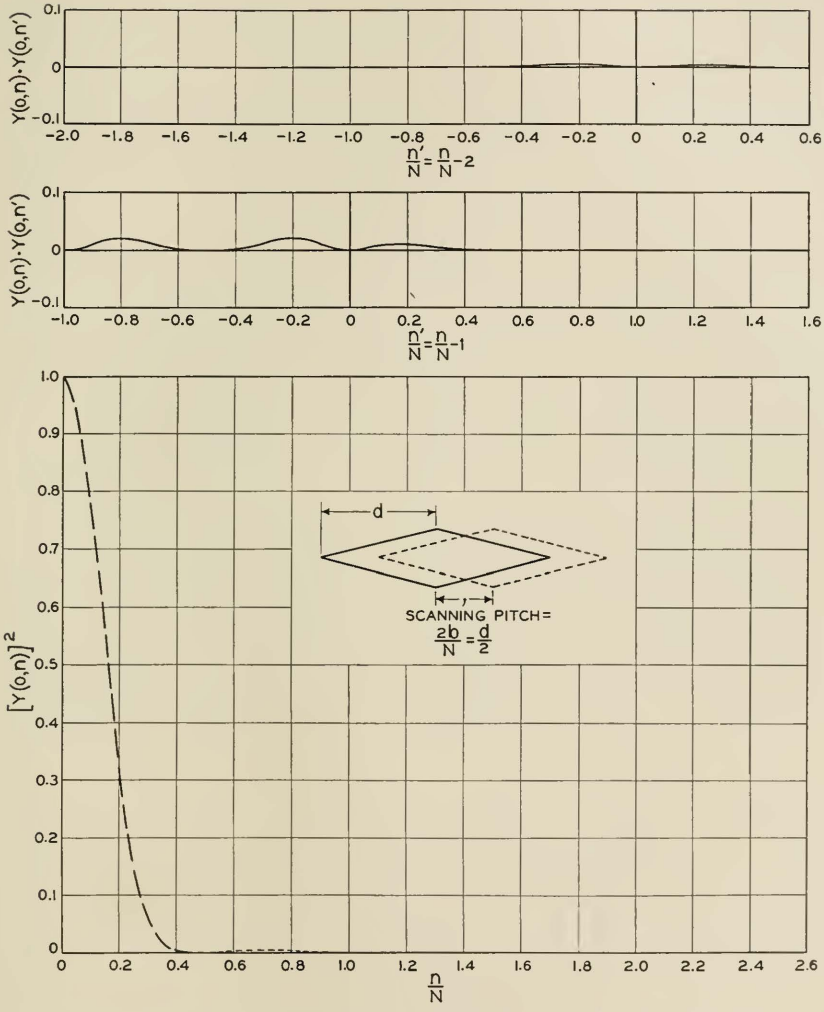


Fig. 29—Diamond shaped aperture with three-quarters diagonal overlap.

Abstracts of Technical Articles from Bell System Sources

The Thermionic Work Function and the Slope and Intercept of Richardson Plots.[1] J. A. BECKER and W. H. BRATTAIN. This article is a critical correlation of the slope and intercept of experimental Richardson lines with the quantities appearing in *theoretical* equations based on thermodynamic and statistical reasoning. The equation for *experimental* Richardson lines is $\log i - 2 \log T = \log A - b/2.3\,T$; A and b are constants characteristic of the surface, i is the electron emission current in amp./cm.^{2}, T is the temperature in degrees K, $\log A$ is the intercept and $-b/2.3$ is the slope of experimental lines. Statistical *theory* based on the Fermi-Dirac distribution of electron velocities in the metal shows that i should be given by $\log i - 2 \log T = \log U(1 - r) - w/2.3\,T$, where U is a universal constant which has the value 120 amp./cm.^{2} °K^2, r is the reflection coefficient, and w is the *work function.* A correlation of the experimental and theoretical equations shows that $b = w - T dw/dt$, and $\log A = \log U(1 - r) - (1/2.3) dw/dT$. Only when r is 0 and the work function is independent of the temperature, is it correct to say that the slope is $-w/2.3$ and that the intercept has the universal value $\log U$. But even when w is a function of T, it follows from a thermodynamic argument that the slope is given by $-h/2.3$, where the heat function h is defined by $h = (L_p/R) - (5/2)T$, L_p is the heat of vaporization per mol at constant pressure. The heat function is related to the work function by the equation $h = w - T dw/dT$.

From experimental and theoretical arguments it is deduced that the reflection coefficient is probably negligibly small. Hence we conclude that *for most surfaces the work function varies with temperature*, since the intercepts of Richardson lines are rarely equal to log 120. This conclusion is to be expected since on Sommerfeld's theory, w depends on the number of free electrons or atoms per cm.^{3}, which in turn varies with temperature due to thermal expansion.

The photoelectric work function should equal the thermionic work function but should not in general be equal to -2.3 times the slope of the Richardson line. The Volta potential between two surfaces having work functions w_1 and w_2 should equal $(w_1 - w_2)k/e$ rather than $2.3k/e$ times the difference between the slopes of the Richardson lines for the two surfaces. The data from photoelectric and Volta potential meas-

[1] *Phys. Rev.*, May 15, 1934.

urements support the conclusion that the work function depends on temperature.

Fundamental Concepts in the Theory of Probability.[2] THORNTON C. FRY. Three commonly accepted definitions of the word "probability" are discussed critically, with regard both to logical soundness and to practical utility. Two major theses are presented: first, that each definition has utilitarian merits which render it especially valuable within its own field; second, that the objection of logical redundancy which is so frequently raised against the Laplacian definition can equally well be raised against the other two definitions.

Wide-Range Recording.[3] F. L. HOPPER. The recent improvements in sound quality resulting from the extension of the frequency and intensity ranges are the results of coordinated activity in recording equipment and processes, reproducing equipment, and theater acoustics. This paper discusses the recording phase of the process. A wide-range recording channel consists essentially of the moving-coil microphone, suitable amplifiers, a new recording lens, and certain electrical networks.

The characteristics of such a system, from the microphone to and including the processed film, are shown. Other factors fundamentally associated with wide-range recording, such as monitoring, film processing, the selection of takes in the review room, and re-recording, are also discussed. The changes brought about by this system of recording result, first, in a greater freedom of expression and action on the part of the actor; and, second, in a much greater degree of naturalness and fidelity than has been previously achieved.

Iron Shielding for Telephone Cables.[4] H. R. MOORE. Voltages of fundamental and harmonic frequencies, induced along communication cables by neighboring power or electric railway systems, can be reduced by the electromagnetic shielding action of the sheath, if this is grounded continuously or at the ends of the exposure. The shielding, particularly at the fundamental frequency, is improved greatly by the provision of a steel tape armor, while a surrounding iron pipe conduit effects a very great improvement at both the fundamental frequency and the higher harmonics.

This paper presents methods for the quantitative prediction of the shielding, expressed by a "shield factor" or the fraction to which

[2] *American Mathematical Monthly*, April, 1934.
[3] *Jour. S. M. P. E.*, April, 1934.
[4] *Electrical Engineering*, February, 1934.

a disturbing voltage is reduced. Necessary impedance data are given for numerous iron-surrounded cable constructions and working charts are supplied for the convenient determination of the shielding obtainable with commercially available steel tape armored cables.

On the basis of data presented in this paper, prediction of the shielding to be obtained from steel tape armored cable sheaths or those inclosed in iron pipes is concluded to be both feasible and practical. With internal impedances measurable on short length samples of a chosen construction, the accuracy of prediction is limited principally by the precision to which the disturbing field and the grounding resistances of the cable sheath may be determined. Either of the constructions discussed is capable of effecting a high order of shielding against low frequency induction and practically complete protection from harmonic disturbances. Field observations on installed cables, both tape armored and in pipe conduit, have verified the computational methods presented.

Propagation of High-Frequency Currents in Ground Return Circuits.[5] W. H. Wise. The electric field parallel to a ground return circuit is calculated without assuming that the frequency is so low that polarization currents in the ground may be neglected. It is found that the polarization currents may be included by replacing the r in Carson's well-known formulas by $r\sqrt{li(\epsilon - 1)/2c\lambda\sigma}$. The problem to be solved is that of calculating the electric field parallel to an alternating current flowing in a straight, infinitely long wire placed above and parallel to a plane homogeneous earth. Carson's derivation of this field is based on three restricting assumptions: (1) The ground permeability is unity; (2) the wave is propagated with the velocity of light and without attenuation; (3) the frequency is so low that polarization currents may be neglected. The first of these restrictions is usually of no consequence and the formula would be quite complicated if the permeability were not made unity. As pointed out in a later paper by Carson, the second restriction amounts merely to assuming reasonably efficient transmission. The effect of the third restriction begins to be noticeable at about 60 kilocycles. The object of the present paper is the removal of the third restriction.

Acoustical Requirements for Wide-Range Reproduction of Sound.[6] S. K. Wolf. The extension of the frequency and volume ranges in recording and reproducing sound has aroused a greater and more critical

[5] *Proc. I. R. E.*, April, 1934.

[6] *Jour. S. M. P. E.*, April, 1934.

consciousness of the importance of theater acoustics. It follows that higher fidelity in reproduction excites greater intolerance of the needless distortion caused by poor acoustics of the theater. To cope with the new situation, engineers have developed new instruments for acoustical analysis, which provide greater precision and facility in detecting defects and in determining the necessary corrections.

In addition to instrumental developments there have been concurrent advances in acoustical theory and practice. The result is that the more stringent requirements imposed on the acoustics of the theater by the enlarged frequency and volume ranges can be fulfilled adequately and practically. The paper discusses the requirements and describes some of the available methods for complying with them.

Contributors to this Issue

C. B. Aiken, B.S., Tulane University, 1923; M.S. in Electrical Communication Engineering, Harvard University, 1924; M.A. in Physics, 1925; Ph.D., 1933. Geophysical research and exploration with Mason, Slichter and Hay, Madison, Wisconsin, 1926–28. Bell Telephone Laboratories, 1928–. Dr. Aiken has been engaged in work on aircraft communication equipment, broadcast receiver design, centralized radio systems and common frequency broadcasting.

A. W. Clement, B.S. in Electrical Engineering, University of Washington, 1925; M.A., Columbia University, 1929. Bell Telephone Laboratories, Apparatus Development Department, 1925–. Mr. Clement has been engaged in the development of various types of transmission networks, such as electric wave filters and equalizers.

Karl K. Darrow, B.S., University of Chicago, 1911; University of Paris, 1911–12; University of Berlin, 1912; Ph.D., University of Chicago, 1917. Western Electric Company, 1917–25; Bell Telephone Laboratories, 1925–. Dr. Darrow has been engaged largely in writing on various fields of physics and the allied sciences.

I. E. Fair, B.S., in Electrical Engineering, Iowa State College, 1929. Bell Telephone Laboratories, Radio Research Department, 1929–. Mr. Fair has been engaged in experimentation on piezo-electric crystals for frequency control.

Frank Gray, B.S., Purdue, 1911; Ph.D., University of Wisconsin, 1916. Western Electric Company, Engineering Department, 1919–25. Bell Telephone Laboratories, 1925–. Dr. Gray has been engaged in work on electro-optical systems.

H. S. Hamilton, B.S. in Electrical Engineering, Tufts College, 1916. American Telephone and Telegraph Company, Engineering Department, 1916–18; Department of Development and Research, 1918–34. Bell Telephone Laboratories, 1934–. Mr. Hamilton has been engaged exclusively in toll transmission work, including telephone repeaters, program transmission and carrier telephone systems.

F. R. Lack, B.Sc., Harvard University, 1925; Engineering Department, Western Electric Company, 1913–22; First Lieutenant, Signal Corps, A.E.F., 1917–19; Harvard University, 1922–25. Bell Tele-

phone Laboratories, 1925–. Mr. Lack has been engaged in experimental work connected with radio communication.

W. P. Mason, B.S. in Electrical Engineering, University of Kansas, 1921; M.A., Columbia University, 1924; Ph.D., 1928. Bell Telephone Laboratories, 1921–. Dr. Mason has been engaged in investigations on carrier transmission systems and more recently in work on wave transmission networks, both electrical and mechanical.

R. C. Mathes, B.Sc., University of Minnesota, 1912; E.E., 1913. Western Electric Company, Engineering Department, 1913–25. Bell Telephone Laboratories, 1925–. Mr. Mathes has been concerned with the early history of the repeater development program, the application of vacuum tube amplifiers in a variety of fields, and the application of voice controlled switching circuits in the toll telephone plant. As Associate Wire Transmission Research Director he carries on investigations relating to the transmission of speech over wire systems.

Pierre Mertz, A.B., Cornell University, 1918; Ph.D., 1926. American Telephone and Telegraph Company, Department of Development and Research, 1919–23, 1926–34. Bell Telephone Laboratories, 1934–. Dr. Mertz has been engaged in special problems in toll transmission, chiefly in telephotography, television, and cable carrier systems.

G. W. Willard, B.A., University of Minnesota, 1924; M.A., 1928; Instructor in Physics, University of Kansas, 1927–28; Student and Assistant, University of Chicago, 1928–30. Bell Telephone Laboratories, 1930–. Mr. Willard's work has had to do with special problems in piezo-electric crystals for frequency control.

S. B. Wright, M.E. in Electrical Engineering, Cornell University, 1919. Engineering Department and Department of Development and Research, American Telephone and Telegraph Company, 1919–34. Bell Telephone Laboratories, 1934–. Mr. Wright is engaged in transmission development work on voice-operated systems and wire connections to radio telephone stations.

VOLUME XIII OCTOBER, 1934 NUMBER 4

THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

An Extension of the Theory of Three-Electrode Vacuum Tube Circuits—*S. A. Levin and Liss C. Peterson* 523

The Electromagnetic Theory of Coaxial Transmission Lines and Cylindrical Shields—*S. A. Schelkunoff* 532

Contemporary Advances in Physics, XXVIII—The Nucleus, Third Part—*Karl K. Darrow* 580

The Measurement and Reduction of Microphonic Noise in Vacuum Tubes—*D. B. Penick* 614

Fluctuation Noise in Vacuum Tubes—*G. L. Pearson* . 634

Systems for Wide-Band Transmission Over Coaxial Lines—*L. Espenschied and M. E. Strieby* . . . 654

Regeneration Theory and Experiment—*E. Peterson, J. G. Kreer, and L. A. Ware* 680

Abstracts of Technical Papers 701

Contributors to this Issue 704

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

50c per Copy *$1.50 per Year*

THE BELL SYSTEM TECHNICAL JOURNAL

Published quarterly by the
American Telephone and Telegraph Company
195 Broadway, New York, N. Y.

EDITORIAL BOARD

Bancroft Gherardi	H. P. Charlesworth	F. B. Jewett
L. F. Morehouse	E. H. Colpitts	O. B. Blackwell
D. Levinger	O. E. Buckley	H. S. Osborne

Philander Norton, *Editor* J. O. Perrine, *Associate Editor*

SUBSCRIPTIONS

Subscriptions are accepted at $1.50 per year. Single copies are fifty cents each. The foreign postage is 35 cents per year or 9 cents per copy.

Copyright, 1934

PRINTED IN U. S. A.

The Bell System Technical Journal

October, 1934

An Extension of the Theory of Three-Electrode Vacuum Tube Circuits

By S. A. LEVIN and LISS C. PETERSON

The relations between input voltage and output current of the three-electrode vaccum tube are discussed when arbitrary feedback is present between grid and plate circuits. Fundamental assumptions are that the amplification factor is constant and conductive grid current absent. The relations developed in the present paper are generalizations of those given by J. R. Carson in I. R. E. Proc. of 1919, page 187. The use of the theory is illustrated by application to a simple modulator circuit. The numerical calculations in this case indicate that neglecting the effects of interelectrode tube capacitances may introduce serious errors.

Introduction

THE relations between input voltage and output current of the three-electrode vacuum tube when connected to impedances in both input and output circuits have been the subject of several papers. One of the first more extensive treatments of this problem was given by J. R. Carson,[1] using a method of successive approximations. The theory was further extended by F. B. Llewellyn,[2] E. Peterson and H. P. Evans,[3] and J. G. Brainerd.[4] The theories given by these authors did not take into account any feedback between input and output circuit except in the first approximation.

The aim of the present paper is to extend the theory of the three-electrode vacuum tube to include the effects of feedback between input and output circuits not only in the first but also in the second and higher approximations. The assumptions underlying Carson's treatment, constancy of the amplification factor and absence of conductive grid current, will be maintained. The extension of the present theory to such cases as treated by Llewellyn, Peterson-Evans and Brainerd still remains to be done.

[1] J. R. Carson: *I. R. E. Proc.*, April, 1919, page 187.
[2] F. B. Llewellyn: *B. S. T. J.*, July, 1926, page 433.
[3] E. Peterson and H. Evans: *B. S. T. J.*, July, 1927, page 442.
[4] J. G. Brainerd: *I. R. E. Proc.*, June, 1929, page 1006.

Theory

Let us consider the circuit arrangement shown in Fig. 1, where Z_1, Z_2 and Z_3 are linear impedances which may include interelectrode admittances. The impressed variable electromotive forces whose instantaneous values are denoted by E_g and E_p are in series with the impedances Z_g and Z_p, respectively. In the absence of these electromotive forces direct currents and voltages are established in the circuit due to constant grid and plate electromotive forces. With the variable electromotive forces impressed incremental currents and voltages are produced. The instantaneous values of these incremental voltages are indicated on Fig. 1 by g, e, v and p. The incremental plate current is J. The positive directions of these quantities are given by the directions of the arrows.

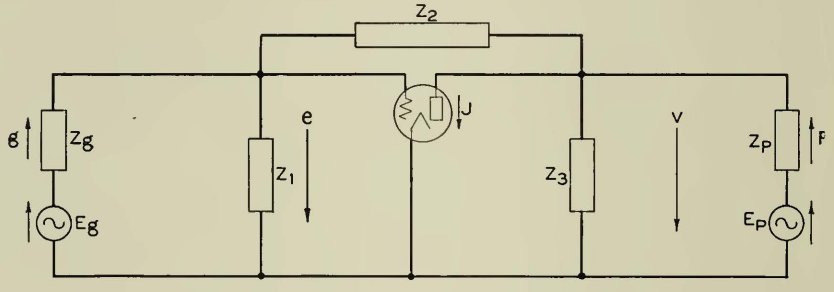


Fig. 1—Three-electrode vacuum tube and circuit.

We will now make two restrictive assumptions: first that the grid is never positive so that conductive grid current is absent, and second that the amplification factor μ is constant.

The basis for the analysis is given by the characteristic tube equation:

$$I = f\left(E_c + \frac{E_b}{\mu}\right), \tag{1}$$

where I is the total instantaneous current flowing from plate to filament; E_c is the total instantaneous potential difference between grid and filament and E_b the total instantaneous potential difference between plate and filament. μ is the amplification factor. The relation between the increments e, v and J is given by the following equation:

$$J = P_1(\mu e + v) + P_2(\mu e + v)^2 + \cdots + P_n(\mu e + v)^n + \cdots, \tag{2}$$

where

$$P_m = \frac{1}{m!}\frac{\partial^m I}{\partial E_b^m}$$

and has to be evaluated at the operating point.[1]

We have further:

$$E_g = g + e, \qquad E_p = p + v. \tag{3}$$

The equations (3) are obtained by applying the circuital laws to the network external to the tube.

We now proceed to a solution of equations (2) and (3) by means of a method of successive approximations. Let

$$J = \sum_1^\infty J_i, \qquad g = \sum_1^\infty g_i, \qquad e = \sum_1^\infty e_i, \tag{4}$$
$$p = \sum_1^\infty p_i, \qquad v = \sum_1^\infty v_i$$

and let us define the relations between the terms in the series (4) as follows:

$$J_1 = P_1(\mu e_1 + v_1), \qquad Eg = g_1 + e_1, \qquad E_p = p_1 + v_1, \tag{5}$$

$$J_2 = P_1(\mu e_2 + v_2) + P_2(\mu e_1 + v_1)^2, \tag{6}$$
$$0 = g_2 + e_2, \qquad 0 = p_2 + v_2,$$

$$J_3 = P_1(\mu e_3 + v_3) + 2P_2(\mu e_1 + v_1)(\mu e_2 + v_2) + P_3(\mu e_1 + v_1)^3, \tag{7}$$
$$0 = g_3 + e_3, \qquad 0 = p_3 + v_3,$$

$$J_4 = P_1(\mu e_4 + v_4) + P_2(\mu e_2 + v_2)^2 + 2P_2(\mu e_1 + v_1)(\mu e_3 + v_3) + 3P_3(\mu e_2 + v_2)(\mu e_1 + v_1)^2 + P_4(\mu e_1 + v_1)^4, \tag{8}$$
$$0 = g_4 + e_4, \qquad 0 = p_4 + v_4,$$

and so forth for subsequent terms.[5]

If we now let

$$R_0 = \frac{1}{P_1}, \tag{9}$$

[1] Loc. cit.

[5] The procedure of finding these equations is as follows: By substituting the first term in each of the series (4) into (2) and (3) and neglecting all terms higher than the first order equations (5) are obtained. By substituting the first two terms in each of the series (4) into (2) and (3), and neglecting terms of higher order than the second and by noting (5) equations (6) are found and so on for the remaining equations.

where R_0 is the internal resistance of the tube, equations (5), (6), (7) and (8) may be rewritten as:

$$R_0 J_1 - v_1 = \mu e_1, \qquad E_g = g_1 + e_1, \qquad E_p = p_1 + v_1, \tag{10}$$

$$R_0 J_2 - v_2 = \mu e_2 + R_0 P_2 (\mu e_1 + v_1)^2, \tag{11}$$
$$0 = g_2 + e_2, \qquad 0 = p_2 + v_2,$$

$$R_0 J_3 - v_3 = \mu e_3 + 2R_0 P_2 (\mu e_1 + v_1)(\mu e_2 + v_2) + R_0 P_3 (\mu e_1 + v_1)^3 \tag{12}$$
$$0 = g_3 + e_3, \qquad 0 = p_3 + v_3,$$

$$R_0 J_4 - v_4 = \mu e_4 + R_0 P_2 (\mu e_2 + v_2)^2 + 2R_0 P_2 (\mu e_1 + v_1)(\mu e_3 + v_3) + 3R_0 P_3 (\mu e_2 + v_2)(\mu e_1 + v_1)^2 + R_0 P_4 (\mu e_1 + v_1)^4, \tag{13}$$
$$0 = g_4 + e_4, \qquad 0 = p_4 + v_4,$$

and so forth.

Equations (10) to (13) admit of simple physical interpretations. Referring first to equations (10) it is clear that the equivalent circuit corresponding to Fig. 1 for first order quantities is given by Fig. 2. Similarly Fig. 3 is the equivalent circuit of Fig. 1 for second order effects and Fig. 4 for third order effects. Higher order effects correspond to similar circuits.

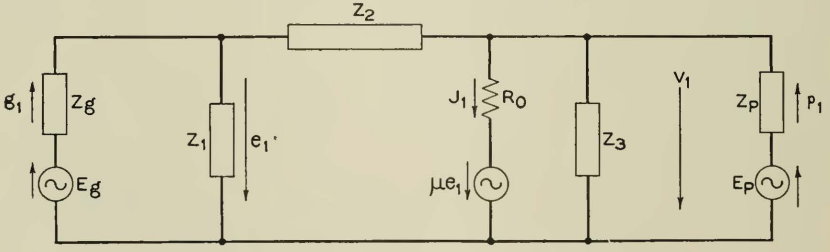


Fig. 2—Equivalent circuit, first order effects.

The equivalence expressed by Fig. 2 is the familiar circuit which has found such wide application, for instance, in amplifier and oscillator work; while the equivalent circuits in Figs. 3 and 4 represent the second and third order effects. With no feedback, that is when Z_2 is infinite, they reduce to the equivalences given by Carson.[1] Comparing now any two equivalent circuits for same order effects with and without feedback we find different values of the electromotive forces appearing in series with the internal tube resistance R_0. Otherwise

[1] Loc. cit., equations (23) and following.

the two circuits are identical except that for one the impedance Z_2 is finite and for the other infinite.

By the aid of the equivalent circuits given, that is by using equations (10), (11), (12), (13) and so forth, the terms in the series (4) can be calculated. These series formally satisfy equations (2) and (3) and are the solutions if they converge.

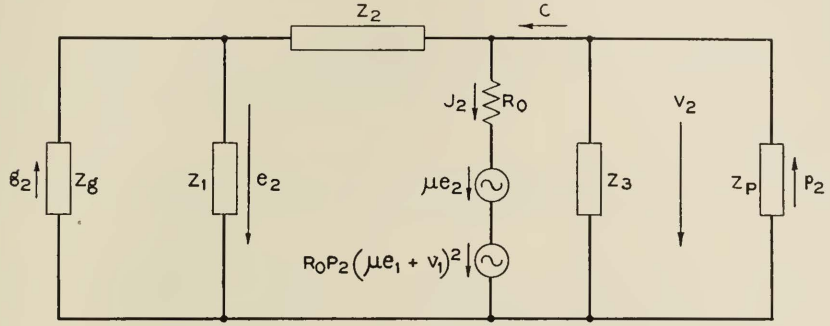


Fig. 3—Equivalent circuit, second order effects.

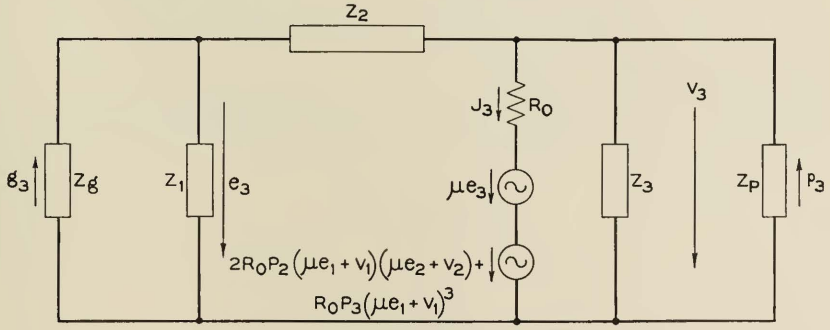


Fig. 4—Equivalent circuit, third order effects.

For the purpose of fixing our ideas we assumed at the start a definite circuit to which the tube was connected. It is obvious, however, that no matter how complicated the linear network is to which the input and output terminals of the tube are connected the procedure given above can be followed.

Application to a Modulator Circuit

As an illustration of the theory just presented we shall calculate the steady state second order effect assuming the circuit configuration to be that given in Fig. 1. In so doing we shall assume that no variable

e.m.f. is impressed in the plate circuit and that the impressed e.m.f. in the grid circuit is given by [6]

$$e_g = K \cos \omega_1 t + S \cos \omega_2 t. \tag{14}$$

We now find the instantaneous value of $\mu e_1 + v_1$ by solving the mesh equations for the equivalent circuit of Fig. 2. The result is:

$$\mu e_1 + v_1 = R_0 \left[\left| \frac{F(\omega_1)}{Z(\omega_1)} \right| K \cos (\omega_1 t - \varphi(\omega_1)) + \left| \frac{F(\omega_2)}{Z(\omega_2)} \right| S \cos (\omega_2 t - \varphi(\omega_2)) \right], \tag{15}$$

where

$$\begin{aligned} F(\omega) &= \frac{Z_1(\mu Z_2 + \mu Z_p' + Z_p')}{(Z_g + Z_1)(Z_g' + Z_2)}, \\ Z(\omega) &= R_0 + Z_p' + \frac{Z_p'(R_0 + \mu Z_g')}{Z_2 + Z_g'}, \\ Z_p' &= \frac{Z_3 Z_p}{Z_3 + Z_p}, \qquad Z_g' = \frac{Z_1 Z_g}{Z_1 + Z_g}, \\ \frac{F(\omega)}{Z(\omega)} &= \left| \frac{F(\omega)}{Z(\omega)} \right| e^{-\varphi(\omega) i} (i = \sqrt{-1}). \end{aligned} \tag{16}$$

In equations (16) we note that Z_1, Z_2, Z_3, Z_g and Z_p all are complex impedances. The driving e.m.f. for the second approximation is $R_0 P_2 (\mu e_1 + v_1)^2$. Letting

$$M = R_0^3 P_2, \tag{17}$$

we get from (15)

$$\begin{aligned} R_0 P_2 (\mu e_1 + v_1)^2 &= M \left[\frac{1}{2} \left(\left| \frac{F(\omega_1)}{Z(\omega_1)} \right|^2 K^2 + \left| \frac{F(\omega_2)}{Z(\omega_2)} \right|^2 S^2 \right) \right. \\ &+ \frac{1}{2} \left| \frac{F(\omega_1)}{Z(\omega_1)} \right|^2 K^2 \cos (2\omega_1 t - 2\varphi(\omega_1)) \\ &+ \frac{1}{2} \left| \frac{F(\omega_2)}{Z(\omega_2)} \right|^2 S^2 \cos (2\omega_2 t - 2\varphi(\omega_2)) \\ &+ \left| \frac{F(\omega_1) F(\omega_2)}{Z(\omega_1) Z(\omega_2)} \right| KS \cos ((\omega_1 - \omega_2) t - \varphi(\omega_1) + \varphi(\omega_2)) \\ &+ \left. \left| \frac{F(\omega_1) F(\omega_2)}{Z(\omega_1) Z(\omega_2)} \right| KS \cos ((\omega_1 + \omega_2) t - \varphi(\omega_1) - \varphi(\omega_2)) \right]. \end{aligned} \tag{18}$$

[6] The extension to any number of sinusoidal e.m.f.'s of arbitrary phases in both plate and grid circuit is obvious.

The driving e.m.f. given by (18) thus consists of a number of sinusoidal components including one of zero frequency. By means of the superposition theorem and the mesh equations we obtain the current and voltage distribution for our equivalent circuit in Fig. 3. Let us for instance calculate the instantaneous current flowing through the impedance consisting of Z_p and Z_3 in parallel and indicated by C in Fig. 3. The result is

$$\begin{aligned} C = M\Bigg[& \frac{\frac{1}{2}\left(\left|\frac{F(\omega_1)}{Z(\omega_1)}\right|^2 K^2 + \left|\frac{F(\omega_2)}{Z(\omega_2)}\right|^2 S^2\right)}{Z(0)} + \cdot \\ & + \frac{\frac{1}{2}\left|\frac{F(\omega_1)}{Z(\omega_1)}\right|^2 K^2}{|Z(2\omega_1)|}\cos\left(2\omega_1 t - 2\varphi(\omega_1) - \psi(2\omega_1)\right) \\ & + \frac{\frac{1}{2}\left|\frac{F(\omega_2)}{Z(\omega_2)}\right|^2 S^2}{|Z(2\omega_2)|}\cos\left(2\omega_2 t - 2\varphi(\omega_2) - \psi(2\omega_2)\right) \\ & + \frac{\left|\frac{F(\omega_1)F(\omega_2)}{Z(\omega_1)Z(\omega_2)}\right| KS}{|Z(\omega_1 - \omega_2)|}\cos\left((\omega_1 - \omega_2)t - \varphi(\omega_1)\right. \\ & \qquad \left. + \varphi(\omega_2) - \psi(\omega_1 - \omega_2)\right) \\ & + \frac{\left|\frac{F(\omega_1)F(\omega_2)}{Z(\omega_1)Z(\omega_2)}\right| KS}{|Z(\omega_1 + \omega_2)|}\cos\left((\omega_1 + \omega_2)t - \varphi(\omega_1)\right. \\ & \qquad \left. - \varphi(\omega_2) - \psi(\omega_1 + \omega_2)\right)\Bigg], \end{aligned} \tag{19}$$

where $\psi(\omega)$ is defined by

$$Z(\omega) = |Z(\omega)|e^{i\psi(\omega)} (i = \sqrt{-1}). \tag{20}$$

Let us now consider the peak value of the current of lower side-band frequency. This value is from (19)

$$MKS\left|\frac{F(\omega_1)F(\omega_2)}{Z(\omega_1)Z(\omega_2)Z(\omega_1 - \omega_2)}\right|. \tag{21}$$

If we write

$$|Z(\omega)| = \left|\frac{R_0}{1 + \frac{R_0 + \mu Z_g'}{Z_2 + Z_g'}} + Z_p'\right|_\omega \left|1 + \frac{R_0 + \mu Z_g'}{Z_2 + Z_g'}\right|_\omega, \tag{22}$$

$$|F(\omega)| = (\mu + 1)\left|\frac{Z_1}{(Z_g + Z_1)(Z_g' + Z_2)}\right|_\omega \left|\frac{\mu}{\mu + 1}Z_2 + Z_p'\right|_\omega, \tag{23}$$

expression (21) becomes:

$$\frac{MKS(\mu+1)^2\left|\frac{Z_1}{(Z_g+Z_1)(Z_g'+Z_2)}\right|_{\omega_1}\times\left|\frac{Z_1}{(Z_g+Z_1)(Z_g'+Z_2)}\right|_{\omega_2}\times\left|\frac{\mu}{\mu+1}Z_2+Z_p'\right|_{\omega_1}\left|\frac{\mu}{\mu+1}Z_2+Z_p'\right|_{\omega_2}}{\left|1+\frac{R_0+\mu Z_g'}{Z_2+Z_g'}\right|_{\omega_1}\left|1+\frac{R_0+\mu Z_g'}{Z_2+Z_g'}\right|_{\omega_2}\left|1+\frac{R_0+\mu Z_g'}{Z_2+Z_g'}\right|_{(\omega_1-\omega_2)}}\times\left|Z_p'+\frac{R_0}{1+\frac{R_0+\mu Z_g'}{Z_2+Z_g'}}\right|_{\omega_1}\left|Z_p'+\frac{R_0}{1+\frac{R_0+\mu Z_g'}{Z_2+Z_g'}}\right|_{\omega_2}\times\left|Z_p'+\frac{R_0}{1+\frac{R_0+\mu Z_g'}{Z_2+Z_g'}}\right|_{(\omega_1-\omega_2)}. \quad (24)$$

With no feedback present, that is when $Z_2 = \infty$, equation (24) reduces to

$$\frac{MKS\mu^2\left|\frac{Z_1}{Z_1+Z_g}\right|_{\omega_1}\left|\frac{Z_1}{Z_1+Z_g}\right|_{\omega_2}}{|Z_p'+R_0|_{\omega_1}|Z_p'+R_0|_{\omega_2}|Z_p'+R_0|_{(\omega_1-\omega_2)}}. \quad (25)$$

But in this case $K\left|\frac{Z_1}{Z_1+Z_g}\right|_{\omega_1}$ is the peak value K' of the grid voltage of frequency ω_1 and similarly $S\left|\frac{Z_1}{Z_1+Z_2}\right|_{\omega_2}$ is the peak value S' of the grid voltage of frequency ω_2. Expression (25) may thus be written:

$$\frac{M\mu^2K'S'}{|Z_p'+R_0|_{\omega_1}|Z_p'+R_0|_{\omega_2}|Z_p'+R_0|_{(\omega_1-\omega_2)}},$$

which is the well known expression given by Carson.[7]

For the purpose of getting an idea of the magnitudes involved let us consider a numerical example. A Western Electric No. 101D vacuum tube may have the following constants when used as a modulator: $R_0 = 9000$ ohms; $\mu = 6$; grid-cathode capacitance $C_1 = 10.5$, plate-grid capacitance $C_2 = 4.8$, and plate-cathode capacitance $C_3 = 8.1$ micromicrofarads. The impedances Z_p and Z_g are assumed to be pure resistances at all frequencies with the values 9000 and 10,000 ohms, respectively. The impressed e.m.f. is of the form given by

[7] Carson: *I. R. E. Proc.*, June, 1921, page 243.

the equation (14) where $\omega_1/2\pi$ is equal to 250,000 and $\omega_2/2\pi$ is equal to 5,000,000 cycles per second. As a reference condition let us take that for which the effect of the interelectrode capacitances is neglected. The plate current for this condition is obtained from equation (25) when Z_1 and Z_3 as well as Z_2 are made infinite. As a next step we compute the plate current from equation (24) when Z_1, Z_2, and Z_3 are the impedances corresponding to the interelectrode capacitances C_1, C_2, and C_3, respectively. It is found that this plate current is 12 db below that obtained in the reference condition. Finally it is of some interest to compute the plate current when the grid-plate capacitance alone is effective. This plate current is obtained from equation (24) by assuming Z_1 and Z_3 to be infinite and Z_2 to be the impedance corresponding to the capacitance C_2. This current is found to be 24 db below that of the reference condition.

The Electromagnetic Theory of Coaxial Transmission Lines and Cylindrical Shields

By S. A. SCHELKUNOFF

A form of circuit which is of considerable interest for the transmission of high frequency currents is one consisting of a cylindrical conducting tube within which a smaller conductor is coaxially placed. Such tubes have found application in radio stations to connect transmitting and receiving apparatus to antennæ. As a part of the development work on such coaxial systems, it has been necessary to formulate the theory of transmission over a coaxial circuit and of the shielding against inductive effects which is afforded by the outer conductor. This paper deals generally with the transmission theory of coaxial circuits and extends the theory beyond the range of present application both as regards structure and frequency.

THE mathematical theory of wave propagation along a conductor with an external coaxial return is very old, going back to the work of Rayleigh, Heaviside and J. J. Thomson. Much important work has been done in developing and extending this theory. Among the problems dealt with in this development may be listed the following: the extension of the theory to systems consisting of a plurality of cylindrical conductors; the investigation of shielding and crosstalk in coaxial systems and the effects of eccentricity; the extension of the particular solution to include the complementary modes of propagation, etc.; and in general the adaptation of the mathematical theory to engineering uses, and its translation into the concepts and language of electric circuit theory. In addition to the author's contribution a substantial part of this mathematical work has been done by the group of engineers associated with Mr. John R. Carson, formerly of the American Telephone and Telegraph Company, now of the Bell Telephone Laboratories, Inc.

The problem is ideally adapted to mathematical investigation, because the conductor shape fits perfectly into the cylindrical system of coordinates, thereby making it entirely feasible to carry out a rigorous discussion on the basis of the electromagnetic theory, instead of using ordinary circuit theory. This has obvious advantages at ultra high frequencies, where the uncertainties of the circuit theory are conspicuous and not easily compensated for. It also proves to be of greater advantage at lower frequencies than one might at first assume. Fortunately, it turns out that the final results obtained by means of field theory can be expressed in a familiar language of circuit

theory, thereby gaining all the simplicity of the latter combined with all the accuracy of the former.

Circularly Symmetric Electromagnetic Fields

In polar coordinates, Maxwell's equations assume the following form:

$$\begin{aligned}
\frac{\partial H_z}{\rho\partial\varphi} - \frac{\partial H_\varphi}{\partial z} &= (g + i\omega\epsilon)E_\rho, \qquad \frac{\partial E_z}{\rho\partial\varphi} - \frac{\partial E_\varphi}{\partial z} = -i\omega\mu H_\rho, \\
\frac{\partial H_\rho}{\partial z} - \frac{\partial H_z}{\partial\rho} &= (g + i\omega\epsilon)E_\varphi, \qquad \frac{\partial E_\rho}{\partial z} - \frac{\partial E_z}{\partial\rho} = -i\omega\mu H_\varphi, \\
\frac{1}{\rho}\left[\frac{\partial(\rho H_\varphi)}{\partial\rho} - \frac{\partial H_\rho}{\partial\varphi}\right] &= (g + i\omega\epsilon)E_z, \\
\frac{1}{\rho}\left[\frac{\partial(\rho E_\varphi)}{\partial\rho} - \frac{\partial E_\rho}{\partial\varphi}\right] &= -i\omega\mu H_z,
\end{aligned} \tag{1}$$

where E and H are respectively the electromotive and magnetomotive intensities.[1]

In general, all six field components depend upon each other. If, however, these quantities are independent of either φ or z, the partial derivatives with respect to the corresponding variable vanish and the original set of equations breaks up into two independent subsets, each involving only three physical quantities. Each of these special fields has important practical applications.

In the *circularly symmetric* case, that is, when the quantities are independent of φ, one of the independent subsets is composed of the first and the third equations on the left of (1), together with the second on the right:

$$\begin{aligned}
\frac{\partial(\rho H_\varphi)}{\partial\rho} &= (g + i\omega\epsilon)\rho E_z, \qquad \frac{\partial H_\varphi}{\partial z} = -(g + i\omega\epsilon)E_\rho, \\
\frac{\partial E_z}{\partial\rho} - \frac{\partial E_\rho}{\partial z} &= i\omega\mu H_\varphi.
\end{aligned} \tag{2}$$

This *circular magnetic* field, with its lines of magnetomotive intensity

[1] In this paper we have adopted a unified practical system of units based upon the customary cgs system augmented by adding one typically electric unit. This system has three obvious advantages: first, theoretical results are expressed directly in the units habitually employed in the laboratory; second, the dimensional character and physical significance of such quantities as $i\omega\mu$ and $g + i\omega\epsilon$ are not obscured as in other systems by suppressing dimensions of some electrical unit such as permeability or dielectric constant; and third, the form of electromagnetic equations is very simple. In this system of units the electromotive intensity E is measured in volts/cm., the magnetomotive intensity H in amperes/cm., the intrinsic conductance g in mhos/cm., the intrinsic inductance μ in henries/cm., and the intrinsic capacity ϵ in farads/cm. Thus, in empty space $\mu = 4\pi 10^{-9}$ henries/cm. or approximately 0.01257 μh/cm. and $\epsilon = (1/36\pi)\cdot 10^{-11}$ farads/cm. or approximately 0.0884 mmf./cm.

forming a system of coaxial circles, is associated with currents flowing in isolated wires as, for example, in a single vertical antenna and under ordinary operating conditions it is also found between the conductors of a coaxial pair (Fig. 1).

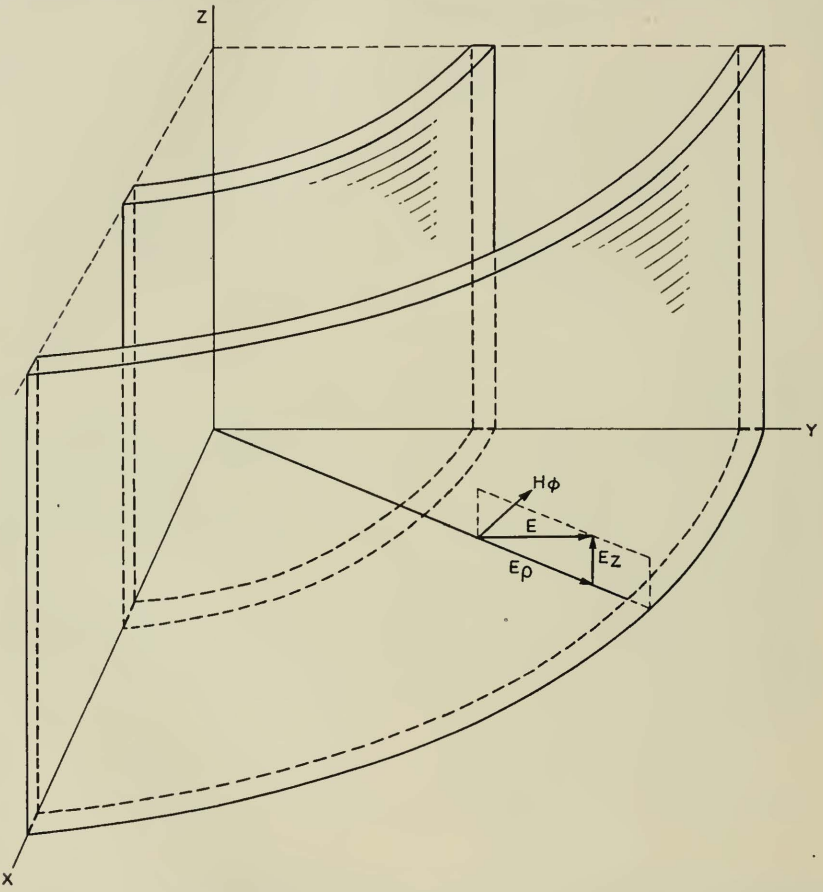


Fig. 1—The relative directions of the field components in a coaxial transmission line.

The remaining three equations of the set (1) form the second group:

$$\frac{\partial(\rho E_\varphi)}{\partial \rho} = -i\omega\mu\rho H_z, \qquad \frac{\partial E_\varphi}{\partial z} = i\omega\mu H_\rho,$$
$$(g + i\omega\epsilon)E_\varphi = \frac{\partial H_\rho}{\partial z} - \frac{\partial H_z}{\partial \rho}, \tag{3}$$

describing the *circular electric* field. Uniformly distributed electric

current in a circular turn of wire is surrounded by a field of this type; in this case, the lines of electromotive intensity form a coaxial system of circles.

Two-Dimensional Fields

By definition, two-dimensional fields are constant in some one direction. If we take the z-axis of our reference system in this direction, all the partial derivatives with respect to z vanish, z disappears from our equations and we can confine our attention to any plane normal to the z-axis.

Once more the set of six electromagnetic equations breaks up into two independent subsets. One of these is [2]

$$E_\rho = -\frac{1}{(g+i\omega\epsilon)\rho}\frac{\partial H_z}{\partial \varphi}, \qquad E_\varphi = -\frac{1}{g+i\omega\epsilon}\frac{\partial H_z}{\partial \rho},$$
$$\frac{1}{\rho}\left[\frac{\partial(\rho E_\varphi)}{\partial \rho}+\frac{\partial E_\rho}{\partial \varphi}\right] = -i\omega\mu H_z. \tag{4}$$

The calculation of what is commonly known as "electrostatic" crosstalk between pairs of parallel wires is based upon these equations. For this reason we shall name the field defined by (4) the *electric field.*

Similarly, the remaining three equations define the magnetic field:

$$H_\rho = -\frac{1}{i\omega\mu\rho}\frac{\partial E_z}{\partial \varphi}, \qquad H_\varphi = -\frac{1}{i\omega\mu}\frac{\partial E_z}{\partial \rho},$$
$$\frac{1}{\rho}\left[\frac{\partial(\rho H_\varphi)}{\partial \rho}+\frac{\partial H_\rho}{\partial \varphi}\right] = -(g+i\omega\epsilon)E_z \tag{5}$$

and are useful in the theory of what is generally known as "electromagnetic" crosstalk.

The distinction between electric and magnetic fields is purely pragmatic and is based upon a necessary and valid engineering separation of general electromagnetic interference into two component parts. In some respects the firmly entrenched terms "electrostatic crosstalk" and "electromagnetic crosstalk" are unfortunate; it would be hopeless, however, to try a change of terminology at this late stage of engineering development.

Further consideration of two-dimensional fields will be deferred until the problem of shielding is taken up later in this paper (page 567).

[2] In passing from the original set (1) we reversed the sign of E_ρ in order to make the set of equations symmetrical. The positive E_ρ is now measured toward the axis.

[3] In these equations, the sign of H_φ was reversed so that the magnetomotive intensity is now positive when it points clockwise. With this convention, the flow of energy is away from the axis when both H_φ and E_z are positive.

Exponential Propagation

While electromotive forces could be applied in such a way that the fields would be of the kind given by (3), in the coaxial transmission line as actually energized the fields are of the circular magnetic type (2) which will claim our special attention in the next few sections.

In order to solve equations (2), we naturally want to eliminate all variables but one. This purpose can be readily accomplished if E_z and E_ρ are substituted from the first and the last equations of the set into the second. Thus, we obtain the following equation for the magnetomotive intensity:

$$\frac{\partial}{\partial \rho}\left[\frac{1}{\rho}\frac{\partial(\rho H_\varphi)}{\partial \rho}\right] + \frac{\partial^2 H_\varphi}{\partial z^2} = \sigma^2 H_\varphi, \tag{6}$$

where

$$\sigma^2 = g\omega\mu i - \omega^2\epsilon\mu. \tag{7}$$

Adopting the usual method of searching for particular solutions of (6) in the form

$$H_\varphi = R(\rho)Z(z), \tag{8}$$

where $R(\rho)$ is a function of ρ alone, and $Z(z)$ a function of z alone, we get

$$\frac{1}{Z}\frac{d^2Z}{dz^2} = \Gamma^2, \tag{9}$$

$$\frac{1}{R}\frac{d}{d\rho}\left[\frac{1}{\rho}\frac{d(\rho R)}{d\rho}\right] = \sigma^2 - \Gamma^2, \tag{10}$$

where Γ is some constant about which we have no information for the time being.

Equation (9) is well known in transmission line theory; its general solution can be written in the form

$$Z = Ae^{\Gamma z} + Be^{-\Gamma z}, \tag{11}$$

where A and B are arbitrary constants. The solutions of (10) are Bessel functions. Since equation (6) is linear, we may invoke the principle of superposition and add any number of particular solutions corresponding to different values of Γ. Thus we can form an infinite variety of other solutions so as to satisfy the physical conditions of various practical problems.

It is seen at once from the first and the last equations of the set (2) that to each H_φ of the form (8) there correspond an E_z and E_ρ of the same form; i.e., there exist circularly symmetric electromagnetic

fields, all of whose components vary exponentially in the direction of the axis of symmetry. Whether any of these fields can be produced individually by some simple physical means is impossible to decide on theoretical grounds alone. It may happen, of course, that the field due to *any* practically realizable source is always a combination of several simple exponential fields. In any case, however, we want to know the properties of pure exponential solutions.

It is convenient to make the exponential character of the quantities E_ρ, E_z and H_φ explicit and write them respectively in the form $E_\rho e^{-\Gamma z}$, $E_z e^{-\Gamma z}$ and $H_\varphi e^{-\Gamma z}$. The new quantities E_ρ, E_z and H_φ are functions of ρ only. If the suggested substitution is made in equations (2), the factor $e^{-\Gamma z}$ cancels out and we have

$$E_\rho = \frac{\Gamma}{g + i\omega\epsilon} H_\varphi, \qquad i\omega\mu H_\varphi = \frac{dE_z}{d\rho} + \Gamma E_\rho,$$
$$\frac{d(\rho H_\varphi)}{d\rho} = (g + i\omega\epsilon)\rho E_z. \tag{12}$$

The quantity Γ is called the *longitudinal propagation constant* or simply the *propagation constant* when no confusion is possible.[4]

Recalling the implied exponential time factor $e^{i\omega t}$, we see that the complete exponential factor in the expressions for the field intensities is $e^{-\Gamma z + i\omega t}$. The propagation constant Γ is often a complex number and can be represented in the form $\alpha + i\beta$ where the real part is called the *attenuation constant* and the imaginary part, the *phase constant.* Thus, $e^{-\alpha z}$ measures the decrease in the amplitudes of the intensities and $e^{-i(\beta z - \omega t)}$, the change of their phases in time as well as in the z-direction. The latter factor suggests that we are dealing with a wave moving in the positive direction of the z-axis with a velocity (ω/β). A wave moving in the opposite direction is obtained by reversing the sign of Γ.

Perfectly Conducting Coaxial Cylinders [5]

Let us now consider one of the simplest problems which, though purely academic in itself, will throw some light on what is likely to happen under less ideal conditions. We suppose that a perfect dielectric is enclosed between two perfectly conducting coaxial cylinders (Fig. 1) whose radii [6] are b and a $(b < a)$. Our problem is to find the symmetric electromagnetic fields which can exist in such a medium.

[4] Another set of exponential solutions is obtained from this by changing Γ into $-\Gamma$.

[5] For a thorough discussion of "complementary" waves in coaxial pairs the reader is referred to John R. Carson [4].

[6] Only the outer radius of the inner conductor and the inner radius of the outer conductor need be considered because in perfectly conducting media electric states are entirely surface phenomena.

In a perfect dielectric $g = 0$ and the preceding set of equations becomes

$$E_\rho = \frac{\Gamma}{i\omega\epsilon} H_\varphi, \qquad i\omega\mu H_\varphi = \frac{dE_z}{d\rho} + \Gamma E_\rho,$$
$$\frac{d(\rho H_\varphi)}{d\rho} = i\omega\epsilon\rho E_z. \tag{13}$$

No force is required to sustain electric current in perfect conductors and the tangential components of the intensities are continuous across the boundaries between different media; therefore, the longitudinal electromotive intensity vanishes where ρ equals either a or b.

Substituting E_ρ from the first equation into the second, solving the latter for H_φ and inserting it into the third equation, we have successively

$$H_\varphi = -\frac{i\omega\epsilon}{m^2}\frac{dE_z}{d\rho}, \tag{14}$$

and

$$\rho\frac{d^2E_z}{d\rho^2} + \frac{dE_z}{d\rho} + m^2\rho E_z = 0, \tag{15}$$

where, for convenience, we let $\Gamma^2 + \omega^2\epsilon\mu = m^2$. The most general solution of the last equation is usually written in the form

$$E_z(\rho) = AJ_0(m\rho) + BY_0(m\rho), \tag{16}$$

where J_0 and Y_0 are Bessel functions of order zero and A and B are constants so far unknown.[7]

The constants A and B can be determined from the fact already mentioned that E_z vanishes on the surface of either conductor, i.e., from the following equations:

$$AJ_0(mb) + BY_0(mb) = 0,$$
and
$$AJ_0(ma) + BY_0(ma) = 0. \tag{17}$$

These equations are certainly satisfied if both constants are equal to 0. If, however, they are not equal to 0 simultaneously, we can determine their ratio from each equation of the above system. These ratios should be the same, of course, and yet they cannot be equal for every value of m. Thus, the *permissible* values of m are the roots of

[7] For large values of the argument these Bessel functions are very much like slightly damped sinusoidal functions; in fact $J_0(x)$ and $Y_0(x)$ are approximately equal, respectively, to $\sqrt{2/\pi x}\cos(x - \pi/4)$ and $\sqrt{2/\pi x}\sin(x - \pi/4)$, provided x is large enough.

the following equation:

$$-\frac{A}{B} = \frac{Y_0(mb)}{J_0(mb)} = \frac{Y_0(ma)}{J_0(ma)} . \tag{18}$$

This equation has an infinite number of roots [8] whose approximate values can be readily determined if we replace Bessel functions by their approximations in terms of circular functions. Thus, we have

$$m_n = \frac{n\pi}{a - b} , \qquad (n = 1, 2, 3, \cdots). \tag{19}$$

This is a surprisingly good approximation for *all* roots if the radius of the outer conductor is less than three times that of the inner; and the larger the n, the better the approximation.[9] The propagation constants are computed from the corresponding values of m_n by means of the following equation,

$$\Gamma_n = \sqrt{m_n^2 - \omega^2\epsilon\mu}. \tag{20}$$

First of all, let us study the simplest solution in which both A and B vanish. In this case, the longitudinal electromotive intensity vanishes identically. The magnetomotive intensity—and the transverse electromotive intensity, as well—also vanishes unless the denominator m^2 in equation (14) equals zero. If all intensities were to vanish, we should have no field and there would be nothing to talk about; hence, we take the other alternative and let

$$\Gamma^2 + \omega^2\epsilon\mu = 0, \qquad \text{i.e.,} \qquad \Gamma = i\omega\sqrt{\epsilon\mu}, \tag{21}$$

the positive sign having been implied in writing equations (13). In air, $\epsilon\mu = (1/c^2)$ where c is the velocity of light in cm.; hence, in air this particular propagation constant equals $i\omega/c$. Since E_z equals zero everywhere, the electromotive intensity is wholly transverse; and the flow of energy being, according to Poynting, at right angles to the electromotive and magnetomotive intensities, the energy transfer is wholly longitudinal.

The above method of determining the propagation constant may be open to suspicion; besides, the method does not tell how to obtain the actual values of the electromagnetic intensities but merely leads to a relation compatible with the existence of such intensities. Therefore, let us obtain the wanted information directly from the funda-

[8] A. Gray and G. B. Mathews, "A Treatise on Bessel Functions" (1922), p. 261.

[9] It is strictly accurate if the radii of the cylinders are infinite, i.e., if we are dealing with a dielectric slab bounded by perfectly conducting planes.

mental equations (13) which assume the following simple form:

$$i\omega\epsilon E_\rho = \Gamma H_\varphi, \qquad i\omega\mu H_\varphi = \Gamma E_\rho, \qquad \frac{d(\rho H_\varphi)}{d\rho} = 0, \tag{22}$$

if E_z vanishes identically. Either of the first two equations determines the ratio of the electromotive intensity to the magnetomotive; the two ratios are consistent only if the condition (21) is satisfied. Then, we have also

$$E_\rho = \frac{\Gamma}{i\omega\epsilon} H_\varphi = \sqrt{\frac{\mu}{\epsilon}} H_\varphi \qquad \text{and} \qquad H_\varphi = \frac{A}{\rho}, \tag{23}$$

where A is some quantity independent of ρ. This constant can be readily calculated from Ampere's law. The magnetomotive force acting along the circumference of any particular cross-section of the inner cylinder equals $2\pi\rho H_\varphi$ amperes, i.e., $2\pi A$; since this M.M.F. should equal the total current I flowing in the inner conductor through the cross-section, the quantity A equals $I/2\pi$. Reintroducing the implied factor $e^{-\Gamma z}$, we have

$$\begin{aligned} H_\varphi &= \frac{I}{2\pi\rho} e^{-\Gamma z}, \\ E_\rho &= \frac{I}{2\pi\rho} \sqrt{\frac{\mu}{\epsilon}} e^{-\Gamma z}. \end{aligned} \tag{24}$$

In practical measurements we are concerned with the total potential difference (V) between the cylinders, rather than with the transverse electromotive intensity. The former is merely the integral of the intensity,

$$V = \int_b^a E_\rho d\rho = \left(\frac{1}{2\pi} \sqrt{\frac{\mu}{\epsilon}} \log \frac{a}{b} \right) I e^{-\Gamma z}. \tag{25}$$

This voltage and the current I vary as voltage and current in a semi-infinite transmission line whose propagation constant is Γ *and whose characteristic impedance is*

$$Z_0 = \frac{V}{I e^{-\Gamma z}} = \frac{1}{2\pi} \sqrt{\frac{\mu}{\epsilon}} \log \frac{a}{b}. \tag{26}$$

At any point z the intensities E_ρ *and* H_φ *have the same values as would the voltage and current at the same distance z from the end of a transmission line whose propagation constant and characteristic impedance are respectively* $i\omega\sqrt{\epsilon\mu}$ *and* $\sqrt{\mu/\epsilon}$.

The connection between electromagnetic theory and line theory is so important that, risking repetition, we wish to emphasize their intimate relationship by deriving the well-known differential equations of the line theory directly from the electromagnetic equations (2) *combined with the assumption that the longitudinal electromotive intensity vanishes everywhere.* We already know that under the assumed conditions the first equation of the system (2) becomes

$$H_\varphi = \frac{I}{2\pi\rho}, \tag{27}$$

where I is the total current flowing in the inner cylinder through a particular cross-section and is some function [10] of z. We can therefore rewrite the last two equations of the system as follows:

$$\frac{\partial E_\rho}{\partial z} = -\frac{i\omega\mu}{2\pi\rho} I, \qquad \frac{1}{2\pi\rho}\frac{\partial I}{\partial z} = -i\omega\epsilon E_\rho. \tag{28}$$

We have merely to integrate both equations with respect to ρ from b to a and substitute the potential difference V for the integral of the transverse electromotive intensity to obtain

$$\frac{\partial V}{\partial z} = -\left(\frac{i\omega\mu}{2\pi}\log\frac{a}{b}\right) I, \qquad \frac{\partial I}{\partial z} = -\frac{2\pi i\omega\epsilon}{\log\frac{a}{b}} V, \tag{29}$$

which are the equations of the transmission line whose distributed series inductance equals $(\mu/2\pi)\log(a/b)$ henries/cm. and shunt capacity $2\pi\epsilon/(\log a/b)$ farads/cm.

With this, we conclude the special case in which the longitudinal electromotive intensity vanishes everywhere, the propagation constant equals $i\omega\sqrt{\epsilon\mu}$, and the velocity of transmission is that of light.

We now turn our attention to the case in which A and B do not vanish. We have already noted that the propagation constants are given by equation (20). Since, in this case, we are interested primarily in the nature of the phenomena rather than in the details of field distribution, we shall simplify our mathematics by supposing the radii of the cylinders to be infinite. Thus, the cylinders become two planes perpendicular to the x-axis, distance a apart. The φ-direction, then, coincides with the y-direction and, therefore, all the intensities are independent of the y-coordinate. Let us choose the z-axis halfway between the planes. The equations describing this two-dimen-

[10] On this occasion, we should remember that a particular type of this function had not yet been ascertained at the time the equations (2) were arrived at.

sional transmission line are

$$\begin{aligned} \frac{\partial H_y}{\partial x} &= i\omega\epsilon E_z, \\ \frac{\partial H_y}{\partial z} &= -i\omega\epsilon E_x, \\ \frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x} &= -i\omega\mu H_y. \end{aligned} \tag{30}$$

If n is an odd integer, these possess the following solutions:

$$\begin{aligned} E_x &= A\,\frac{\Gamma_n}{i\omega\epsilon}\sin\frac{n\pi x}{a}, \\ E_z &= A\,\frac{n\pi}{i\omega\epsilon a}\cos\frac{n\pi x}{a}, \\ H_y &= A\,\sin\frac{n\pi x}{a}; \end{aligned} \tag{31}$$

and if n is an even integer,

$$\begin{aligned} E_x &= A\,\frac{\Gamma_n}{i\omega\epsilon}\cos\frac{n\pi x}{a}, \\ E_z &= -A\,\frac{n\pi}{i\omega\epsilon a}\sin\frac{n\pi x}{a}, \\ H_y &= A\,\cos\frac{n\pi x}{a}, \end{aligned} \tag{32}$$

where

$$\Gamma_n = \sqrt{\frac{n^2\pi^2}{a^2} - \omega^2\epsilon\mu} = \pi\sqrt{\frac{n^2}{a^2} - \frac{4}{\lambda^2}}, \tag{33}$$

and λ is the wave-length corresponding to the frequency f.

Let us now define the *longitudinal impedance* (Z_z) as the ratio of E_x to H_y,

$$Z_z = \frac{\Gamma_n}{i\omega\epsilon}, \tag{34}$$

and the *transverse impedance* (the impedance in the x-direction) as the ratio of E_z to H_y,

$$\begin{aligned} Z_x &= \frac{n\pi}{i\omega\epsilon a}\cot\frac{n\pi x}{a}, && \text{if } n \text{ is odd}, \\ Z_x &= -\frac{n\pi}{i\omega\epsilon a}\tan\frac{n\pi x}{a}, && \text{if } n \text{ is even}. \end{aligned} \tag{35}$$

It will be observed that, depending on the frequency, the longitudinal

propagation constant Γ_n is either real or purely imaginary; it vanishes if $a = n(\lambda/2)$, that is, if the spacing between the planes is a whole number of half wave-lengths. When the propagation constant is real, the longitudinal impedance is purely imaginary, and vice versa, when the propagation constant is purely imaginary, the longitudinal impedance is real. In the former case, no energy is transmitted longitudinally but merely surges back and forth, and in the latter case we have a true transmission line. The transverse impedance is purely imaginary at all frequencies and, hence, the energy merely fluctuates to and fro.

If the frequency is sufficiently low, all of these higher order propagation constants are real and all the energy is transmitted in the *principal mode* described by equations (21) to (29). The rôle of the higher propagation constants consists in redistributing the energy near the sending terminal,[11] that is, in terminal distortion. But as the frequency gets high enough to make the wave-length less than $2a$, the next transmission mode may become prominent, and so forth up the infinite ladder of transmission modes.

Imperfect Coaxial Conductors [12]

We shall now suppose that the conductors are not perfect; i.e., the conductivity instead of being infinite, is merely large. Assuming that our solutions are continuous functions of conductivity (this can be proved), we conclude: first, there exists an infinite series of propagation constants approaching the values given in the preceding section as the conductivity tends to infinity; second, one of these propagation constants, namely that approaching $i\omega\sqrt{\epsilon\mu}$, is very small unless the conductivity is too small. In the immediately succeeding sections we shall be concerned only with electromagnetic fields corresponding to this particular propagation constant.

Let us now prove that the simple expression for the magnetomotive intensity in the dielectric between perfectly conducting cylinders is still true for all practical purposes, even if the conductors are merely good, and even when there are more than two of them. Since the lines of force are circles, coaxial with the conductors, and since H_φ is independent of φ, the total magnetomotive force acting along any one of the circles equals H_φ times the circumference of the circle $(2\pi\rho)$. This M.M.F. also equals the total current I passing through the area of the circle. Therefore, the magnetomotive intensity is $(I/2\pi\rho)$ amperes/cm. This expression is true at any point in the conductors as

[11] And near the receiving terminal as well, if the line is finite.

[12] The general theory of wave propagation in a multiple system of imperfect coaxial conductors is amply covered by John R. Carson and J. J. Gilbert [2, 3].

well as in the dielectric between them. In a conductor the total current I passing through the area of the circle is a function of ρ since the current is distributed throughout the entire cross-section of the conductor. Strictly speaking, the same is true of any circle in the dielectric. There is one important difference, however; the conduction current passing through such circles is the same and the displacement current is usually so small that it can be legitimately neglected. Thus, in the dielectric, we have to an extremely high degree of accuracy unless ρ is very large

$$H_\varphi = \frac{I}{2\pi\rho}, \tag{36}$$

where I is merely a constant, namely, the total *conduction* current passing through the area of the circle of radius ρ.

That the longitudinal displacement current can be neglected, unless the conductivity of the conductors is small, has been already indicated in the opening paragraph. The following comparison is an aid to the mathematical argument. The density of the longitudinal conduction current is gE and that of the displacement current is $i\omega\epsilon E$. Near the boundary, E is substantially the same in the conductor and in the dielectric. In copper, $g = (1/1.724)10^6$ and in air $\epsilon = (1/36\pi)10^{-11}$. Thus, even at very high frequencies, the density of the displacement current is very small compared to that of the conduction current. On the other hand, the conduction current is ordinarily distributed over a small area while the displacement current may flow across a large area. The latter area would have to be very large, however, before it could even begin to compensate for the extremely low current density.

Electromotive Intensities in Dielectrics

With the aid of equations (12) and (36), we can now calculate the electromotive intensities in the dielectric between two conductors. Thus, the transverse intensity is

$$E_\rho = \frac{\Gamma I}{2\pi(g + i\omega\epsilon)\rho}. \tag{37}$$

Substituting this in the second equation of the set (12), we obtain the following differential equation for the longitudinal intensity:

$$\frac{dE_z}{d\rho} = \left[i\omega\mu - \frac{\Gamma^2}{g + i\omega\epsilon} \right] \frac{I}{2\pi\rho}, \tag{38}$$

where μ is the permeability of the dielectric. Integrating with respect

to ρ, we have

$$E_z = \frac{1}{2\pi}\left[i\omega\mu - \frac{\Gamma^2}{g + i\omega\epsilon} \right] I \log \frac{\rho}{b'} + A, \tag{39}$$

where A is a constant to be determined from the boundary conditions.[13]

THE POTENTIAL DIFFERENCE BETWEEN TWO COAXIAL CYLINDERS

Equation (36) relates the transverse electromotive intensity to the total current flowing in the inner conductor. In practice, however, we are interested in the difference of potential between the conductors, that is, in the transverse electromotive force rather than the electromotive intensity. This potential difference V is obtained at once from equation (37) by integration:

$$V = \int_{b'}^{a''} E_\rho d\rho = \frac{\Gamma I}{2\pi(g + i\omega\epsilon)} \int_{b'}^{a''} \frac{d\rho}{\rho} = \frac{\Gamma I \log \frac{a''}{b'}}{2\pi(g + i\omega\epsilon)}. \tag{40}$$

This transverse E.M.F. produces a transverse electric current which is partly a conduction current—if the dielectric is not quite perfect—and partly a displacement (or "capacity") current.

Now, the total transverse current per centimeter length of line is

$$I_\rho = 2\pi\rho(g + i\omega\epsilon)E_\rho.$$

Then, by equation (37), we have

$$I_\rho = \Gamma I. \tag{41}$$

Therefore, equation (40) becomes

$$V = \frac{\log \frac{a''}{b'}}{2\pi(g + i\omega\epsilon)} I_\rho. \tag{42}$$

The ratio of a current to the electromotive force that produces it is called *admittance*. Hence, the distributed *radial admittance* per

[13] The following system of notation will be adhered to throughout the remainder of the paper: The inner radius of any cylindrical conductor is denoted by a, and its outer radius by b. When several coaxial conductors are used, they are differentiated by superscripts; a', a'', $a^{(3)}$, $\cdots$ referring to their inner radii, for example, and b', b'', $b^{(3)}$, $\cdots$ to their outer radii. This convention also applies to conductivities, permeabilities, and other physical constants of the conductors in question.

For convenience, we have written the ratio of ρ to the outer radius of the inner conductor in place of ρ; this change affects only the arbitrary constant A which will eventually be assigned the value required by the boundary conditions. When written in this form, the first term of E_z vanishes on the surface of the inner conductor which is a convenience in determining the value of A.

unit length between two cylindrical conductors is

$$Y = \frac{2\pi(g + i\omega\epsilon)}{\log \frac{a''}{b'}} \equiv G + i\omega C, \tag{43}$$

the symbols G and C being used in the usual way to designate the distributed radial conductance and capacity. Writing these separately, we have

$$G = \frac{2\pi g}{\log \frac{a''}{b'}}, \qquad C = \frac{2\pi\epsilon}{\log \frac{a''}{b'}}. \tag{44}$$

Returning to (40), we find that V can be written in the form

$$V = \frac{\Gamma}{Y} I.$$

But the ratio of the transverse electromotive force V to the longitudinal current I is known as the longitudinal *characteristic impedance* of the coaxial pair. Its value is obviously Γ/Y.

The External Inductance

In dealing with parallel wires it is customary to use the term "external inductance" for the total magnetic flux in the space surrounding the pair.[14] We shall adopt the same usage in connection with coaxial pairs. Strictly speaking, we must therefore consider it as being composed of two parts: one being the flux *between* the cylinders, the other the flux in the space surrounding them. But the longitudinal displacement current is negligible by comparison with the conduction current, and effects due to it have been consistently ignored throughout this part of our study. To the same order of approximation, the flux outside the pair is negligible by comparison with that between them, whence we find the "external inductance" to be

$$L_e = \frac{\mu \int_{b'}^{a''} H_\varphi d\rho}{I} = \frac{\mu}{2\pi} \log \frac{a''}{b'} \text{ henries/cm.} \tag{45}$$

[14] While this definition is very descriptive, it is not strictly accurate unless the wires are perfectly conducting. The correct definition should read as follows: The external inductance of a parallel pair is the measure (per unit current) of magnetic energy stored in the space surrounding the pair. The reason the simpler definition fails for imperfectly conducting parallel wires is because some of the lines of magnetic flux lie partly inside and partly outside the wires. This does not happen in connection with coaxial pairs even when they are not perfectly conducting. Hence we are warranted in using the simpler idea.

Comparing this with equation (44), we have the following relation between the external inductance and the capacity

$$CL_e = \epsilon\mu. \tag{46}$$

Propagation Constants of Coaxial Pairs

Since the relation between electromotive intensity and current is linear, we are justified in writing the intensities at the adjacent surfaces of the pair in the form

$$E_z(b') = Z_b'I, \qquad E_z(a'') = Z_a''I, \tag{47}$$

where Z_b' and Z_a'' depend only upon the material of the conductors and the geometry of the system. These quantities will be called *surface impedances* of the inner and outer conductors, respectively.

Inserting (47) in (39) we obtain

$$\begin{aligned} A &= Z_b'I, \\ \frac{1}{2\pi}\left[i\omega\mu - \frac{\Gamma^2}{g + i\omega\epsilon} \right] I \log\frac{a''}{b'} + A &= -Z_a''I, \end{aligned} \tag{48}$$

by means of which A and Γ may be expressed in terms of Z_b' and Z_a''. If we solve the first of these for A and substitute the value thus derived in the second we get, by virtue of (45),

$$\frac{\Gamma^2}{2\pi(g + i\omega\epsilon)} \log\frac{a''}{b'} = Z_a'' + Z_b'' + i\omega L_e, \tag{49}$$

or, by (43)

$$\Gamma^2 = YZ, \tag{50}$$

where for brevity we have written

$$Z = Z_a'' + Z_b' + i\omega L_e. \tag{51}$$

Direct Conversion of the Circularly Symmetric Field Equations into Transmission Line Equations

As the practical applications of Maxwell's theory become more numerous, it becomes increasingly important to formulate its exact connection with transmission line theory. With this purpose in mind, let us attempt to throw (2) into the form of the transmission line equations.

The obvious plan of attack is to introduce into (2) the transverse voltage V and the longitudinal current I, in place of the intensities E and H. The total current is introduced by substituting $(I/2\pi\rho)$ for H_φ, and the total voltage by integrating the set of equations (2) in the transverse direction. The first equation gives us nothing of

importance.[15] The second and third equations, on the other hand, give

$$\frac{i\omega\mu I}{2\pi}\log\frac{a''}{b'} = E_z''(a) - E_z'(b) - \frac{\partial V}{\partial z}, \tag{52}$$

$$V = -\frac{\log\frac{a''}{b'}}{2\pi(g + i\omega\epsilon)}\frac{\partial I}{\partial z}.$$

But, upon substituting (45), (47) and (51) in the first of these equations and (43) in the second, we get

$$\frac{\partial V}{\partial z} = -ZI, \qquad \frac{\partial I}{\partial z} = -YV, \tag{53}$$

where Z and Y are to be interpreted respectively as the distributed series impedance and shunt admittance.

Current Distribution in Cylindrical Conductors

So far, we have been dealing with electromagnetic intensities in dielectrics. We now turn our attention to conductors and determine their current distributions with the ultimate view of calculating their surface impedances. One of our sources of information is the familiar set of equations (12). In these equations, however, we now let $\epsilon = 0$ since the displacement current in conductors is negligibly small by comparison with the conduction current. From these equations, we eliminate electromotive intensities and thus obtain a differential equation for the magnetomotive intensity. The latter is in fact equation (6) with only one difference: the exponential factor $e^{-\Gamma z}$ has been explicitly introduced and cancelled so that the equation has become

$$\frac{d}{d\rho}\left[\frac{1}{\rho}\frac{d(\rho H_\varphi)}{d\rho}\right] = (\sigma^2 - \Gamma^2)H_\varphi,$$

or

$$\frac{d^2H_\varphi}{d\rho^2} + \frac{1}{\rho}\frac{dH_\varphi}{d\rho} - \frac{H_\varphi}{\rho^2} = (\sigma^2 - \Gamma^2)H_\varphi, \tag{54}$$

where

$$\sigma^2 = g\omega\mu i = 2\pi g\mu f i.$$

This σ will be called the intrinsic propagation constant of solid metal.

[15] Our standard practice of neglecting the longitudinal displacement currents has given us the general rule that $2\pi\rho H_\varphi = I$ is independent of ρ. Using this relation in the first of equations (2.2), we get

$$(g + i\omega\epsilon)E_z \doteq 0;$$

but this merely reflects the fact that $g + i\omega\epsilon$ is very small.

The attenuation and the phase constants are each equal to $\sqrt{\pi g \mu f}$. The intrinsic propagation constants of metals are large quantities except at low frequencies as the accompanying table indicates.

PROPAGATION CONSTANT OF COMMERCIAL COPPER

$g = 5.800\ 10^5$ mhos/cm.
$\mu = 0.01257\ \mu h$/cm.

f	$\frac{\sigma}{1+i} = \sqrt{\pi g \mu f}$
0	0.0
1	0.1513
10	0.4785
100	1.513
10,000	15.13
1,000,000	151.3
100,000,000	1513.

On the other hand, Γ is very small; if air is the dielectric between the conductors, Γ is of the order of $(1/3)i\omega\ 10^{-10}$. Hence, even at high frequencies Γ^2 is negligibly small by comparison with σ^2 and we can rewrite (54) as follows:

$$\frac{d}{d\rho}\left[\frac{1}{\rho}\frac{d}{d\rho}(\rho H_\varphi)\right] = \sigma^2 H_\varphi. \tag{55}$$

This is Bessel's equation and its solution can be written down at once [16] as

$$H_\varphi = AI_1(\sigma\rho) + BK_1(\sigma\rho), \tag{56}$$

where the functions $I_1(u)$ and $K_1(u)$ are the *modified Bessel functions of the first order and respectively of the first and second kind.* For large values of the argument we have approximately

$$\begin{aligned} I_1(u) &= \frac{e^u}{\sqrt{2\pi u}}\left(1 - \frac{3}{8u}\right), \\ K_1(u) &= \sqrt{\frac{\pi}{2u}}\, e^{-u}\left(1 + \frac{3}{8u}\right); \end{aligned} \tag{57}$$

[16] It is interesting to note that in the case of a fairly thin hollow conductor whose inner radius is not too small there exist very simple approximate solutions of (55). Under these circumstances ρ varies over such a small range that no serious error is introduced in treating the factors $(1/\rho)$ and ρ in (55) as constants, and the equation becomes

$$\frac{d^2 H_\varphi}{d\rho^2} = \sigma^2 H_\varphi,$$

which is satisfied by the exponential functions $e^{\sigma\rho}$ and $e^{-\sigma\rho}$. The larger the value of ρ and the faster the change in H_φ with ρ, the better is the approximation.

while for small values

$$I_1(u) = \frac{u}{2} + \frac{u^3}{16},$$
$$K_1(u) = \frac{1}{u} + \frac{u}{2}\log\frac{u}{2}. \tag{58}$$

The function $I_1(u)$ becomes infinite and $K_1(u)$ vanishes when u is infinite.[17] When u is zero, $I_1(u)$ vanishes and $K_1(u)$ becomes infinite.

The longitudinal electromotive intensity is calculated from the third equation (12) with the aid of the following rules for differentiation of modified Bessel functions of any order n:

$$\frac{d}{dx}(x^n I_n) = x^n I_{n-1},$$
$$\frac{d}{dx}(x^n K_n) = -x^n K_{n-1}. \tag{59}$$

Thus,

$$E_z = \eta[AI_0(\sigma\rho) - BK_0(\sigma\rho)], \tag{60}$$

where

$$\eta = \frac{\sigma}{g} = \frac{i\omega\mu}{\sigma}. \tag{61}$$

For reasons which will appear later, this quantity η will be called the *intrinsic impedance of solid metal.*

The current density is merely the product of the intensity E_z and the conductivity g.

In a general way the behavior of the functions of zero order is similar to that of the functions whose order is unity. Thus, for large values of the argument,

$$I_0(u) = \frac{e^u}{\sqrt{2\pi u}}\left(1 + \frac{1}{8u}\right),$$
$$K_0(u) = \sqrt{\frac{\pi}{2u}}\,e^{-u}\left(1 - \frac{1}{8u}\right), \tag{62}$$

and for small values

$$I_0(u) = 1 - \frac{u^2}{4},$$
$$K_0(u) = -\log u + 0.116. \tag{63}$$

[17] This statement is correct only as long as the real part of u is positive. This is so in our case because out of two possible values of the square root representing σ we can always choose the one with the positive real part.

Surface Impedance of a Solid Wire

On page 547 we defined the surface impedances of a coaxial pair as the ratios of the longitudinal electromotive intensities on the adjacent surfaces of the cylinders to the total currents flowing in the respective conductors. In that place, however, we were unable to give explicit formulæ for the impedances so defined because we did not yet have a precise value for E_z. Now that this omission has been supplied, we are prepared to compute Z_b' and Z_a''.

We consider the case of a *solid* inner cylinder surrounded by *any coaxial return*, and seek to determine the constants A and B in (60). Since the E.M.I. must be finite along the axis of the wire we must make $B = 0$, because the K-function becomes infinite when $\rho = 0$. On the surface of the wire the magnetomotive intensity is $I/2\pi b$ if I is the total current in the wire. By equation (56) this intensity equals $AI_1(\sigma b)$; hence,

$$A = \frac{I}{2\pi b I_1(\sigma b)}$$

and the final expression for the electromotive intensity within the wire is

$$E_z(\rho) = \frac{\eta I_0(\sigma\rho)}{2\pi b I_1(\sigma b)} I. \tag{64}$$

Thus, we have the following expression for the surface impedance of the solid wire:

$$Z_b = \frac{E_z(b)}{I} = \frac{\eta I_0(\sigma b)}{2\pi b I_1(\sigma b)}, \text{ ohms/cm.} \tag{65}$$

As the argument increases, the modified Bessel functions of the first kind (the I-functions) become more and more nearly proportional to the exponential functions of the same argument. Thus, if the absolute value of σb exceeds 50, the Bessel functions in the preceding equation cancel out and the following simple formula holds within 1 per cent:

$$Z_b = \frac{\eta}{2\pi b} = \frac{1}{2b}\sqrt{\frac{\mu f}{\pi g}}(1 + i), \text{ ohms/cm.} \tag{66}$$

This surface impedance consists of a resistance representing the amount of energy dissipated in heat, and a reactance due to the magnetic flux in the wire itself. Separating (66) into these two parts, we have, approximately,

$$R_b = \omega L_b = \frac{1}{2b}\sqrt{\frac{\mu f}{\pi g}}.$$

However, most of the error in (66) occurs in the real part. If more accurate approximations for Bessel functions are used, then

$$\begin{aligned} R_b &= \frac{1}{2b}\sqrt{\frac{\mu f}{\pi g}} + \frac{1}{4\pi g b^2}, \\ \omega L_b &= \frac{1}{2b}\sqrt{\frac{\mu f}{\pi g}}; \end{aligned} \tag{67}$$

these are correct within 1 per cent if $|\sigma b| > 6$. The surface inductance L_b equals $(1/4\pi b)\sqrt{\mu/\pi g f}$ henries/cm.; it decreases as the frequency increases.

If the wire is so thin or the frequency is so low that $|\sigma b| < 6$, equation (65) has to be used. Its use in computations is quite simple, however, because the argument σb is a complex number of the form $u\sqrt{i}$; and the necessary functions have been tabulated. Lord Kelvin introduced the symbols ber u and bei u for the real and the imaginary parts of $I_0(u\sqrt{i})$, so that we now write

$$I_0(u\sqrt{i}) = \text{ber } u + i \text{ bei } u. \tag{68}$$

Differentiating, we have

$$\sqrt{i}\, I_0'(u\sqrt{i}) = \sqrt{i}\, I_1(u\sqrt{i}) = \text{ber}' u + i \text{ bei}' u,$$

and therefore

$$I_1(u\sqrt{i}) = \frac{\text{ber}' u + i \text{ bei}' u}{\sqrt{i}}. \tag{69}$$

If we insert these values in (65), and recall that the d.-c. resistance of a solid wire is $1/\pi g b^2$, and that $\sigma = g\eta$, we obtain at once

$$\begin{aligned} \frac{Z_b}{R_{\text{d.-c.}}} = \pi g b^2 Z_b = \frac{u}{2}\frac{\text{ber } u \text{ bei}' u - \text{bei } u \text{ ber}' u}{(\text{ber}' u)^2 + (\text{bei}' u)^2} \\ + i\frac{u}{2}\frac{\text{ber } u \text{ ber}' u + \text{bei } u \text{ bei}' u}{(\text{ber}' u)^2 + (\text{bei}' u)^2}, \end{aligned} \tag{70}$$

where u is the absolute value of σb. The accompanying graph illustrates the real and imaginary parts of this equation [18] (Fig. 2).

The Surface Impedances of Hollow Cylindrical Shells [19]

In the case of a hollow conductor whose inner and outer radii are respectively equal to a and b, the return coaxial path for the current

[18] For equation (70) and various approximations see E. Jahnke and F. Emde.

[19] In the case of self-impedances the more general equations of two parallel cylindrical shells were deduced by Mrs. S. P. Mead. For the special formulæ concerning self-impedances of coaxial pairs see A. Russell.

may be provided either outside the given conductor or inside it or partly inside and partly outside. We designate by Z_{aa} the surface impedance with internal return, and by Z_{bb}, that with external return. These impedances are equal only at zero frequency; but if the con-

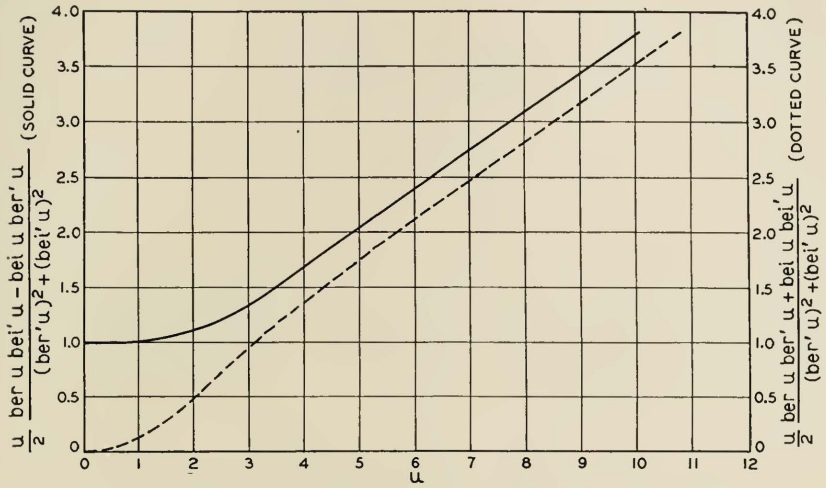


Fig. 2—The skin effect in solid wires. The upper curve represents the ratio of the a-c. resistance of the wire to its d-c. resistance and the lower curve the ratio of the internal reactance to the d-c. resistance.

ductor is thin, they are nearly equal at all frequencies. If the return path is partly internal and partly external, we have in effect two transmission lines with a distributed mutual impedance Z_{ab} due to the mingling of the two currents in the hollow conductor common to both lines. However, since this quantity Z_{ab} is not the total mutual impedance between the two lines unless the hollow conductor is the only part of the electromagnetic field common to them, it is better to call Z_{ab} the *transfer impedance* from one surface of the conductor to the other.

In order to determine these impedances, let us suppose that of the total current $I_a + I_b$ flowing in the hollow conductor, the part I_a returns inside and the rest outside. Since the total current enclosed by the inner surface of the given conductor is $-I_a$, and that enclosed by the outer surface is I_b, the magnetomotive intensity takes the values $-(I_a/2\pi a)$ and $(I_b/2\pi b)$, respectively, at these surfaces. This information is sufficient to determine the values of the constants A and B in the equation (59) governing current distribution. In fact, we

have

$$\begin{aligned} AI_1(\sigma a) + BK_1(\sigma a) &= -\frac{I_a}{2\pi a}, \\ AI_1(\sigma b) + BK_1(\sigma b) &= \frac{I_b}{2\pi b}, \end{aligned} \tag{71}$$

and therefore

$$\begin{aligned} A &= \frac{K_1(\sigma b)}{2\pi aD} I_a + \frac{I_1(\sigma a)}{2\pi bD} I_b, \\ B &= -\frac{I_1(\sigma b)}{2\pi aD} I_a - \frac{I_1(\sigma a)}{2\pi bD} I_b, \end{aligned} \tag{72}$$

where

$$D = I_1(\sigma b)K_1(\sigma a) - I_1(\sigma a)K_1(\sigma b). \tag{73}$$

Substituting these into the second equation of the set (59), we obtain the longitudinal electromotive intensity at any point of the conductor. We are interested, however, in its values at the surfaces since these values determine the surface impedances. Equating ρ successively to a and b, we obtain

$$\begin{aligned} E_z(a) &= Z_{aa}I_a + Z_{ab}I_b, \\ E_z(b) &= Z_{ba}I_a + Z_{bb}I_b, \end{aligned} \tag{74}$$

where [20]

$$\begin{aligned} Z_{aa} &= \frac{\eta}{2\pi aD}[I_0(\sigma a)K_1(\sigma b) + K_0(\sigma a)I_1(\sigma b)], \\ Z_{bb} &= \frac{\eta}{2\pi bD}[I_0(\sigma b)K_1(\sigma a) + K_0(\sigma b)I_1(\sigma a)], \\ Z_{ab} &= Z_{ba} = \frac{1}{2\pi gabD}. \end{aligned} \tag{75}$$

The results embodied in equation (74) can be stated in the following two theorems:

Theorem 1: If the return path is wholly external ($I_a = 0$) or wholly internal ($I_b = 0$), the longitudinal electromotive intensity on that surface of a hollow conductor which is nearest to the return path equals the corresponding surface impedance per unit length multiplied by the total current flowing in the conductor; and the intensity on the other surface equals the transfer impedance per unit length multiplied by the total current.

[20] To obtain the last equation, it is necessary to use the identity

$$I_0(x)K_1(x) + K_0(x)I_1(x) = \frac{1}{x}.$$

Theorem 2: If the return path is partly external and partly internal the separate components of the intensity due to the two parts of the total current are calculated by the above theorem and then added to obtain the total intensities.

At high frequencies, or when the conductors are very large, (75) can be replaced by much simpler approximate expressions.[21] If, however, we are compelled to use the rigorous equations in numerical computations, it is convenient to express the Bessel functions in terms of Thomson functions. Two of these, the ber and bei functions, or Thomson functions of the first kind, have already been introduced. The functions of the second kind are defined in an entirely analogous fashion as

$$K_0(x\sqrt{i}) = \ker x + i \operatorname{kei} x. \tag{76}$$

Differentiating, we have

$$\sqrt{i}\, K_0'(x\sqrt{i}) = -\sqrt{i}\, K_1(x\sqrt{i}) = \ker' x + i \operatorname{kei}' x, \tag{77}$$

so that

$$K_1(x\sqrt{i}) = -\frac{\ker' x + i \operatorname{kei}' x}{\sqrt{i}}. \tag{78}$$

All these subsidiary functions have been tabulated;[22] but the process of computing the impedances is laborious nevertheless.

The Complex Poynting Vector[23]

In the preceding sections we have been able to determine the surface impedances of the coaxial conductors by reducing the field equations to the form of transmission line equations, and interpreting various terms accordingly. However, if the conductors are eccentric or of irregular shape, the *effective* surface impedances are more conveniently calculated by the use of the modified Poynting theorem.

This theorem states that, if E and H are the complex electromotive and magnetomotive intensities at any point, and if E^* and H^* are the conjugate complex numbers, then[24]

$$\int\int [EH^*]dS = g\int\int\int (EE^*)dv + i\omega\mu \int\int\int (HH^*)dv. \tag{79}$$

[21] See portion of this text under the heading "Approximate Formulæ for the Surface Impedance of Tubular Conductors," page 557.

[22] British Association Tables, 1912, pp. 57–68; 1915, pp. 36–38; 1916, pp. 108–122.

[23] For an early application of the Complex Poynting vector see Abraham v. Föppl, Vol. 1 (Ch. 3, Sec. 3).

[24] The brackets signify the vector product and the parentheses the scalar product of the vectors so enclosed. The inward direction of the normal to the surface is chosen as the positive direction. The division by 4π does not occur if the consistent practical system of units is used as it is done in this paper.

To get an insight into the significance of this equation, let us consider a conductor which is part of a single-mesh circuit, and extend our integrals over the region occupied by this conductor. Then the first integral on the right of (79) represents twice the power dissipated in heat in the conductor, while $\mu \int\int\int (HH^*)dv$ is four times the average amount of magnetic energy stored in it.

On the other hand, when we look at the conductor from the standpoint of circuit theory, these two quantities are respectively RI^2 and LI^2; R and L being by definition the "resistance" and "inductance" of the conductor. Hence we have the equation,

$$\int\int [EH^*]_n dS = (R + i\omega L)I^2 = ZI^2, \tag{80}$$

from which the *impedance* Z can be computed when the field intensities are known at the surface of the conductor.

If, on the other hand, the conductor is part of a two-mesh circuit and I_1 and I_2 are the amplitudes of the currents in meshes 1 and 2 respectively, the average amount of energy dissipated in heat per second can be regarded as made up of three parts, two of which are proportional to the squares of these amplitudes, while the third is proportional to their product. The first two of these parts being dependent on the magnitude of the current flowing in one mesh only are attributed to the *self-resistance* of the conductor to the corresponding current; the third part is attributed to the *mutual resistance* of the conductor. Designating the self-resistances by R_{11} and R_{22} and the mutual resistance by R_{12}, we represent the energy dissipated in heat in the form $1/2(R_{11}I_1^2 + 2R_{12}I_1I_2 + R_{22}I_2^2)$. Similarly, the average amount of energy stored in the conductor can be represented in the form $1/4(L_{11}I_1^2 + 2L_{12}I_1I_2 + L_{22}I_2^2)$, where L_{11} and L_{22} are called respectively *self-inductances* and L_{12} *mutual inductance.* In this case, equation (79) can be written as follows:

$$\int\int [EH^*]_n dS = Z_{11}I_1^2 + 2Z_{12}I_1I_2 + Z_{22}I_2^2, \tag{81}$$

where the quantities Z_{11}, Z_{22} and Z_{12} are respectively the *self-impedances* and the *mutual impedance* of the conductor.

In general, if the conductor is part of a k-mesh circuit, we can obtain all its self-and mutual impedances by evaluating the integral $\int\int [EH^*]_n dS$ over its surface, and picking out the coefficients of various combinations of I's.

We shall have an occasion to apply these results in computing the effect of eccentricity upon the resistance of parallel cylindrical conductors.

Approximate Formulæ for the Surface Impedance of Tubular Conductors

The exact formulæ (75) for the internal impedances of a tubular conductor are hard to use for numerical computations, but simple approximations can be easily obtained if the modified Bessel functions are replaced by their asymptotic expansions and the necessary division performed as far as the second term. Thus, we have

$$\begin{aligned} Z_{bb} &= \frac{\eta}{2\pi b}\left[\coth \sigma t + \frac{\pi}{2\sigma}\left(\frac{3}{a}+\frac{1}{b}\right)\right], \\ Z_{aa} &= \frac{\eta}{2\pi a}\left[\coth \sigma t - \frac{\pi}{2\sigma}\left(\frac{3}{b}+\frac{1}{a}\right)\right], \\ Z_{ab} &= \frac{\eta}{2\pi\sqrt{ab}}\operatorname{csch} \sigma t, \end{aligned} \tag{82}$$

where t is the thickness of the tube. Separating the real and imaginary parts, we have

$$\begin{aligned} R_{bb} &= \frac{1}{2b}\sqrt{\frac{\mu f}{\pi g}}\frac{\sinh u + \sin u}{\cosh u - \cos u} + \frac{a+3b}{16\pi g a b^2}, \\ R_{aa} &= \frac{1}{2a}\sqrt{\frac{\mu f}{\pi g}}\frac{\sinh u + \sin u}{\cosh u - \cos u} - \frac{b+3a}{16\pi g b a^2}, \\ R_{ab} &= \frac{1}{\sqrt{ab}}\sqrt{\frac{\mu f}{\pi g}}\frac{\sinh \frac{u}{2}\cos\frac{u}{2} + \cosh\frac{u}{2}\sin\frac{u}{2}}{\cosh u - \cos u}, \\ \omega L_{bb} &= \frac{1}{2b}\sqrt{\frac{\mu f}{\pi g}}\frac{\sinh u - \sin u}{\cosh u - \cos u}, \\ \omega L_{aa} &= \frac{1}{2a}\sqrt{\frac{\mu f}{\pi g}}\frac{\sinh u - \sin u}{\cosh u - \cos u}, \\ \omega L_{ab} &= \frac{1}{\sqrt{ab}}\sqrt{\frac{\mu f}{\pi g}}\frac{\sinh \frac{u}{2}\cos\frac{u}{2} - \cosh\frac{u}{2}\sin\frac{u}{2}}{\cosh u - \cos u} \\ |Z_{ab}| &= \frac{\sqrt{\mu f}}{\sqrt{\pi g a b(\cosh u - \cos u)}}, \end{aligned} \tag{83}$$

where $u = t\sqrt{2g\omega\mu}$.

It is obvious that in the equations for the self-resistances, the second terms represent the first corrections for curvature and vanish altogether if the conductors are plane. Although these formulæ were derived by using asymptotic expansions which are valid only when the argument is large, i.e., at high frequencies, the results are good even at low

frequencies, provided the tubular conductor is not too thick. Thus, if the frequency is 0, the first term in the above expression for R_{bb} becomes $1/2\pi gbt$ which is the d.-c. resistance of the tube if its curvature is neglected. The second term only partially corrects for curvature, the error being of the order of $t^2/8b^2$. Hence, if the thickness of the tube is not more than 25 per cent of its *high-frequency* radius, that is, the radius of the surface nearest the return path, the error is less than 1 per cent. The formula for the mutual impedance is exceedingly good down to zero frequency for all ordinary thicknesses.

If the frequency is very high, further approximations can be made and the formulæ simplified as follows:

$$
\begin{aligned}
R_{bb} &= \frac{1}{2b}\sqrt{\frac{\mu f}{\pi g}} + \frac{a+3b}{16\pi g a b^2}, \\
R_{aa} &= \frac{1}{2a}\sqrt{\frac{\mu f}{\pi g}} - \frac{b+3a}{16\pi g b a^2}, \\
R_{ab} &= \frac{\sqrt{2}}{\sqrt{ab}}\sqrt{\frac{\mu f}{\pi g}}\, e^{-(u/2)} \cos\left(u - \frac{\pi}{4}\right), \\
\omega L_{bb} &= \frac{1}{2b}\sqrt{\frac{\mu f}{\pi g}}, \\
\omega L_{aa} &= \frac{1}{2a}\sqrt{\frac{\mu f}{\pi g}}, \\
\omega L_{ab} &= \frac{\sqrt{2}}{\sqrt{ab}}\sqrt{\frac{\mu f}{\pi g}}\, e^{-(u/2)} \cos\left(u + \frac{\pi}{4}\right), \\
|Z_{ab}| &= \frac{\sqrt{2}}{\sqrt{ab}}\sqrt{\frac{\mu f}{\pi g}}\, e^{-(u/2)}.
\end{aligned}
\tag{84}
$$

If the ratio of the diameters of the tube is not greater than 4/3, then we have the following formula for the surface transfer impedance:

$$
\frac{|Z_{ab}|}{R_{\text{d-c.}}} = \frac{u}{\sqrt{\cosh u - \cos u}}, \tag{85}
$$

which is correct to within 1 per cent at any frequency. This ratio is illustrated in Fig. 3. The ratios of the mutual resistance and the mutual reactance to the d.-c. resistance are shown in Fig. 4.

In the case of self-resistances, we let

$$
R_0 = \frac{1}{2\pi g r t}, \tag{86}
$$

where r is the high frequency radius of the tube. Thus R_0 is the d.-c. resistance of the tube if the curvature is neglected. Then we have approximately

$$\frac{R}{R_0} = \frac{u}{2} \cdot \frac{\sinh u + \sin u}{\cosh u - \cos u} \pm \frac{t}{2r}, \tag{87}$$

if the tube is fairly thin. The curvature correction is positive if the

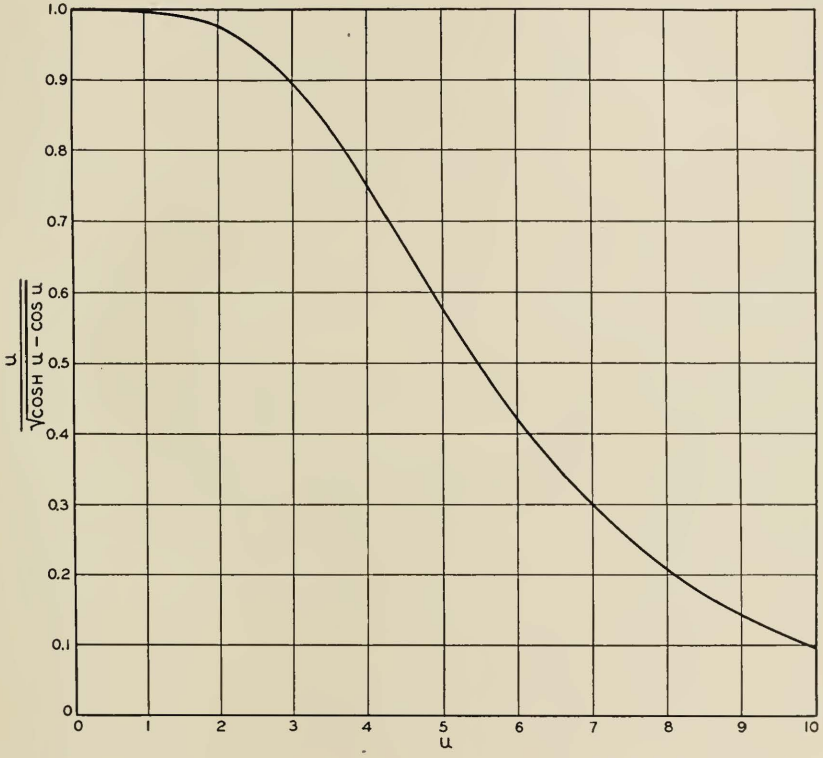


Fig. 3—The transfer impedance from one surface of a cylindrical shell to the other. The curve represents its ratio to the d-c. resistance.

return path is external, and negative if it is internal. The graph of the first term is shown in Fig. 5.

An interesting observation can be made at once from the formulæ (83) for the self-resistances of a tubular conductor. If the frequency is kept fixed and the thickness of the conductor is increased from 0, its resistance (with either return) passes through a sequence of maxima and minima.[25] The first minimum occurs when $u = \pi$, i.e., when

[25] The general fluctuating character of this function was noted by Mrs. S. P. Mead [12].

$t = \sqrt{\pi}/(2\sqrt{g\mu f})$; the first maximum occurs when $u = 2\pi$, etc. This fluctuation in resistance is due to the phase shift in the current density as we proceed from the surface of the conductor to deeper layers. The "optimum" resistance is $R_0((\pi/2) \tanh \pi/2) = 1.44R_0$, plus or

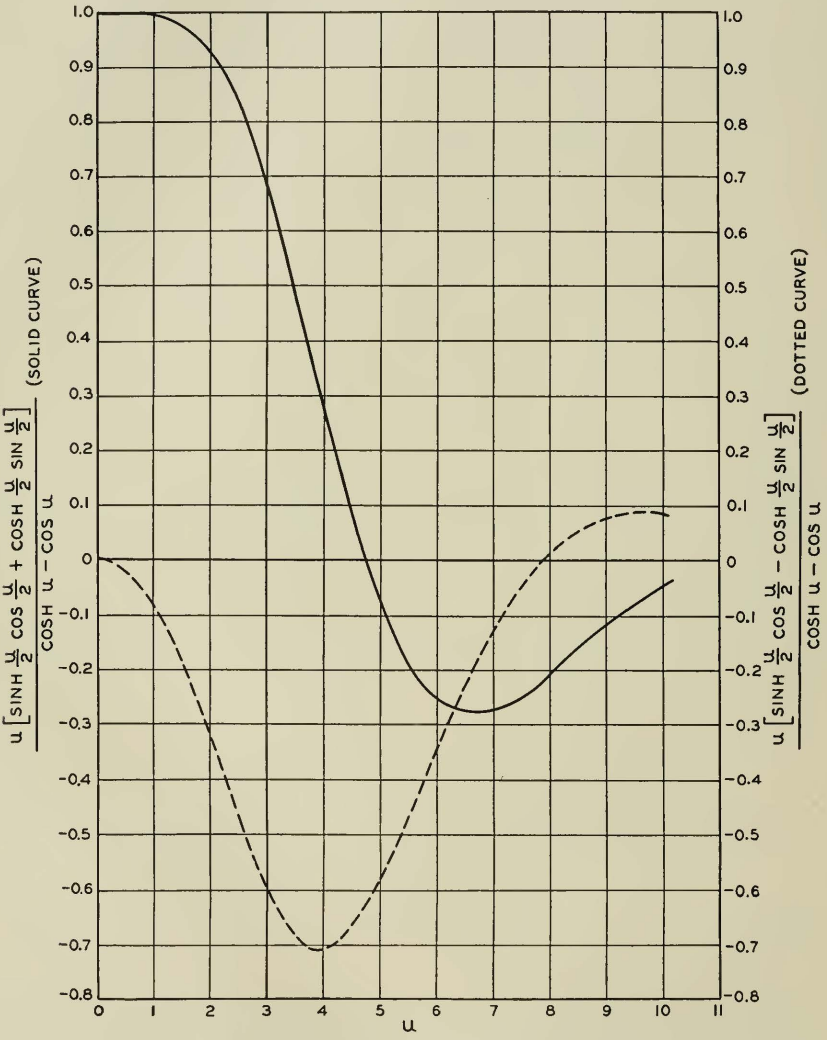


Fig. 4—The ratios of the transfer resistance and transfer reactance of a cylindrical shell to its d-c. resistance.

minus the curvature correction $t/2r$. If curvature is disregarded, the ratio of the optimum resistance to the resistance of the infinitely thick conductor with the same internal diameter as the hollow con-

ductor is $\tanh \pi/2 = 0.92$. When $u = 2\pi$, the ratio reaches its first maximum $\coth \pi = 1.004$. At 1 megacycle the optimum thickness of a copper conductor is about 0.1038 mm.

By a method of successive approximations, H. B. Dwight has ob-

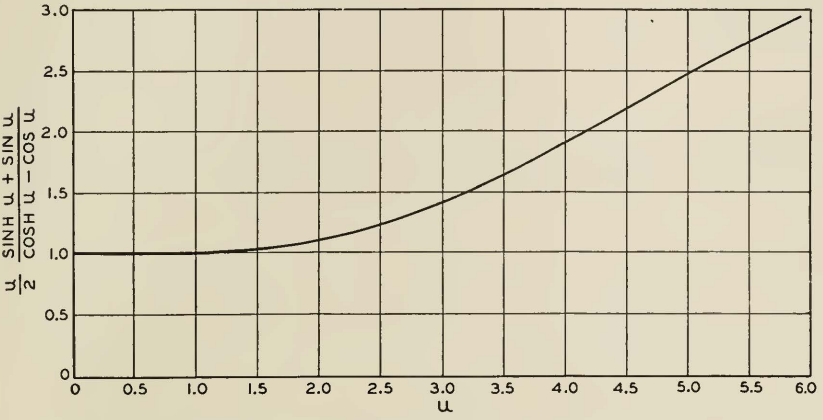


Fig. 5—The skin effect in cylindrical shells. The curve represents the ratio of the a-c. resistance of a typical shell to its d-c. resistance.

tained the impedance of a tubular conductor with an external coaxial return.[26] His final results appear as the ratio of two infinite power series, which converge for all values of the variables involved, though they can be used advantageously in numerical computations only when the frequencies are fairly low and the convergence is rapid. We shall merely indicate how Dwight's formula and other similar formulæ can be obtained directly from the exact equations (75).

Let us replace the outer radius b of (75) by $a + t$, where t is the thickness of the wall, and replace the various Bessel functions by their Taylor series in t:

$$\begin{aligned} I_0(\sigma b) &= I_0(\sigma a + \sigma t) = \sum_{n=0}^{\infty} \frac{(\sigma t)^n}{n!} I_0^{(n)}(\sigma a), \\ K_0(\sigma b) &= K_0(\sigma a + \sigma t) = \sum_{n=0}^{\infty} \frac{(\sigma t)^n}{n!} K_0^{(n)}(\sigma a), \\ I_1(\sigma b) &= I_0'(\sigma b) = \sum_{n=0}^{\infty} \frac{(\sigma t)^n}{n!} I_0^{(n+1)}(\sigma a), \\ -K_1(\sigma b) &= K_0'(\sigma b) = \sum_{n=0}^{\infty} \frac{(\sigma t)^n}{n!} K_0^{(n+1)}(\sigma a). \end{aligned} \tag{88}$$

[26] "Skin Effect in Tubular and Flat Conductors," *A. I. E. E. Journal*, Vol. 37 (1918), p. 1379.

We thus obtain

$$Z_{bb} = \frac{2\eta}{b} \frac{\sum_{n=0}^{\infty} A_n \frac{(\sigma t)^n}{n!}}{\sum_{n=0}^{\infty} A_{n+1} \frac{(\sigma t)^n}{n!}}, \tag{89}$$

where A_n is defined as

$$A_n = \begin{vmatrix} I_0'(\sigma a) & I_0^{(n)}(\sigma a) \\ K_0'(\sigma a) & K_0^{(n)}(\sigma a) \end{vmatrix}. \tag{90}$$

In spite of the complicated appearance of (90) the A's are in reality very simple functions of σa, as the accompanying list (91) will show.[27]

$$\begin{aligned} A_0 &= \frac{1}{\sigma a}, \qquad A_1 = 0, \qquad A_2 = \frac{1}{\sigma a}, \qquad A_3 = -\frac{1}{\sigma^2 a^2}, \\ A_4 &= \frac{1}{\sigma a} + \frac{3}{\sigma^3 a^3}, \qquad A_5 = -\frac{2}{\sigma^2 a^2} - \frac{12}{\sigma^4 a^4}, \\ A_6 &= \frac{1}{\sigma a} + \frac{9}{\sigma^3 a^3} + \frac{60}{\sigma^5 a^5}. \end{aligned} \tag{91}$$

The formula (90) can be made more rapidly convergent by partially summing the numerator and the denominator by means of hyperbolic functions. Thus, the numerator becomes

$$\sqrt{\frac{a}{b}} \cosh \sigma t + \frac{\sinh \sigma t}{2\sigma a}\left[1 - \frac{3t}{4a} + \cdots \right] - \frac{3t^2}{a^2} + \cdots,$$

and the denominator

$$\sqrt{\frac{a}{b}} \sinh \sigma t + \frac{3(\sigma t \cosh \sigma t - \sinh \sigma t)}{8(\sigma a)^2} + \cdots.$$

The reader will readily see that there would be no difficulty in using this method to obtain other expansions somewhat similar to (89). For example, we might write $a = b - t$ in (75) and express our results in terms of the outer radius. In this respect the method that we have used has greater flexibility than Dwight's; but there seems to be little advantage gained from it, since the simple formulæ (82) are sufficient for most practical purposes.

[27] The values given in (91) are *exact*, not approximate. One of them, namely,

$$A_1 = \begin{vmatrix} I_0'(\sigma a) & I_0(\sigma a) \\ K_0'(\sigma a) & K_0(\sigma a) \end{vmatrix} \equiv \frac{1}{\sigma a},$$

is one of the fundamental identities found in all books on Bessel functions. The rest are consequences of analogous, though ess familiar, identities. The general expressions for the coefficients A_n were obtained by H. Pleijel [20].

Internal Impedances of Laminated Conductors

So far we have supposed that all conductors were homogeneous. We shall now consider a somewhat more general conductor composed of n coaxial layers of different substances. As before, we are interested in finding expressions for the internal impedances; besides, we may wish to know how the total current is distributed between the different layers of the conductor.

To begin with, let us suppose that a coaxial return path is provided *outside* the given conductor. We number our layers consecutively and call the inner layer the first. Let $Z_{aa}{}^{(m)}$ and $Z_{bb}{}^{(m)}$ be the surface impedances of the mth layer, the first when the return is internal, the other when it is external; and let $Z_{ab}{}^{(m)}$ be the transfer impedance from one surface to the other. Formulæ for these impedances have already been obtained in the section under "The Surface Impedances of Hollow Cylindrical Shells," page 552. Also, let $z_{bb}{}^{(m)}$ be the *surface impedance of the first m layers* with external return; that is, the ratio of the longitudinal electromotive intensity at the outer surface of the mth layer to the total current I_m in all m layers.[28]

By hypothesis, there is no return path inside the laminated conductor as a whole. Hence, when we fix our attention on any one layer alone, say the mth, we may say that the current *in this layer* returns partly through the $m - 1$ layers within it, and partly outside. In the $m - 1$ inner layers, however, the current is assumed to be I_{m-1} in the outward direction—or what amounts to the same thing $-I_{m-1}$ in the return direction. Hence we conclude that, of the current $I_m - I_{m-1}$ in the layer under discussion, I_m returns *outside* and $-I_{m-1}$ *inside*. Substituting these values in Theorem 2 on page 555, we find that the electromotive intensity along the inner surface of the layer is $Z_{ab}{}^{(m)}I_m - Z_{aa}{}^{(m)}I_{m-1}$.

But the *inner* surface of the mth layer is the *outer* surface of the composite conductor comprising the $m - 1$ inner layers, and by Theorem 1, the electromotive intensity on this outer surface is $z_{bb}{}^{(m-1)}I_{m-1}$. As the two must be equal, we obtain an equation from which we can determine the ratio of the current flowing in the first $m - 1$ layers to that flowing in m layers. This is

$$\frac{I_{m-1}}{I_m} = \frac{Z_{ab}{}^{(m)}}{Z_{aa}{}^{(m)} + z_{bb}{}^{(m-1)}}. \tag{92}$$

In this formula for the effect of an extra layer on the current dis-

[28] In this notation, the current flowing in the mth layer is $I_m - I_{m-1}$. It should also be noted that $Z_{bb}{}^{(1)} = z_{bb}{}^{(1)}$.

tribution, it will be noted that the denominator is the impedance (with internal return) of the added layer plus the original impedance.

We now consider the electromotive intensity on the outer surface of the mth layer, which is $z_{bb}^{(m)} I_m$ on the one hand, and $(Z_{bb}^{(m)} I_m - Z_{ab}^{(m)} I_{m-1})$ on the other. Thus, we have the following equation,

$$z_{bb}^{(m)} = Z_{bb}^{(m)} - Z_{ab}^{(m)} \frac{I_{m-1}}{I_m} = Z_{bb}^{(m)} - \frac{[Z_{ab}^{(m)}]^2}{Z_{aa}^{(m)} + z_{bb}^{(m-1)}}, \quad (93)$$

expressing the effect of an additional layer upon the impedance of the conductor.

This equation is a convenient reduction formula. Starting with the first layer (for which $z_{bb}^{(1)} = Z_{bb}^{(1)}$), we add the remaining layers one by one and thus obtain the impedance of the complete conductor in the form of the following continued fraction:

$$z_{bb}^{(n)} = Z_{bb}^{(n)} - \frac{[Z_{ab}^{(n)}]^2}{Z_{aa}^{(n)} + Z_{bb}^{(n-1)} +} \frac{[Z_{ab}^{(n-1)}]^2}{Z_{aa}^{(n-1)} + Z_{bb}^{(n-2)} +} \cdots \frac{[Z_{ab}^{(2)}]^2}{Z_{aa}^{(2)} + Z_{bb}^{(1)}}. \quad (94)$$

We can also get a reduction formula for the transfer impedance between the inner and outer surfaces of the composite conductor formed by the first m layers. To do so, it is only necessary to note that, since the inner surface of the first $m - 1$ layers is also the inner surface of the first m layers as well, the electromotive intensity on that surface can be expressed either as $z_{ab}^{(m-1)} I_{m-1}$ or as $z_{ab}^{(m)} I_m$. Thus, we have

$$z_{ab}^{(m)} = z_{ab}^{(m-1)} \frac{I_{m-1}}{I_m} = \frac{Z_{ab}^{(m)} z_{ab}^{(m-1)}}{Z_{aa}^{(m)} + z_{bb}^{(m-1)}}. \quad (95)$$

By noting that $z_{ab}^{(1)} = Z_{ab}^{(1)}$, we can determine successively the transfer impedances across the first two layers, the first three, and so on. This formula is not quite as simple as (94), owing to the presence of $z_{bb}^{(m-1)}$ in its denominator, and it is therefore not expedient to evaluate $z_{ab}^{(m)}$ explicitly; but it is not prohibitively cumbersome from the numerical standpoint when the computations are made step by step.

Although in deducing equations (93) and (95) we supposed that the added layer was homogeneous, the equations are correct even if this layer consists of several coaxial layers, provided $Z_{aa}^{(m+1)}$ and $Z_{ab}^{(m+1)}$ are interpreted as the impedances of the added non-homogeneous layer in the absence of the original core of m layers. These latter

impedances themselves have to be computed by means of equations (94) and (95).

If the return path is inside the laminated conductor, instead of outside, formulæ (92) and (93) still hold, provided we interchange a and b, and count layers from the outside instead of the inside, so that $m = 1$ is the outermost, rather than the innermost, layer.

The basic rule for determining the surface impedances of laminated conductors can be put into the following verbal form:

Theorem 3: Let two conductors, both of which may be made up of coaxial layers, fit tightly one inside the other. Any surface self-impedance of the compound conductor equals the individual impedance of the conductor nearest to the return path diminished by the fraction whose numerator is the square of the transfer impedance across this conductor and whose denominator is the sum of the surface impedances of the two component conductors if each is regarded as the return path for the other. The transfer impedance of the compound conductor is the fraction whose numerator is the product of the transfer impedances of the individual conductors and whose denominator is that of the self-impedance.

If two coaxial conductors are short-circuited at intervals, short compared to the wave-length, the above theorem holds even if the conductors do not fit tightly one over the other, provided we add in the denominators a third term representing the inductive reactance of the space between the conductors.

Disks as Terminal Impedances for Coaxial Pairs

So far we have been concerned only with infinitely long pairs. We now take up a problem of a different sort; namely, the design of a disk which, when clapped on the end of such a pair, will not give rise to a reflected wave.

The line of argument will be as follows: To begin with, we shall assume a disk of arbitrary thickness h, compute the field which will be set up in it, and then adjust the thickness so as to make this field match that which would exist in the dielectric of an infinite line.

The field in the disk has to satisfy equation (2) where $i\omega\epsilon$ can be disregarded by comparison with g. Thus, we have

$$\frac{\partial H_\varphi}{\partial z} = -gE_\rho, \qquad \frac{1}{\rho}\frac{(\rho H_\varphi)}{\partial \rho} = gE_z,$$
$$\frac{\partial E_z}{\partial \rho} - \frac{\partial E_\rho}{\partial z} = i\omega\mu H_\varphi. \tag{96}$$

In the dielectric between the coaxial conductors, the longitudinal displacement current density is very small; in fact, it would be zero if the conductors were perfect. This current density is continuous across the surface of the disk and, therefore, gE_z is exceedingly small. Hence, the second of the above equations becomes approximately

$$\frac{\partial(\rho H_\varphi)}{\partial \rho} = 0; \tag{97}$$

so that

$$H_\varphi = \frac{P}{\rho}, \tag{98}$$

where P is independent of ρ but may be a function of z. Under these conditions, the remaining two equations are

$$\frac{\partial E_\rho}{\partial z} = -i\omega\mu H_\varphi, \qquad \frac{\partial H_\varphi}{\partial z} = -gE_\rho. \tag{99}$$

From the form of these equations and from (98), we conclude that the general expressions for the intensities in the disk are

$$H_\varphi = \frac{Ae^{\sigma z} + Be^{-\sigma z}}{\rho}, \qquad E_\rho = \frac{\sigma[Be^{-\sigma z} - Ae^{\sigma z}]}{g\rho}, \tag{100}$$

where $\sigma = \sqrt{g\omega\mu i}$.

On the outside flat surface of the disk (given by $z = h$ where h is the thickness of the plate), the magnetomotive intensity is very nearly zero;[29] therefore,

$$Ae^{\sigma h} + Be^{-\sigma h} = 0. \tag{101}$$

From this we obtain

$$A = -Ce^{-\sigma h}, \qquad B = Ce^{\sigma h}, \tag{102}$$

where C is some constant. Thus equations (100) can be written as follows:

$$H_\varphi = \frac{C \sinh \sigma(h - z)}{\rho},$$

$$E_\rho = \frac{\sigma C \cosh \sigma(h - z)}{g\rho}; \tag{103}$$

and at the boundary between the disk and the dielectric of the transmission line ($z = 0$), we have

$$\frac{E_\rho}{H_\varphi} = \frac{\sigma}{g} \coth \sigma h. \tag{104}$$

[29] On account of the negligibly small longitudinal current in the disk.

On the other hand, if there is to be no reflection this must equal $\sqrt{\mu/\epsilon}$ by equation (24). Hence

$$\frac{\sigma}{g}\coth \sigma h = \sqrt{\frac{\mu}{\epsilon}}. \tag{105}$$

If σh is small, coth σh equals approximately $1/\sigma h$, and

$$h = \frac{1}{g}\sqrt{\frac{\epsilon}{\mu}}\text{ cm}. \tag{106}$$

Under these conditions, the generalized flux of energy across the inner surface of the disk is, in accordance with the text under "The Complex Poynting Vector," page 555, and equation (14),

$$\int_0^{2\pi}\int_{b'}^{a''} E_\rho H_\varphi{}^* \rho\, d\rho\, d\varphi = \frac{1}{2\pi}\sqrt{\frac{\mu}{\epsilon}}\log\frac{a''}{b'}\, I^2. \tag{107}$$

Thus, the impedance of this disk is a pure resistance equal to the characteristic impedance of the coaxial pair.

Cylindrical Waves and the Problem of Cylindrical Shields [30]

It is well known that when two transmission lines are side by side, to a greater or lesser extent they interfere with each other. This interference is usually analyzed into "electromagnetic crosstalk" and "electrostatic crosstalk."

Thus, electric currents in a pair of parallel wires produce a magnetic field with lines of force perpendicular to the wires. These lines cut the other pair of wires and induce in them electromotive forces, thereby producing what is usually called the "electromagnetic crosstalk"; this crosstalk is seen to be proportional to the current flowing in the first pair. The "electrostatic crosstalk," on the other hand, is caused by electric charges induced on the wires of the second system; these charges are proportional to the potential difference existing between the wires of the "disturbing" transmission line.

The distinction between two types of crosstalk is valid, although the terminology is somewhat unfortunate; the word "electromagnetic" is used in too narrow a sense and the word "electrostatic" is a contradiction in terms since electric currents and charges in a transmission line are variable. The terms "impedance crosstalk" and "admittance crosstalk" would be preferable because the former is due to a distributed mutual series impedance between two lines and the latter is produced by a distributed mutual shunt admittance.

[30] Since this paper was written, a related paper has been published by Louis V. King [18]. However, the physical picture here developed appears to be new. The earliest writer who treated the problem of electromagnetic shielding is H. Pleijel [21].

The crosstalk between two parallel pairs (this applies to twisted pairs as well) can be reduced by enclosing each pair in a cylindrical metallic shield. It is the object of this and the following two sections to develop a theory for the design of such shields.

This theory is based upon an assumption that in so far as the radial movement of energy toward and away from the wires is concerned we can disregard the non-uniform distribution of currents and charges along the length of the wires. No serious error is introduced thereby as long as the radius of the shield is small by comparison with the wave-length. The field around the wires is considered, therefore, as due to superposition of two two-dimensional fields of the types given by equations (4) and (5).

The actual computation of the effectiveness of a given shield will be reduced to an analogous problem in Transmission Line Theory.

Equations (4) and (5) are too general as they stand. Strictly speaking the effect of a shield upon an arbitrary two-dimensional field cannot be expressed by a single number. The field at various points outside the shield will be reduced by it in different ratios. However, any such field can be resolved into "cylindrical waves," each of which is reduced by the shield everywhere in the same ratio. Moreover, to all practical purposes the field produced by electric currents (or electric charges) in a pair of wires is just such a pure cylindrical wave.

Since both E and H are periodic functions of the coordinate φ, they can be resolved into Fourier series. The name "cylindrical waves" will be applied to the fields represented by the separate terms of the series. As the name indicates the wave fronts of these waves are cylindrical surfaces, although owing to relatively low frequencies and long wave-lengths used in practice the progressive motion of these waves is not clearly manifested except at great distances from the wires.

Turning our attention specifically to magnetic cylindrical waves *of the nth order*, and writing the field components tangential to the wave fronts in the form $E \cos n\varphi$ and $H \cos n\varphi$, we have from equations (5):

$$\frac{dE}{d\rho} = -i\omega\mu H, \quad \frac{d(\rho H)}{d\rho} = -\left[(g + i\omega\epsilon)\rho + \frac{n^2}{i\omega\mu\rho}\right]E. \quad (108)$$

From these we obtain

$$\rho^2\frac{d^2E}{d\rho^2} + \rho\frac{dE}{d\rho} = [i\omega\mu(g + i\omega\epsilon)\rho^2 + n^2]E. \quad (109)$$

This equation, being of the second order, possesses two independent solutions: one for diverging cylindrical waves and the other for reflected waves. The ratio of E to H in the first case and its negative in the second will be called the *radial impedance* offered by the medium to cylindrical waves.

In the next section we shall determine radial impedances in dielectrics and metals and show that for all practical purposes the attenuation of cylindrical waves in metals is exponential. The significance of the radial impedance is the same as that of the characteristic impedance of a transmission line. When a cylindrical wave passes from one medium into another, a *reflection* takes place unless the radial impedances are the same in the two media. Thus if E_0 and H_0 are the *impressed* intensities (at the boundary between the two media), E_r and H_r the reflected and E_t and H_t, the transmitted intensities, we have

$$E_0 + E_r = E_t \qquad \text{and} \qquad H_0 + H_r = H_t, \tag{110}$$

since both intensities must be continuous. On the other hand, if k is the ratio of the impedance in the first medium to that in the second, then equations (110) become

$$kH_0 - kH_r = H_t \qquad \text{and} \qquad H_0 + H_r = H_t. \tag{111}$$

Solving we obtain

$$H_t = \frac{2k}{k+1} H_0 \qquad \text{and} \qquad E_t = \frac{2}{k+1} E_0. \tag{112}$$

The reflection loss will be defined as

$$R = 20 \log_{10} \frac{|H_0|}{|H_t|} \text{ decibels.} \tag{113}$$

When a wave passes through a shield, it encounters two boundaries and if the shield is electrically thick, that is, if the attenuation of the wave in the shield is so great that secondary reflections can be disregarded without introducing a serious error, the total reflection loss is the sum of the losses at each boundary. The first loss can be computed directly from (112) and the second from the same equation if we replace k by its reciprocal. Thus, the total reflection loss for electrically thick shields is

$$R = 20 \log_{10} \frac{|k+1|^2}{4|k|} \text{ decibels.} \tag{114}$$

When the ratio of the impedances is very large by comparison with unity, the formula becomes

$$R = 20 \log_{10} \frac{|k|}{4}, \tag{115}$$

and when k is very small, then

$$R = 20 \log_{10} \frac{1}{4|k|}. \tag{116}$$

In the next section we shall see that to all practical purposes, the wave in the shield is attenuated exponentially. If α is the attenuation constant in nepers and if t is the thickness of the shield, then the attenuation loss is

$$A = 8.686\alpha t \text{ decibels} \tag{117}$$

and the total reduction in the magnetomotive intensity due to the presence of the shield is

$$S = R + A. \tag{118}$$

The electromotive intensity is reduced in the same ratio.

But if the shield is not electrically thick, a correction term has to be added to the reflection loss. This correction term can be shown to be [31]

$$C = 20 \log_{10} \left| 1 - \frac{(k-1)^2}{(k+1)^2} e^{-2\Gamma t} \right| \text{ decibels}, \tag{119}$$

and if k is very large or very small by comparison with unity then

$$C \doteq 3 - 8.686\alpha t + 10 \log_{10} (\cosh 2\alpha t - \cos 2\beta t). \tag{120}$$

Equation (120) does not hold down to $t = 0$; when Γt is nearly zero, then

$$C = 20 \log_{10} \left| 1 - \frac{(k-1)^2}{(k+1)^2} \right|. \tag{121}$$

So far we supposed that the shields were coaxial with the source. If this is not so, it is always possible to replace any given line source within the shield by an equivalent system of line sources coaxial with the shield and emitting cylindrical waves of proper orders. Mathematically this amounts to a change of the origin of the coordinate system. In the next section we shall see that the shielding effectiveness is not the same for all cylindrical waves. This means, of course, that if the shield is not coaxial with the source, the total reduction in

[31] Here, $\Gamma = \alpha + i\beta$ is the propagation constant in the shield.

the field depends upon the position of the measuring apparatus. The variation is very small, however, unless the source is almost touching the shield and it can be stated that approximately the shielding effectiveness is independent of the position of the source.

It is interesting to observe from the accompanying tables that while the attenuation loss is greater in iron than in copper, the reflection loss is greater at a copper surface. In fact, at some frequencies the impedances of iron and air nearly match and practically no reflection takes place. Hence, a thin copper shield may be more effective than an equally thin iron shield. And if a composite shield is made of copper and iron, the shield will be more effective if copper layers are placed on the outside to take advantage of the added reflection.

TABLE I

THE ABSOLUTE VALUE OF THE RADIAL IMPEDANCE OFFERED BY AIR TO CYLINDRICAL MAGNETIC WAVES OF THE FIRST ORDER (IN MICROHMS)

f	Radius = 0.5 cm.	1 cm.	2 cm.
1 cycle	0.0395	0.07896	0.1579
10 cycles	0.395	0.790	1.58
100 cycles	3.95	7.90	15 8
1 kilocycle	39.5	79.0	158.
10 kilocycles	395.	790.	1,580.
100 kilocycles	3,950.	7,900.	15,800.
1 megacycle	39,500.	79,000.	158,000.
10 megacycles	395,500.	790,000.	1.58 ohms
100 megacycles	3.95 ohms	7.9 ohms	15.8 ohms

TABLE II

THE INTRINSIC IMPEDANCE OF CERTAIN METALS $(\eta)/(\sqrt{i})$ IN MICROHMS

f	Copper $g = 5.8005 \times 10^5$ mhos/cm. $\mu = 0.01257\ \mu h$/cm.	Lead $g = 4.8077 \times 10^4$ mhos/cm. $\mu = 0.01257\ \mu h$/cm.	Aluminum $g = \frac{1}{3} 10^6$ mhos/cm. $\mu = 0.01257\ \mu h$/cm.	Iron $g = 10^5$ mhos/cm. $\mu = 1.257\ \mu h$/cm. = (100 relative to copper)
1 cycle	0.369	1.28	0.487	8.88
10 cycles	1.17	4.05	1.54	28.1
100 cycles	3.69	12.8	4.87	88.8
1 kilocycle	11.7	40.5	15.4	281.
10 kilocycles	36.9	128.	48.7	888.
100 kilocycles	117.	405.	154.	2,810.
1 megacycle	369.	1,280.	487.	8,880.
10 megacycles	1,170.	4,050.	1,540.	28,100.
100 megacycles	3,690.	12,800.	4,870.	88,800.

TABLE III

The Intrinsic Propagation Constant of Certain Metals

f	Copper $g = 5.8005 \times 10^5$ mhos cm. $\mu = 0.01257$ μh/cm.		Lead $g = 4.8077 \times 10^4$ mhos/cm. $\mu = 0.01257$ μh/cm.		Aluminum $g = \frac{1}{3}10^6$ mhos/cm. $\mu = 0.01257$ μh/cm.		Iron $g = 10^5$ mhos/cm. $\mu = 1.257$ μh/cm. = (100 relative to copper)	
	$\frac{\sigma}{\sqrt{i}}$ in $\frac{\text{nepers}}{\text{cm.}}$	$\frac{\sigma}{\sqrt{i}}$ in db/cm.	$\frac{\sigma}{\sqrt{i}}$ in $\frac{\text{nepers}}{\text{cm.}}$	$\frac{\sigma}{\sqrt{i}}$ in db/cm.	$\frac{\sigma}{\sqrt{i}}$ in $\frac{\text{nepers}}{\text{cm.}}$	$\frac{\sigma}{\sqrt{i}}$ in db/cm.	$\frac{\sigma}{\sqrt{i}}$ in $\frac{\text{nepers}}{\text{cm.}}$	$\frac{\sigma}{\sqrt{i}}$ in db/cm.
1 cycle	0.214	1.86	0.0616	.535	0.162	1.41	0.888	7.72
10 cycles	0.677	5.88	0.195	1.69	0.513	4.46	2.81	24.4
100 cycles	2.14	18.6	0.616	5.35	1.62	14.1	8.88	77.2
1 kilocycle	6.77	58.8	1.95	16.9	5.13	44.6	28.1	244.
10 kilocycles	21.4	186.	6.16	53.5	16.2	141.	88.8	772.
100 kilocycles	67.7	588.	19.5	169.	51.3	446.	281.	2,440.
1 megacycle	214.	1,860.	61.6	535.	162.	1,410.	888.	7,720.
10 megacycles	677.	5,880.	195.	1690.	513.	4,460.	2810.	24,400.
100 megacycles	2140.	18,600.	616.	5350.	1620.	14,100.	8880.	77,200.

Cylindrical Waves in Dielectrics and Metals

In good dielectrics g is small by comparison with $\omega\epsilon$ and the first term on the right in (109) very nearly equals $(2\pi\rho/\lambda)^2$ where λ is the wave-length. But we are interested in wave-lengths measured in miles and shields with diameters measured in inches; thus we shall write (109) in the following approximate form:

$$\rho^2 \frac{d^2E}{d\rho^2} + \rho \frac{dE}{d\rho} - n^2E = 0. \tag{122}$$

When $n \neq 0$ there are two independent solutions

$$E_1 = \rho^{-n} \quad \text{and} \quad E_2 = \rho^n; \tag{123}$$

and when $n = 0$,

$$E_1 = \log \rho \quad \text{and} \quad E_2 = 1. \tag{124}$$

The corresponding expressions for H are, by (108),

$$H_1 = \frac{n\rho^{-n-1}}{i\omega\mu} \quad \text{and} \quad H_2 = -\frac{n\rho^{n-1}}{i\omega\mu}, \tag{125}$$

in the first case, and

$$H_1 = \frac{1}{i\omega\mu\rho} \quad \text{and} \quad H_2 = 0, \tag{126}$$

in the second.

The second case in which E_1 and H_1 are the electromotive and magnetomotive intensities in the neighborhood of an isolated wire carrying electric current is of interest to us only in so far as it helps to interpret (123) and (125). If we were to consider $2n$ infinitesimally thin wires equidistributed upon the surface of an infinitely narrow cylinder, the adjacent wires carrying equal but oppositely directed currents of strength sufficient to make the field different from zero, and calculate the field, we should obtain expressions proportional to E_1 and H_1. An actual cluster of $2n$ wires close together would generate principally a cylindrical wave of order n; the strengths of other component waves of order $3n$, $5n$, etc. rapidly diminish as the distance from the cluster becomes large by comparison with the distance between the adjacent wires of the cluster. For the purposes of shielding design we can regard a pair of wires as generating a cylindrical wave of the first order ($n = 1$). The radial impedance of an nth order wave is

$$Z_\rho = \frac{E_1}{H_1} = \frac{i\omega\mu\rho}{n}, \tag{127}$$

and that of the corresponding reflected wave has the same value. It should be noted that by the "reflected" cylindrical wave in the space enclosed by a shield, we mean the sum total of an infinite number of successive reflections. Each of the latter waves condenses on the axis and diverges again only to be re-reflected back; in a steady state all these reflected waves interfere with each other and form what might be called a "stationary reflected wave." Not being interested in any other kind of reflected waves we took the liberty of omitting the qualification.

In conductors the attenuation of a wave due to energy dissipation is much greater (except at extremely low frequencies) than that due to the cylindrical divergence of the wave. Hence, in the shield we can regard the wave as plane and write (108) in the following approximate form:

$$\frac{dE}{d\rho} = -i\omega\mu H, \qquad \frac{dH}{d\rho} = -gE. \tag{128}$$

In form, these are exactly like ordinary transmission line equations. Hence, in a shield the radial impedance is simply the intrinsic impedance of the metal,

$$Z_\rho = \eta = \sqrt{\frac{i\omega\mu}{g}} \text{ ohms}, \tag{129}$$

and the propagation constant,

$$\sigma = \sqrt{i\omega\mu g} = \sqrt{\pi f\mu g}\,(1 + i) \text{ nepers/cm}. \tag{130}$$

The exact value of the radial impedance in metals can be found by solving (108). Thus, we can obtain

$$Z_\rho = -\eta \frac{K_n(\sigma\rho)}{K_n'(\sigma\rho)} \tag{131}$$

for diverging waves, and

$$Z_\rho = \eta \frac{I_n(\sigma\rho)}{I_n'(\sigma\rho)} \tag{132}$$

for the reflected waves.

Cylindrical waves of the electric type can be treated in the same manner. It turns out that the transmission laws in metals are identical with those for magnetic waves. The radial impedance in perfect dielectrics, on the other hand is given by

$$Z_\rho = \frac{n}{i\omega\epsilon\rho}. \tag{133}$$

This is enormous by comparison with the impedance in metals, thereby explaining an almost perfect "electrostatic" shielding offered by metallic substances. Even when the frequency is as high as 100 kc. the radial impedance of air 1 cm. from the source is about 36×10^6 ohms while the impedance of a copper shield is only 117×10^{-6} ohms. The reflection loss is approximately 220 db.

Power Losses in Shields

As we have shown in the text under "The Complex Poynting Vector," page 555, the average power dissipated in a conductor is the real part of the integral $\Phi = 1/2 \int\int [EH^*]_n dS$ taken over the surface of the conductor. If the source of energy is inside a shield, the integration need be extended only over its inner surface, because the average energy flowing outward through this surface is almost entirely dissipated in the shield, the radiation loss being altogether negligible. If a cylindrical wave whose intensities at the inner surface of the shield of radius "a" are

$$H_\varphi = H_0 \cos n\varphi, \qquad H_\rho = H_0 \sin n\varphi, \qquad E_z = \eta H_\varphi, \tag{134}$$

$\eta = i\omega\mu a/n$ being the radial impedance in the dielectric, is impressed upon the inner surface of the shield, a reflected wave is set up. The resultant of the magnetomotive intensities in the two is readily found to be $(2k/k + 1)H_0$, where k is the ratio of the radial impedance of the dielectric column inside the shield to the impedance Z looking into the shield. If the shield is electrically thick, the impedance Z is obviously the radial impedance of the shield; otherwise it is modified somewhat by reflection from the outside of the shield. The average power loss in the shield per centimeter of length is, then, the real part of

$$\Phi = \frac{2\pi a k k^* Z}{(k + 1)(k^* + 1)} H_0 H_0^*. \tag{135}$$

This becomes simply

$$\Phi = 2\pi a Z H_0 H_0^*, \tag{136}$$

if the frequency is so high that k is large as compared with unity.

If the source of the impressed field is a pair of wires along the axis of the shield, the magnetomotive intensity on the surface of the shield can be shown to be

$$H_\varphi = \frac{l}{2\pi a^2} I \cos \varphi, \tag{137}$$

where l is the separation between the axes of the wires. Therefore,

$$\Phi = \frac{kk^*l^2Z}{2\pi a^3(k+1)(k^*+1)} I^2. \tag{138}$$

Resistance of Nearly Coaxial Tubular Conductors

When two tubular conductors are not quite coaxial, a proximity effect [32] appears which disturbs the symmetry of current distribution and therefore somewhat increases their resistance. This effect can be estimated by the following method of successive approximations. To begin with, we assume a symmetrical current distribution in the *inner* conductor. The magnetic field outside this conductor is then the same as that of a simple source along its axis and can be replaced by an equivalent distribution of sources situated along the axis of the outer conductor. The principal component of this distribution is a simple source of the same strength as the actual source and does not enter into the proximity effect. The next largest component is a double source given by

$$\begin{aligned} E_z &= \frac{i\omega\mu lI}{2\pi r}\cos\theta, \\ H_\theta &= \frac{lI}{2\pi r^2}\cos\theta, \end{aligned} \tag{139}$$

where l is the interaxial separation, r is the distance of a typical point of the field from the axis of the outer conductor, and θ is the remaining polar coordinate.

This field is impressed upon the inner surface of the outer conductor [33] and the resulting power loss equals, by equation (136), the real part of

$$\Phi = 2\pi a\eta\left(\frac{l}{2\pi a^2}\right)^2 I^2 = \frac{l^2}{2\pi a^3}\eta I^2, \tag{140}$$

where at high frequencies $\eta = \sqrt{i\omega\mu/g}$ is simply the intrinsic impedance of the outer conductor.[34] This loss increases the resistance of the outer tube by the amount,

$$\Delta R_a = \frac{l^2}{a^3}\sqrt{\frac{\mu f}{\pi g}}, \tag{141}$$

[32] For promixity effect in parallel wires external to each other, the reader is referred to the following papers: John R. Carson [1], C. Manneback [9], S. P. Mead [12].

[33] The radius of this surface is designated by a.

[34] At low frequencies η has to be replaced by the radial impedance looking into the shield.

in excess of the concentric resistance $R_a = (1/2a)\sqrt{\mu f/\pi g}$ given by (84). The relative increase is, therefore,

$$\frac{\Delta R_a}{R_a} = \frac{2l^2}{a^2}. \tag{142}$$

The magnetic field (139) is partially reflected from the outer tube, impressed upon the inner conductor, partially refracted into it and dissipated there. Using (110) and (111) we can show that the reflected field is

$$\begin{aligned} H_\theta &= \frac{lI}{2\pi a^2}\cos\theta, \\ E_z &= \frac{i\omega\mu lI}{2\pi a^2}\rho\cos\theta. \end{aligned} \tag{143}$$

This field converges to the axis of the outer conductor. In order to estimate its effect upon the inner conductor, it is convenient to replace it by an equivalent field converging toward the axis of the inner conductor. By properly changing the origin of the coordinate system this equivalent field can be shown to be

$$\begin{aligned} E_z &= -\frac{i\omega\mu lI}{2\pi a^2}(l + \rho\cos\varphi), \\ H_\varphi &= -\frac{lI}{2\pi a^2}\cos\varphi. \end{aligned} \tag{144}$$

Applying once more (138) (replacing there a by the radius b of the inner conductor), we find that the power loss due to this field is given by the real part of

$$\Phi = \frac{bl^2}{2\pi a^4} I^2, \tag{145}$$

so that the absolute increase in resistance of the inner conductor is

$$\Delta R_b = \frac{bl^2}{a^4}\sqrt{\frac{\mu f}{\pi g}} \tag{146}$$

which must be added to the concentric resistance of the inner conductor $R_b = (1/2b)\sqrt{\mu f/\pi g}$. The relative increase is therefore

$$\frac{\Delta R_b}{R_b} = 2\left(\frac{b}{a}\right)^2\left(\frac{l}{a}\right)^2. \tag{147}$$

It is unnecessary to carry the process further.

Considering the pair as a whole, the resistance when concentric is $R = R_a + R_b$, and the increase due to eccentricity is $\Delta R = \Delta R_a + \Delta R_b$ thus giving a percentage increase,

$$\frac{\Delta R}{R} = 2\left(\frac{b}{a}\right)\left(\frac{l}{a}\right)^2. \tag{148}$$

It is obvious that, so long as b and l are small compared with a, this percentage increase is very small.

From the well-known formulæ for the inductance and the capacity between parallel cylindrical conductors, we find that the characteristic impedance of a nearly coaxial pair is given in terms of the characteristic impedance of the coaxial pair by

$$Z = Z_0\left[1 - \frac{e^2k^2}{(k^2 - 1)\log k}\right], \tag{149}$$

where the "eccentricity" e is defined as the ratio of the interaxial separation to the inner radius of the outer conductor and k as the ratio of the inner radius of the outer conductor to the outer radius of the inner conductor. Combining (149) and (148) we have for the attenuation of the nearly coaxial pair:

$$\alpha = \alpha_0\left[1 + \frac{2e^2}{k} + \frac{e^2k^2}{(k^2 - 1)\log k}\right]. \tag{150}$$

References

Papers

1. John R. Carson, "Wave Propagation Over Parallel Wires: The Proximity Effect," *Phil. Mag.*, Vol. 41, Series 6, pp. 607–633, April, 1921.
2. John R. Carson and J. J. Gilbert, "Transmission Characteristics of the Submarine Cable," *Jour. Franklin Institute*, p. 705, December, 1921.
3. John R. Carson and J. J. Gilbert, "Transmission Characteristics of the Submarine Cable," *Bell Sys. Tech. Jour.*, pp. 88–115, July, 1922.
4. John R. Carson, "The Guided and Radiated Energy in Wire Transmission," *A.I.E.E. Jour.*, pp. 908–913, October, 1924.
5. John R. Carson, "Electromagnetic Theory and the Foundations of the Electric Circuit Theory," *Bell Sys. Tech. Jour.*, January, 1927.
6. S. Butterworth, "Eddy Current Losses in Cylindrical Conductors, with Special Applications to the Alternating Current Resistances of Short Coils," *Phil. Trans., Royal Soc. of London*, pp. 57–100, September, 1921.
7. H. B. Dwight, "Skin Effect and Proximity Effect in Tubular Conductors," *A.I.E.E. Jour.*, Vol. 41, pp. 203–209, March 1922.
8. H. B. Dwight, "Skin Effect and Proximity Effect in Tubular Conductors," *A.I.E.E. Trans.*, Vol. 41, pp. 189–195, 1922.
9. C. Manneback, "An Integral Equation for Skin Effect in Parallel Conductors," *Jour. of Math. and Physics*, April, 1922.
10. H. B. Dwight, "A Precise Method of Calculation of Skin Effect in Isolated Tubes," *A.I.E.E. Jour.*, Vol. 42, pp. 827–831, August, 1923.
11. H. B. Dwight, "Proximity Effect in Wires and Thin Tubes," *A.I.E.E. Jour.*, Vol. 42, pp. 961–970, September, 1923; *Trans.*, Vol. 42, pp. 850–859, 1923.

12. Mrs. S. P. Mead, "Wave Propagation Over Parallel Tubular Conductors: The Alternating Current Resistance," *Bell Sys. Tech. Jour.*, pp. 327–338, April, 1925.
13. Chester Snow, "Alternating Current Distribution in Cylindrical Conductors," *Scientific Papers of the Bureau of Standards*, No. 509, 1925.
14. John R. Carson and Ray S. Hoyt, "Propagation of Periodic Currents over a System of Parallel Wires," *Bell Sys. Tech. Jour.*, pp. 495–545, July, 1927.
15. John R. Carson, "Rigorous and Approximate Theories of Electrical Transmission Along Wires," *Bell Sys. Tech. Jour.*, January, 1928.
16. John R. Carson, "Wire Transmission Theory," *Bell Sys. Tech. Jour.*, April, 1928.
17. A. Ermolaev, "Die Untersuchung des Skineffektes—Drahten mit Complexer Magnetischer Permeabilitate," *Archiv. f. Elektrotechnik*, Vol. 23, pp. 101–108, 1929.
18. Louis V. King, "Electromagnetic Shielding at Radio Frequencies," *Phil. Mag.*, Vol. 15, Series 7, pp. 201–223, February, 1933.
19. E. J. Sterba and C. B. Feldman, "Transmission Lines for Short-Wave Radio Systems," *Proc. I. R. E.*, July, 1932, and *Bell Sys. Tech. Jour.*, July, 1932.
20. H. Pleijel, "Beräkning af Motstånd och Självinduktion," Stockholm, K. L. Beckmans Boktryckeri, 1906.
21. H. Pleijel, "Electric and Magnetic Induction Disturbances in Parallel Conducting Systems," 1926, *Ingeniorsvetenskapsakademiens Handlingar* NR 49.
22. J. Fisher, "Die allseitige in zwei Kreiszylindrishen, konaxial geschichteten Stoffen bei axialer Richtung des Wechselstromes," Jahrbuch der drahtlosen Telegraphie und Telephonie, Band 40, 1932, pp. 207–214.

Books

J. Clerk Maxwell, "Electricity and Magnetism," Vols. 1 and 2.
O. Heaviside, "Electrical Papers."
Sir William Thomson, "Mathematical and Physical Papers."
Lord Rayleigh, "Scientific Papers."
Sir J. J. Thomson, "Recent Researches in Electricity and Magnetism."
A. Russell, "A Treatise on the Theory of Alternating Currents."
John R. Carson, "Electric Circuit Theory and the Operational Calculus."
R. W. Pohl, "Physical Principles of Electricity and Magnetism."
Max Abraham and R. Becker, "The Classical Theory of Electricity and Magnetism."
E. Jahnke and F. Emde, "Tables of Functions," B. G. Teubner, 1933.

Note: This list of references is by no means complete. Only the more recent papers dealing with some phase of the subject treated here are included.

Contemporary Advances in Physics, XXVIII
The Nucleus, Third Part *

By KARL K. DARROW

This article deals first with the newer knowledge of alpha-particle emission: that common and striking form of radioactivity, in which massive atom-nuclei disintegrate of themselves, emitting helium nuclei (alpha-particles) and also corpuscles of energy in the form of gamma-rays or high-frequency light. There follows a description of the contemporary picture of the atom-nucleus, in which this appears as a very small region of space containing various charged particles, surrounded by a potential-barrier; and the charged particles within, or those approaching from without, are by the doctrine of quantum mechanics sometimes capable of traversing the barrier even when they do not have sufficient energy to surmount it. The exponential law of radioactivity—to wit, the fact that the choice between disintegration and survival, for any nucleus at any moment, seems to be altogether a matter of pure chance—then appears not as a singularity of nuclei, but as a manifestation of the general principle of quantum mechanics: the principle that the underlying laws of nature are laws of probability. Moreover it is evident that transmutation of nuclei by impinging charged particles, instead of beginning suddenly at a high critical value of the energy of these particles, should increase very gradually and smoothly with increasing energy, and might be observed with energy-values so low as to be incomprehensible otherwise; and this agrees with experience.

Diversity of Energies in Alpha-Particle Emission

ON EVERYONE who studied radioactivity some twenty years ago, there was impressed a certain theorem, an attractively simple statement about the energy of alpha-particles: it was asserted that all of these which are emitted by a single radioactive substance come forth from their parent atoms with a single kinetic energy and a single speed. When beams of these corpuscles were defined by slits and deflected by fields for the purpose of measuring charge-to-mass ratio, nothing clearly contradicting this assertion was observed: the velocity-spectrum of the deflected corpuscles appeared to consist of a single line. In studying the progression of alpha-particles across dense matter, it was indeed observed that not all of those proceeding from a single substance had sensibly the same range. It is, however, to be expected that if two particles should start with equal energy into a sheet of (let us say) air or mica, they would usually traverse unequal distances before being stopped; for the stopping of either would

* This is the second and concluding section of "The Nucleus, Third Part," begun in the July, 1934 *Technical Journal.*

"The Nucleus, First Part" was published in the July, 1933 issue of the *Bell Sys. Tech. Jour.* (12, pp. 288–330), and "The Nucleus, Second Part" in the January, 1934 issue (13, pp. 102–158).

be brought about by its encounters with atoms and the electrons which atoms contain; and there would be statistical variations between the numbers of atoms and electrons which different particles would encounter after plunging into such a sheet. The probable effect of these variations can be computed; and it was shown at an early date that for at least some of the substances emitting alpha-rays—α-emitters—the diversity in ranges of the particles is no broader than should be expected. The curve of distribution-in-range of an α-ray beam often consists of a single peak or hump, and the shape and breadth of the hump are consistent with the assumption that it is due entirely to the "straggling" (the name applied to the statistical variations aforesaid) of particles all possessed initially of a single speed.

The vanishing of this beautiful but too-simple theorem from physics is due to experiments of three types. First, it was found that when all of the well-known α-particles of about 8.6 cm. range from ThC′ were completely intercepted by a stratum of matter of rather more than 8.6-cm. A.E. (air-equivalent [12]), and the detecting apparatus was adjusted to a sensitiveness much greater than would have been tolerable for the main beam, a very few particles—a few millionths of the number in the 8.6-cm. flock—were still coming through. Some of these had ranges as great as 11.5 cm., immensely greater than could be ascribed to straggling. These are the "long-range" alpha-particles, other examples of which have been discovered with RaC′ and (very lately) with AcC′. Next the colossal new magnet at Bellevue near Paris was employed by Rosenblum for deflecting α-particles and observing their velocity-spectrum, and the unprecedented dispersion and resolving-power (to employ optical terms) of this superb apparatus disclosed that for several α-emitters (the list now comprises eight) the spectrum consists of two or several lines instead of only one. The "groups" of alpha-particles to which these lines bear witness lie closer to one another in energy than the aforesaid long-range particles lie to the medium-range ones, wherefore they are often said to constitute the "fine-structure" of the alpha-rays; but it is probable that a more significant basis for distinction lies in the fact that the long-range corpuscles are relatively scanty, while the various lines of a fine-structure system are not so greatly unequal in intensity.

[12] I recall that while the range of an alpha-particle of given speed depends on the density and nature of the substance which it is traversing, the student is usually dispensed from taking account of this by the fact that the investigators nearly always state, not the actual thickness of the actual matter which they used, but the thickness of air at a standard pressure and temperature (usually 760 mm. Hg and 15° C.) which would have the same effect.

Another great magnet of a peculiar and original construction, developed at the Cavendish, was then applied both to spectra displaying fine structure and to the spectrum of RaC′, with notable success; while the technique of determining distribution-in-range curves has been improved to such an extent that it now almost rivals the magnets in its capacity of distinguishing separate groups in an alpha-ray beam. The theorem of the unique speed is therefore like so many another theorem of physics; it was valid so long as the delicacy of the experimental methods was not refined beyond a certain point, its validity ceased as soon as that point was passed.

To enter now into detail:

The *long-range particles* were discovered by observing scintillations, a method of singular delicacy and value, but having great disadvantages: all the observations being ocular, it is wearisome and taxing, not every eye is capable of it, and there is no record left behind except in the observer's memory or notes. Tracks of some of these particles were later photographed in the Wilson chamber, but it is a long research to procure even a few hundred of such photographs, and yet even a few hundred are not sufficiently many for plotting a really good distribution-in-range curve (the disagreements between the early work with scintillations and the subsequent work with Wilson chambers are rather serious). The best available curve is that which Rutherford, Ward and Lewis obtained with the method of the "differential ion-chamber," of which the principle is as follows:

When an alpha-particle (or, for that matter, a proton) traverses a sheet of matter, its ionizing power or ionization per-unit-length-of-path—we may take one mm. as a convenient unit of air-equivalent—varies in a characteristic way with the length of path which the particle has yet to traverse before being stopped completely. The ionization-curve at first is nearly horizontal, then rises to a pretty sharp maximum, then falls rapidly to zero.[13] Suppose now that the particle traverses a pair of shallow ionization-chambers, each containing a gas of which the thickness amounts to not more than a few mm. of air-equivalent (the same for both) and the two separated by a metal wall of negligible air-equivalent. Suppose further that the metal wall is both the negative electrode of the one chamber and the positive electrode of the other, and that it is connected to the electrometer or other detecting device. The charge which is perceived is then the difference between the ionizations in the two chambers. If these are traversed by a particle which is yet far from the end of its range, the difference

[13] The curve for protons is exhibited in Fig. 7 of "The Nucleus, Second Part," p. 124.

will be small and perhaps imperceptible; if by a particle which is approaching its maximum ionizing-power, the difference will be appreciable and of one sign; if by a particle which is coming to the end of its range, the difference will be considerable and of the opposite sign. So the differential chamber and its detecting device (in these experiments, an oscillograph connected through an amplifier, reacting appreciably to the passage of a single particle) are sensitive above all to particles which are nearing the ends of their ranges; and if a small number of such corpuscles be mingled with even a much greater number of faster charged particles—be they alpha-particles, be they protons, be they the fast electrons produced by gamma-rays—this circumstance, which would cripple any other method, will be almost without effect on it.[14] If the readings of the electrometer are plotted against the air-equivalent of the thickness of matter between the source (of alpha-particles) and the chamber, the resulting curve should not be much distorted from the ideal distribution-in-range curve.

The curve obtained in this way for the long-range particles of RaC′ exhibits a notable peak at range 9.0 cm.; to one side thereof a very much lower hump at smaller range (7.8); to the other side a wavy curve with four distinct maxima, which Rutherford and his colleagues deem to be the superposition not of four peaks only, but of seven. I show this portion (Fig. 5) to illustrate the analysis of such a curve for groups. Even the tallest of the peaks just mentioned is a mere molehill compared to the mountain which the principal group of RaC′ —the 6.9-cm. α-particles which were formerly the only ones known—would form if it could be plotted on the same sheet of paper; for the abundances of the 7.8-cm., 9.0-cm. and 6.9-cm. groups stand to one another as 1 : 44 : 2,000,000.

These, however, are not the latest words concerning the α-spectrum of RaC′. The energies of these groups might be deduced from their ranges, but for this purpose it is necessary to use an empirical curve of energy *vs.* range which at the time of the foregoing spectrum-analysis had been extended only up to range 8.6 cm. It was desirable to measure the energies of some of these groups directly, not only for their intrinsic interest but in order to carry onward that empirical energy-*vs.*-range relation. Recourse must therefore be had to a deflection-method.

Now in the usual form of magnet employed in deflection-experiments the field pervades the whole of the space between the solid disc-shaped faces of two pole-pieces. Were the pole-pieces to be so hollowed out

[14] Also a proton near the end of its range can be distinguished from an alpha-particle near the end of *its* range, on account of the difference in maximum ionization-per-unit-length ("The Nucleus, Second Part," pp. 124–125).

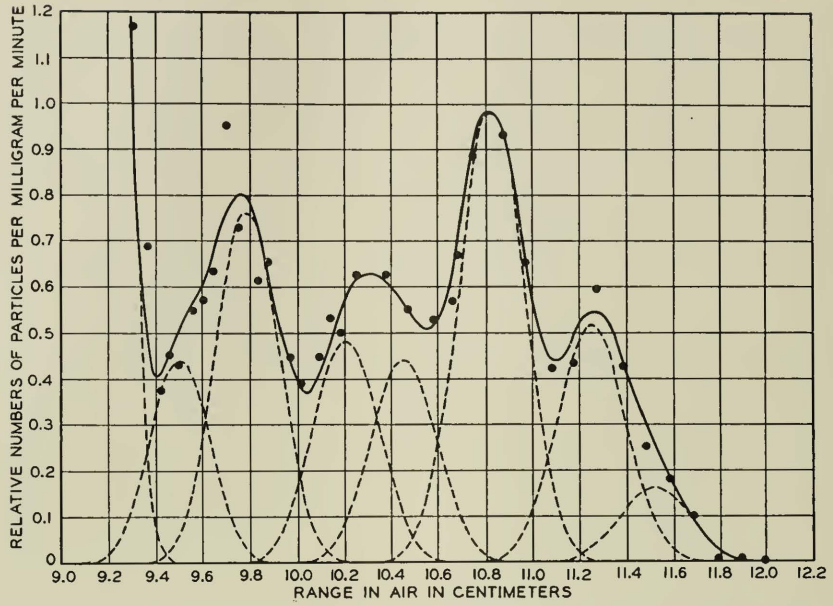


Fig. 5—Distribution-in-range of the long-range alpha-particles proceeding from RaC′, determined with the differential ionization-chamber (Rutherford Ward & Lewis; *Proc. Roy. Soc.*).

that their disc-shaped faces were reduced to narrow circular rings, there would be a great economy in magnetizable metal and a great reduction of weight and volume of the apparatus, as well as other advantages. This need not impair the availability of the magnet for analyzing a beam of α-particles, provided that the α-emitter can be located in the narrow annular space between the faces of the rings, and provided that the magnetization of the metal can be varied sufficiently widely. For then, for each group of α-particles there will be a certain value of the field-strength, whereby those particles which start out in directions nearly perpendicular to the field and tangent to the rings will be swept around in circular paths which are confined within that narrow annular space where alone the field exists. Somewhere in that space the detector should be placed; and the curve of its reading *vs.* field-strength H should show a peak for every group, and from the abscissa of the peak and the radius of the rings the energy of the group may readily be computed.

Such a magnet was built after Cockcroft's design at the Cavendish; the radius of its rings is 40 cm., they are 5 cm. broad and 1 cm. apart (these figures are the dimensions of the annular space), and the field-strength was adjustable up to 10,000 (later 12,000) gauss which was

sufficient for α-particles of energy up to and even beyond 10.6 MEV (millions of electron-volts) and range up to and even beyond 11.5 cm. Figure 6 exhibits the outward aspect, Fig. 7 the cross-section of this device (one sees how the armature is fully contained within the rings). The annular space and everything within it was evacuated (being walled in by the ring B seen in the figures); the detector—a simple ionization-chamber connected through a linear amplifier to an oscillograph—was set 180° around the annulus from the source. This device in the hands of Rutherford, Lewis and Bowden proved itself able to furnish even a better spectrum than the scheme of the differential ionization-chamber; all of the peaks indicated by the former curve were clearly separated, a hump which had suggested two groups was resolved into three maxima, and an extra group was discovered—*twelve* altogether! (Incidentally, the empirical energy-*vs.*-range curve of α-particles had previously been extended with the same device by Wynn-Williams and the rest, to energy = 10.6 MEV and range = 11.6 cm.)

The long-range spectrum of RaC′ is thus of no mean complexity. There will be occasion later for quoting its actual energy-values. Of the long-range spectrum of ThC′ there is relatively little to be said; evidently it has not been studied so intensively as the other, but it seems to be comparatively simple, for only two groups have been recognized. One of the groups of ThC′ has about the same range as the highest group of RaC′, so that between them they comprise the most energetic subatomic particles ever yet discovered (about 10.6 MEV) apart from those of the cosmic rays and those resulting from certain artificial transmutations. As for AcC′, it is one of the constituents of the mixture of radioactive bodies known as actinium active deposit, from which α-particles of a range of about 10 cm. have lately been observed in the Institut du Radium.

Thus far I have written as though RaC′ and ThC′ were isolable substances, of which one may obtain pure samples and analyze at leisure the α-rays thereof. The truth, however, is far otherwise; for the difficulties of making one radioactive substance practically free from others, serious in most cases, are utterly insuperable in these. Both RaC′ and ThC′ are so very ephemeral (their half-lives are too small to measure, and are guessed from the Geiger-Nuttall relation as 10^{-6} and 10^{-11} second respectively) that they can never be dissevered from their mother-elements RaC and ThC which are also α-emitters. Sometimes one finds the long-range particles designated as belonging to RaC or ThC, and indeed I have nowhere found stated any compelling reason for attributing them to the C′-elements rather than the C-

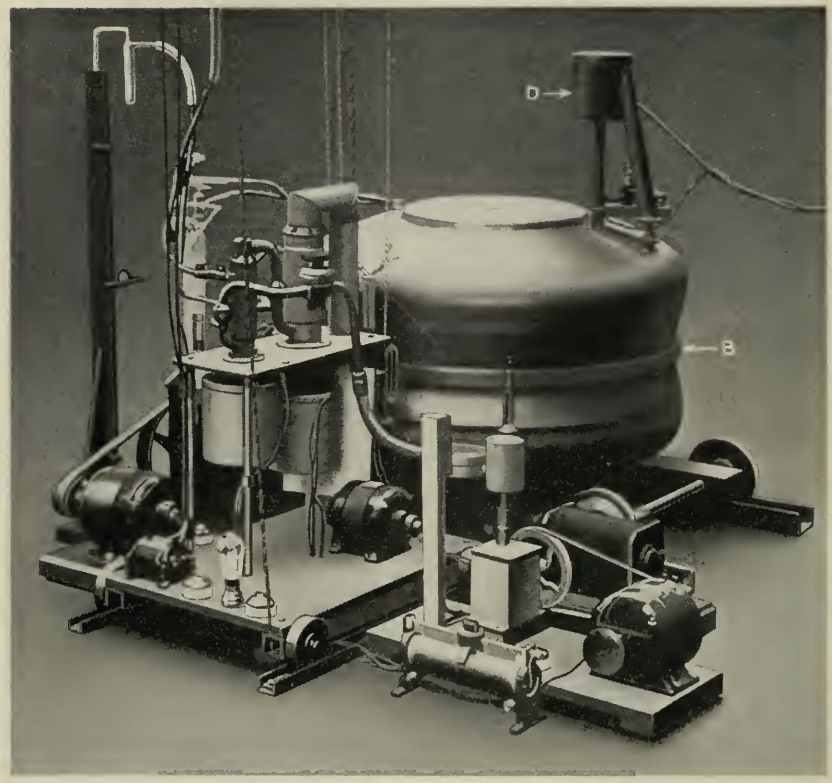


Fig. 6—Annular magnet employed for analysis of alpha-ray spectra. (After Rutherford, Wynn-Williams, Lewis & Bowden; *Proc. Roy. Soc.*).

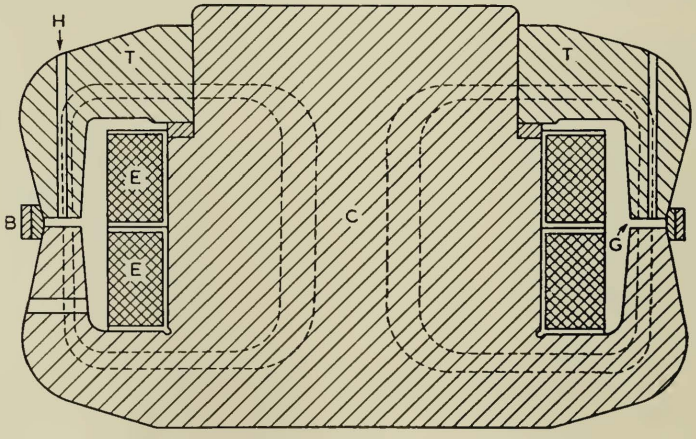


Fig. 7—Cross-section of the annular magnet used for analysis of alpha-ray spectra.

elements, apart from what is known about the correlated gamma-rays (see footnote 21).

Fine-structure was discovered, as I said before, with the great magnet of Bellevue. This has solid pole-pieces (75 cm. in diameter!) instead of rings; it is not necessary to adjust the field-strength step by step so as to bring group after group to a narrow detector; all the groups

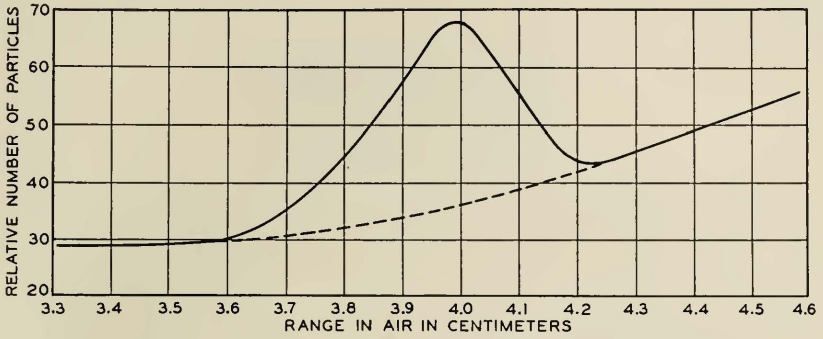


Fig. 8—Alpha-ray spectrum of RaC; peak observed with differential ionization-chamber, never before detected because of immensely greater number of particles in RaC′ peak just off the diagram to the right; asymmetry indicating fine-structure. (Rutherford, Ward & Wynn-Williams; *Proc. Roy. Soc.*).

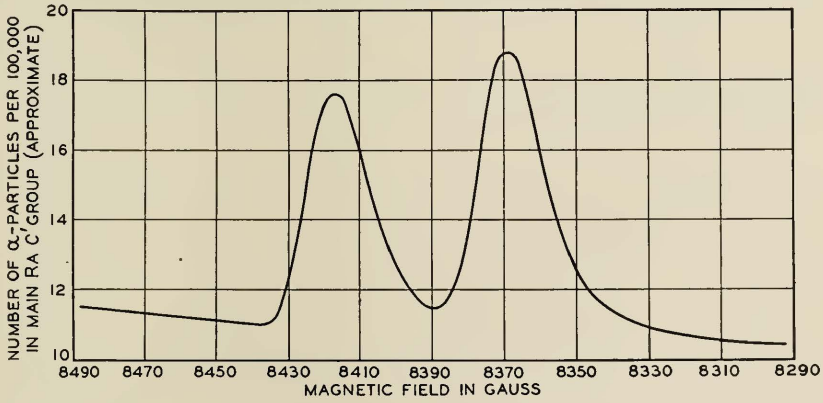


Fig. 9—Fine-structure of alpha-ray spectrum of RaC; as asymmetric peak of Fig. 8 resolved into two nearly equal peaks by annular magnet. (Rutherford, Wynn-Williams, Lewis & Bowden; *Proc. Roy. Soc.*).

of various speeds are deviated simultaneously in circular arcs each of its own particular radius, and simultaneously fall upon a photographic plate, producing what looks precisely like a line-spectrum in optics (Figs. 10, 11). The example in Fig. 10 relates to ThC, the earliest to

be analyzed; four lines only are visible upon the reproduction, but some plates after long exposure have shown as many as six.* The fine-structure of AcC consists of a pair of lines, which were detected as peaks in the distribution-in-range curve obtained at the Cavendish with a differential ionization-chamber. Instead of showing this curve I have chosen the corresponding curve for RaC, albeit it shows only a single hump (Fig. 8).[15] The unsymmetrical shape of this hump, how-



Fig. 10—Fine-structure of alpha-ray spectrum of ThC (not completely brought out in picture) obtained with Bellevue magnet. (S. Rosenblum.)

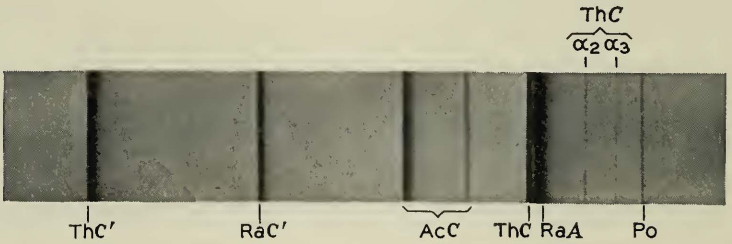


Fig. 11—Alpha-ray spectra of several elements (those of Po and AcC shifted slightly to the right with respect to the rest). (S. Rosenblum, *Origine des rayons gamma*, Hermann & Cie.).

ever, implies that it really consists of a pair of overlapping peaks; and so it does; for when the Cavendish magnet was applied to an α-ray beam from this element, the curve of detector-reading *vs.* field-strength displayed two equal peaks quite sharply separate (Fig. 9).

According to Rosenblum's latest census (February 2, 1934) there are now eight known examples of fine-structure: from the radium series, Ra (two groups), RaC (2); from the thorium series, RdTh (2), ThC (6); from the actinium series, RdAc (no fewer than eleven groups, the richest case of all!), AcX (3), An (3), AcC (2). According to Lewis and Wynn-Williams, there are (or were, in the spring of 1932)

* I am indebted to Dr. Rosenblum for a print from which Fig. 10 was made.

[15] This curve was the first to disclose the α-rays of RaC, previously known only by inference (though it was very compelling inference). Being of somewhat lesser range than the much more numerous (to be precise, 3000 times as numerous) α-particles emanating from the RaC′ atoms with which RaC is always inevitably mingled, they were completely hidden from observation by any method known before the use of the differential chamber and the powerful magnet.

at least five cases in which the distribution-in-range curve obtained with the differential chamber shows a single symmetrical peak suggesting only one group: from the radium series, Rn and RaA; from the thorium series, Tn and ThA; from the actinium series, AcA. Altogether there are twenty-three [16] known alpha-emitters, so that nearly half of the total remain to be investigated to this end. It may be significant that out of the four known alpha-emitters having odd atomic number, the high proportion of three at least is known to display fine-structure (the fourth, Pa, being as yet uninvestigated).

Interrelations of Alpha-Ray Spectra and Gamma-Ray Spectra

Evidently, if two atoms of the same radioactive substance were to emit alpha-particles of different speeds, there would be three obvious possibilities. The resultant nuclei might be different: in this case we should expect (though not with certainty) that they would be the starting-points of different radioactive series, and we should speak of "branching." The initial nuclei might have been different, in which case it would have been improper to speak of them as belonging to the same substance. Finally one at least of the two atoms might also emit gamma-ray photons, of energies complementary to those of the alpha-particles, in such a way that the total amount of energy released by the one atom would be the same as the total amount released by the other.

The first of these possibilities is not to be excluded *a priori* (since there are known cases of branching, though in them the alternative is between emission of an alpha-particle and emission of an electron) and neither is the second. The third, however, is the most agreeable, since if realized it allows us to believe that in the transformation of radium (to take one example) every radium nucleus is like every other before its change begins and every resulting (radon) nucleus is like every other after its change is over. Now alpha-ray emission and gamma-ray emission often occur together, which suggests that often the third possibility is the one which is realized; but this cannot be proved without measuring the energies or the wave-lengths of the gamma-rays.

The simplest cases are those in which the alpha-ray spectrum consists of two lines only. Here and always, there is an inconvenient complication at the start: when an alpha-particle is emitted, the residual nucleus recoils, and it is the sum of the kinetic energies of

[16] Not including Sm and other elements of atomic number lower than 81. The rest are depicted (together with the beta-ray emitters of atomic numbers 81 and greater) in Fig. 21.

the two (not that of the alpha-particle alone!) which must be taken into account.[17] Denote by U_1 and U_2 the values of this sum for the faster and for the slower alpha-particles. Does the gamma-ray spectrum then consist of a single line of which the photon-energy $h\nu$ is equal to $(U_1 - U_2)$?

In the case of AcC, the difference $(U_1 - U_2)$ is 0.35 or 0.36 MEV. There is an intense gamma-ray line proceeding from actinium active deposit (comprising AcC), and the energy of its photons is concordant. In the case of RaC, the difference $(U_1 - U_2)$ is only 0.04 MEV, and the search for so relatively soft a radiation of photons is difficult. In the case of Ra the conditions are more favorable, and here the history is worth retelling. Long before the earliest analysis of alpha-ray spectra, radium was known to emit feeble gamma-rays of photon-energy about 0.19 MEV. An estimate of their intensity was made in 1932 by Stahel; he concluded that the photons are less than one-tenth as numerous as the alpha-particles already known. Search was thereupon made by Rosenblum for fine-structure in the alpha-ray spectrum of radium. Two lines appeared on the plate after five minutes' exposure: they were due to groups proceeding one from radium and the other from its daughter-element radon. On plates exposed for hours there appeared yet another line. The values of U_1 and U_2 being computed for this and for the stronger radium group, the difference was found to be close to 0.185 MEV, with a sufficient latitude to be concordant with the estimate for the photons.

As the number of alpha-ray lines increases beyond two, the prospects rapidly become formidable; for a spectrum of n such lines suggests n possible states of the residual nucleus, and every one of these might "combine" (in the technical sense of the word) with every one below it in the energy-scale, making a total of $n(n - 1)/2$ gamma-ray lines to be expected. Even so, anyone acquainted only with optical spectra might think it no difficult matter to photograph (say) the gamma-ray spectrum of ThC, and see whether it consists in just 15 lines in just the right places to correspond with the six alpha-particle groups. But one does not photograph gamma-ray spectra—one photographs the beta-ray spectra of the electrons ejected by the gamma-rays from atoms, and tries to deduce the photon-energies $h\nu$ from the electron-energies.[18] The atoms may be those of the radioactive substance

[17] By multiplying the kinetic energy of the alpha-particle by the factor $(1 + m/M)$, where m stands for the mass of the alpha-particle and M for that of the recoiling nucleus. This point was overlooked by a number of people before it was noticed by Feather.

[18] I have dealt with this procedure at length in the article "Radioactivity," No. XII of this series (*Bell Sys. Tech. Jour.*, 6, 55–99, 1927).

itself (either the very ones which are emitting the gamma-rays, in which case the rays are said to undergo "internal conversion," or their neighbours) or they may be those of other elements mixed with the radioactive substances, or those of nearby solids or gases on which the gamma-rays fall. Each gamma-ray line is responsible for several different beta-ray lines, a circumstance which makes the analysis more difficult at the beginning though it makes the inference more reliable in the end. There may be gamma-rays having nothing to do with alpha-particle emission, and there may be gamma-rays from several different radioactive substances inextricably mixed up together, so that the problem of analyzing the spectrum of one transformation is preceded (or, more truly, accompanied) by that of distinguishing it from the intermingled spectra of others.[19] The experimental errors in the estimates of U-values and $h\nu$-values may be so large that apparent agreements are actually unreliable. Altogether, the comparison of a rich alpha-ray spectrum with a rich gamma-ray spectrum is an exceedingly intricate business, the outcome of which is not to be summarized in a few sentences. To give a mere notion of the sort of conclusion which is reached, I quote some lines from Rutherford, Lewis and Bowden, in their comparison of the thirteen-line alpha-ray spectrum and the very rich gamma-ray spectrum of RaC′:

"When we consider in broad terms the data which have been presented, there can be no doubt that there is a high correlation between the alpha-particle levels which have been observed and the emission of gamma-rays. In more important cases the numerical agreement is well within the experimental error of measurement, while the relation between the intensity of the alpha-ray groups and the gamma-rays associated with them is of the right order of magnitude to be expected on general theoretical grounds. In other cases the agreement is very uncertain, and more definite information on the gamma-rays is required to make the deductions trustworthy. It is unfortunate that we have been unable to detect the alpha-particle groups corresponding to certain postulated levels [*i.e.* postulated from the classification of the gamma-ray lines] . . . ."

Thus it appears that there are excellent agreements between $h\nu$-values and $(U_i - U_j)$ values, and yet nothing approaching a perfect one-to-one-correspondence. Nevertheless, the general programme is fixed: to assume that each nucleus possesses a system of stationary

[19] It is interesting to notice that after the $h\nu$-values of certain gamma-rays emitted from mixtures of ThC and ThC″ had been found to agree with values of $(U_i - U_j)$ taken from the alpha-ray spectrum of ThC, these gamma-rays were proved to proceed from ThC in its transformation into ThC″, whereas till then they had been supposed to proceed from ThC″ in its transformation into ThD (Meitner & Philipp, Ellis).

states and energy-levels, to assume further that $h\nu$-values and $(U_i - U_j)$-values are alike the differences between these energy-levels, and to ascribe apparent defects of correlation to special circumstances by virtue of which certain gamma-rays and certain alpha-rays are too feeble to be detected. Should these ideas prove untenable, we shall probably have to suppose that the nucleus is even more different from the extra-nuclear world than we have hitherto admitted.

Now arises the important question: when alpha-particles and photons both are emitted in the course of the complete transformation of one nucleus into another, which comes first? Despite the immeasurable shortness of the times which are involved, this is in principle an answerable question. For as I have mentioned already, gamma-rays are detected and their photon-energies are measured by examining the spectrum of the electrons which they eject from the orbital electron-layers of atoms, chiefly from the layers surrounding those very nuclei whence the photons themselves proceed. Now if the photons come before the alpha-particles, say for example in the transformation of ThC into ThC″, these electrons will come from the electron-layers of ThC atoms; in the contrary case, from the layers of ThC″. It is possible to distinguish from which they do come, even when the energy of the photons is not independently known and must itself be derived from the same data.[20]

The classical and crucial experiment of this type was performed about ten years ago by Meitner, and it proved that the gamma-rays emitted during the transformation of RdAc into AcX and during that of AcX into An spring forth *after* the alpha-particle has departed and the nucleus has become that of the daughter-element. These are two of the cases in which the alpha-ray spectrum exhibits fine-structure; and it is now generally supposed that the rule extends to all such cases. The stationary states or energy-levels deduced from the $(U_i - U_j)$-values and the $h\nu$-values then would pertain to the "final" or daughter nucleus. In the instances where all the alpha-particle groups except the main one are designated "long-range groups"—RaC′ and ThC′ (the quotation above from Rutherford, Lewis and Bowden refers to the former of these)—Gamow argues that the gamma-rays are emitted before the alpha-particles; the energy-levels deduced from the alpha-ray and the gamma-ray spectra would then pertain to

[20] See the previously-cited article "Radioactivity," pp. 94–96. I mention in passing that sometimes the "internal conversion" of photons whereby electrons are ejected is apparently so much the rule, that no appreciable fraction of the gamma-rays of some particular energy (or energies) escape from the atoms at all; in which cases it becomes expedient to speak not of gamma-rays at all, but of an immediate transfer of energy from the nucleus to the orbital electrons (a policy which may be applied to all cases of internal conversion).

the "initial" or mother nucleus. It is not clear from the literature whether this hypothesis has been fully tested in the manner of Meitner's tests aforesaid, but presumably it was adopted in calculating the $h\nu$-values from the electron-energies, so that the agreements between $h\nu$ and $(U_i - U_j)$ support it.[21]

The search for interrelations among the energy-levels, the different $h\nu$-values and the different U-values belonging to individual transformations has of course already begun. Rutherford and Ellis find that the frequencies of many of the lines in the gamma-ray spectrum of RaC′ can be fitted by assigning various integer values to p and q and constant values to E_1 and E_2 in the formula $pE_1 + qE_2$; while H. A. Wilson finds that if the U-values or the $h\nu$-values are added together in pairs, an amazing number of the pairs are equal to integer multiples (the integer multipliers ranging from 16 to 54) of the amount 0.385 MEV—this even if the two members of a pair are taken from different spectra!

The Quantum-Mechanical Theory and the Crater Model of the Nucleus [22]

Anyone who is acquainted with the contemporary atom-model in its present or in its earlier stages, with its congeries of charged particles revolving in or jumping between definitely-prescribed and quantized orbits, governed by attractions and repulsions both classical and unimaginable—any such person will probably be looking for a nucleus-model of the same variety but built on a very much smaller scale,

[21] I learn by letter from Dr. Ellis that in the case of RaC′, some at least of the gamma-rays which agree with the $(U_i$-$U_j)$-values of the long-range alpha-particles are definitely proved in this fashion to proceed from nuclei of atomic number 84 (that of RaC′) as distinguished from 83, 82 or 81; the proof is especially strong for the most intense gamma-ray, of photon-energy 0.607 MEV. Perhaps this is the most powerful evidence that the long-range particles come from RaC′ rather than RaC.

The half-periods of RaC′ and ThC′ are exceedingly short, 10^{-6} sec. and 10^{-11} sec. respectively; had it been otherwise, objection might be made to Gamow's contention on the ground that atoms in states, from which they are liable to depart by emitting radiation, generally do depart from those states and emit that radiation within a period of the order of 10^{-7} or 10^{-8} sec. There is no *a priori* certainty that this principle applies to nuclei, but if it does it may suffice to explain why the long-range particles are observed only from these very short-lived nuclei, why they are so scanty even in these cases (nearly always the photon *is* emitted before the alpha-particle gets ready to leave, so that the latter nearly always leaves with low energy instead of high), and why the fine-structure of other alpha-ray spectra is related to energy-levels of the final instead of the initial nucleus (Gamow). Incidentally it strengthens the case for ascribing the long-range particles to the C′-products instead of the C-products.

[22] Quantum mechanics was first applied to the nucleus-model here to be described, independently and almost simultaneously, by Gurney and Condon and by Gamow. Rather than by all three names together, it seems preferable to denote this model thus interpreted by a neutral descriptive term, such as "crater model" (an allusion to the aspect of the graph obtained when the curve of Fig. 12 is coupled with its mirror-image in the yz-plane).

with protons and possibly neutrons figuring among the revolving particles. He will be looking too far into the future, and will be disappointed with the present. The present nucleus-model consists of little more than a single curve—a curve which, moreover, relates only to the fringe of the nucleus and to the region surrounding it, and for want of knowledge is not extended into the central region or nucleus proper where the constituent particles must be. The theory which it serves is a theory not of the nucleus as a stable system of corpuscles, but of the escape of some from among these corpuscles and the entry of new ones—a theory professing to deal only with the entry and the escape, not at all with the events succeeding the one or preceding the other.

The curve purports to portray the electrostatic potential, as function of r the distance measured from the centre of the nucleus, from $r = \infty$ inward to a minimum distance which is indeed very small even in the atomic scale—10^{-13} cm. or less—but still definitely not zero, since the components of the nucleus must be presumed to be normally at distances yet smaller. When it is plotted as in Fig. 12, its ordinate

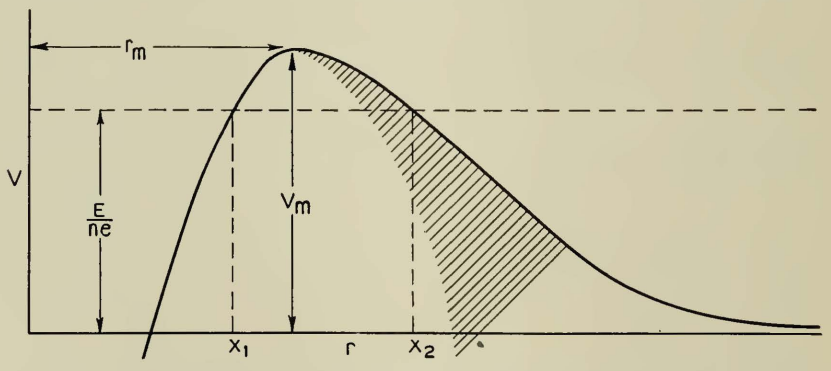


Fig. 12—Nuclear potential-curve postulated for explaining transmutation (without allowance for resonance) and radioactivity.

at any r is a measure of the amount of kinetic energy which a positively-charged particle approaching the nucleus must sacrifice—i.e. which must be converted into potential energy—in order to come from infinity to r. Traced from infinity inward, the curve must follow at first the function *const.*/r, corresponding to the inverse-square law of force; for it is known, both from experiments on alpha-particle scattering (which supplied the foundation for the contemporary atom-model) and from the successes of the theory of atomic spectra, that beyond a certain distance a nucleus is surrounded by an inverse-

square force-field. This is the obstacle, or at any rate, a part of the obstacle, which an oncoming proton or deuton or alpha-particle must overcome in order to reach the nucleus and achieve transmutation. One may picture it as a hill, up which the ball must roll to reach the castle at the top—and *down* which the ball will roll if it starts from the top, shooting outward towards infinity as the fast-flying alpha-particle.

Now to assume the inverse-square force as prevailing all the way inward to $r = 0$ would be to postulate a point-nucleus without room for parts or structure, surrounded by a hill of infinite height which no approaching positive particle could climb; all of which is inadmissible. Departure from the inverse-square law is actually shown by some experiments on scattering of alpha-particles which pass very close to the nucleus, and these indications are to be heeded in tracing that part of the curve of Fig. 12 which lies to the right of the maximum; but the maximum itself and the sharp descent to its left are dictated by no such observations, and to postulate them is to make the theory which is now to find its employment and its test. That there should be such a maximum and such a descent is of course the most natural supposition to make. If there are several particles of positive charge which stay for a finite time within the nucleus, there must be something which restrains them from flying away. This something must either be an agency of a type as yet unknown, or else be described by a potential-curve with a maximum at what we may henceforward call the boundary of the nucleus; and the latter assumption is to be preferred till proved unusable.

Applying classical ideas to this "model" (if the word be not considered too presumptuous) of a nucleus, one is led at once to two predictions, which may be sharply formulated if we adopt symbols such as V_m and r_m for the two parameters indicated on Fig. 12, *viz.* the "height of the potential-barrier" and the "radius of the potential-barrier" as they are commonly called, the latter being also called the "radius of the nucleus." These are:

1. If the nucleus emits a particle of positive charge $+ 2e$, the kinetic energy with which this particle is endowed when it completes its escape cannot be less than $2eV_m$; consequently, when it is observed that atoms of a certain element emit alpha-particles with kinetic energy K_0, the height of the potential-barrier for that element cannot surpass $K_0/2e$; consequently, when the force-field about the nuclei of such atoms is explored by the classical method of studying the scattering of alpha-particles projected against a sheet of that element, it must be found that the region of repulsive force, and *a fortiori* the

inverse-square field, do not extend far enough inward for the integral to surpass the value $K_0/2e$.

2. When particles of charge ne (n = any integer) are projected at a sheet of any specific element, they cannot enter the nuclei at all unless their kinetic energy exceeds the critical value neV_m; and if the curve of number-entering-nuclei *vs.* kinetic-energy can in any way be deduced from any experiments, it should rise fairly sharply from the axis of kinetic energy at this critical abscissa.

(It will have been noticed that I expressed both of these predictions as though the escape or the entry of a particle made no difference to the height of the potential-barrier, which is the universal practice. This is obviously too crude an assumption; the error in it must be graver the smaller the atomic number of the element, therefore graver in theorizing about the transmutation of light elements than in theorizing about the radioactivity of heavy ones; it must be rectified in future.)

The former of these predictions can be sharply and unquestionably tested; and it proves to be wrong. Uranium I. emits alpha-particles of kinetic energy K_0 equal to 4 MEV; but Rutherford suspected from scattering-experiments on other heavy elements, and subsequently proved by such experiments upon uranium itself, that the inverse-square force-field extends so far inwards as to involve a height of potential barrier at least twice as great as $K_0/2e$; so that an emerging alpha-particle should possess at least 8 MEV of kinetic energy derived from coasting down the hill, and even this is merely a lower limit to the estimate, since the hill may be higher and the particle might come with some excess of energy over its brow!

The second prediction is not so readily tested. If all of the charged particles (protons or deutons or alpha-particles, say) projected at the postulated sheet of matter were directed straight towards the centres of nuclei, and arrived at the potential-hills without suffering any prior loss of energy elsewhere, the fraction entering through the potential barriers would rise suddenly from zero to unity as the kinetic energy K of the particles was raised to neV_m, and any phenomenon depending solely upon entry would make its advent suddenly if at all. Unfortunately this does not occur in any experiment now possible or likely ever to become possible. If the sheet of matter is a monatomic layer, most of the oncoming particles will be going towards the gaps between the nuclei, and the initial directions of the rest will be pointed towards all parts of the cross-sections of the nuclei, only an infinitesimal fraction going straight toward the centres. Designate by p the perpendicular distance from a centre to the line-of-initial-motion of

an oncoming particle; it is evident that the minimum kinetic energy permitting of entry will increase with p, starting from neV_m and rising to infinity as p rises from 0 to r_m. The relation between fraction-of-particles-entering-nuclei—call it P_e—and kinetic energy K could be calculated, given specific assumptions about the values of V_m and r_m and the trend of the potential-curve. Without undertaking the calculation, it is easy to see that the vertical rise of what I will hereafter call the "ideal" curve—the curve of probability-of-entry-at-central-impact *vs.* K—will be distorted into a bending slope, starting, however, at the same critical abscissa neV_m. If the sheet of matter is a thick layer, there will of course be a much greater fraction of the impinging particles of which the initial paths point straight toward some nucleus or other, but the fraction achieving entry will not be raised in the same ratio, for the particles going toward nuclei embedded deep in the layer will lose some or the whole of their velocity in passing through the intervening matter.[23] This also will contribute to converting the vertical rise into a gradual bend. Still it does not seem possible that if the ideal curve had such a shape, the experimental ones could rise with so extreme a gradualness as does the one of Fig. 17 or those of Figs. 16 and 17 in the Second Part; for these suggest no sudden beginning at all, but rather they have the characteristic aspect of curves asymptotic to the axis of abscissas, as if their apparent starting-points could be pushed indefinitely closer to the origin by pushing up indefinitely the sensitiveness of the apparatus. Neither does it seem possible that V_m can be so low as their starting-points imply.

There is, however, another difficulty: these curves refer not directly to P_e, but to number of transmutations, or to be precise (for precision is essential in these matters) to the number of particles producing transmutations involving the ejection of fragments having certain ranges. Call this number P_t. It is easiest to conduct the argument as though P_t were proportional to P_e—as if an observable transmutation could result only from the entry of a particle through the potential-barrier of a nucleus, and as if the number of transmutations of any special type were strictly proportional to the number of entries, the factor of proportionality being independent of K. Yet few assumptions are less plausible. It is far more reasonable to suppose that the probability of a particle bringing about a transmutation when it enters a nucleus is not invariably unity, but is instead some function $f_1(K)$. It is reasonable also to suppose that a particle passing close to the potential-barrier but not traversing it may yet be able to touch off an internal explosion or eruption leading to a transmutation. Denote

[23] Contrast the two curves of Fig. 17.

by $f_2(K)(1 - P_e(K))$ the number of cases in which this happens. The least which we can take for granted is some general relation of the form,

$$P_t = f_1(K) \cdot P_e(K) + f_2(K)(1 - P_e(K)), \tag{20}$$

and the variations of f_1 and f_2 may contribute still further to blotting out all signs of the hypothetical vertical rise in the ideal curve. Moreover, f_2 might be appreciable at values of K smaller than neV_m, thus blotting out every sign of the critical energy-value at which entry commences.

Thus with regard to the second prediction, the situation is this: the experimental curves of number-of-observed-transmutations *vs.* kinetic-energy-of-impinging-particles rise so smoothly and so gradually from the axis as to give not the slightest support to the idea that entry into the nucleus commences suddenly at a critical value of K; moreover, transmutation commences to be appreciable—for several elements, at least—when K is still so small that K/ne is only a small fraction of the least value which can reasonably [24] be ascribed to V_m, in view of what we know from alpha-particle-scattering about the circumnuclear fields of these or similar elements. This again might be due to the hypothetical effect to which the term $f_2(K)$ in the equation alludes, but it seems far too prominent for that! With it is to be linked the fact that alpha-particles emerge from nuclei with kinetic energy less than $2eV_m$. The potential-hill seems not to be so high either for entering or for emerging particles, as it is for those which only skirt its slopes!

Now if in theorizing about potential-hills and particles we substitute quantum mechanics for classical mechanics, these phenomena cease to be things contrary to expectation, and become instead the very things to be expected.

This is one of the situations—regrettably frequent in the present-day theoretical physics—where neither pictures nor words are adequate. The nearest description which can be made with words is probably somewhat as follows: We set out to ascertain whether a particle of charge ne and kinetic energy K, coming from infinity straight toward the nucleus (I simplify the problem as much as possible) will surmount the potential-hill of height V_m. Were we to conceive it as a particle conforming to classical mechanics, we should arrive at the answers: yes, if $K \geqq neV_m$—no, if $K < neV_m$. But we are to turn away from

[24] It is true that the elements of which the circumnuclear fields have been most carefully explored by alpha-particles are not in general the same as those for which transmutability has been observed down to very low values of K; but boron and carbon figure on both the lists, Riezler having studied the scattering of alpha-particles by these (*Proc. Roy. Soc.*, **134**, 154–170, 1932).

the particle for awhile, and to conceive a train of waves advancing from infinity towards the nucleus. The phase-speed and the frequency of this wave-train are prescribed by definite rules making them dependent upon E, and the train is governed by a prescribed wave-equation in which figures the function $V(r)$ of Fig. 12. On solving this equation in the prescribed fashion we find that it requires the wave-train to continue (though reduced in amplitude) past the top of the hill if K is greater than neV_m. This is partially satisfactory, for the particle when it is reintroduced is to be associated with the waves, and everything would be spoiled if the particle could go where the waves cannot. But also, the equation requires the wave-train to continue past the top of the hill when K is less than neV_m. True, it does not wholly pass; there is a reflected as well as a transmitted beam, and the ratio of reflected to incident amplitude goes very rapidly up towards unity and the ratio of transmitted to incident amplitude goes very rapidly down toward zero as K drops downward from the value neV_m. All the same there *is* this wave-train beyond the hill with an amplitude greater than zero; and the association of particles with waves is apparently spoiled, for the waves can go where the particle cannot.

At this point, however, it is the rule of theoretical physicists to give the precedence to the waves, and declare that *where the waves go there the particle must go also, whether it can* (by classical mechanics) *or cannot.* Since some of the waves are beyond the hill, the particle also must be able to traverse the hill, even though its kinetic energy is insufficient for it to climb to the top. But since the waves beyond the hill have a smaller amplitude than those coming up from infinity, it is not certain that the particle will pass through, but merely possible. The chance or probability of its passing through is determined chiefly (not fully) by the ratio of the squared-amplitudes of the waves on the two sides of the hill, and this is what must be computed by quantum-mechanics. How the particle gets over or through the hill—where and what it is and how it is moving while it is getting through—these are questions which the theorist usually declares to be unanswerable in principle, and having so declared, he does not attempt to visualize this part of the process.

Into Fig. 12 the diagonal lines have been introduced in a crude attempt to make graphic as much as possible of the theory. The length of the sloping line drawn from any point P of the curve is meant as a sort of inverse suggestion of the chance which a particle of charge $+ ne$ has of entering the nucleus if its energy E is equal to ne times the ordinate of P: the longer the line, the less the chance of

entry! (This energy E will be the same as the initial kinetic energy K already so often mentioned, which the particle has before it starts to climb the hill.) Perhaps it is not too fanciful to think of these lines as the posts of a fence standing up vertically from the curve, the varying height of which is a rough indication of the varying difficulties which particles of various energies have in getting through.

To predict successfully how the height of this metaphorical fence, the probability of transmission or of penetration, varies with V or E would be a magnificent triumph of nuclear theory, but it is vain to hope for such a success in the immediate future. Almost certainly the top of the fence curves much more rapidly upward than the drawing suggests, and also there is good reason to think that there may be gaps in the fence somewhere like those depicted in Fig. 13, where the

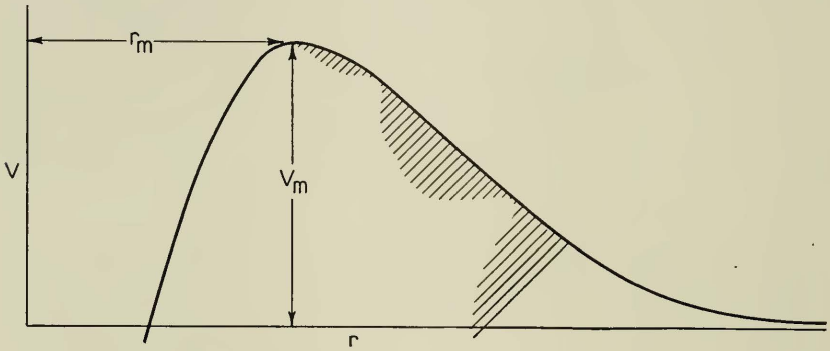


Fig. 13—Nuclear potential-curve postulated for explaining transmutation, with allowance for "resonance."

probability of penetration rises to values remarkably near to unity. Such at least are the features of certain one-dimensional potential-fields (do not forget that Figs. 12 and 13 refer to three-dimensional potential-fields having spherical symmetry!) which have isolated potential-hills or hills-adjoined-by-valleys.

Three of these cases are displayed in Figs. 14, 15, 16. Take the first for definiteness. One plunges *in medias res* by writing down at once Schroedinger's wave-equation:

$$d^2\Psi/dx^2 + (8\pi^2 m/h^2)[E - neV(x)]\Psi = 0, \qquad (21)$$

in which $V(x)$ stands for the potential-function exhibited in the figure,[25] while the meanings of ne, m and E have probably already been

[25] I deviate from the otherwise-universal usage of employing V for the potential energy of the particle, for the reason that the latter depends on the charge of the particle, while the potential-function is supposed (no doubt inaccurately) not to depend on it.

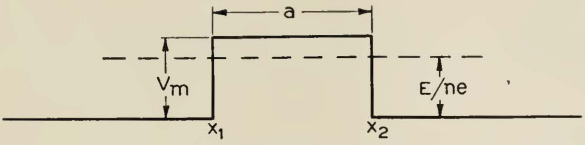


Fig. 14—Illustrating an artificial case of a potential-curve with a single square-topped potential-hill.

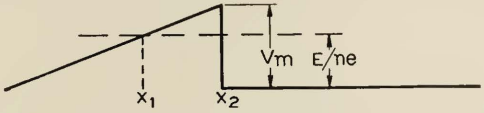


Fig. 15—Illustrating an artificial case of a potential-curve with a pointed hill.

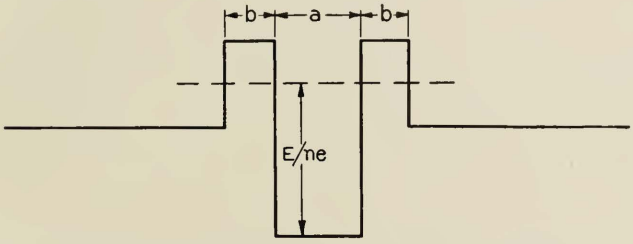


Fig. 16—Illustrating an artificial case of a potential-curve with a valley between two hills.

guessed by the reader—they are constants to which any values may be assigned, and the eventual result of the mathematical operations is going to be taken as referring to a stream of particles of charge ne, mass m and energy E.

The problem is stated as that of finding a solution of (21) for whatever value is chosen for E—a solution everywhere single-valued, bounded, continuous, and possessed of a continuous first derivative, such being the general requirement in quantum mechanics. Not, however, *any* solution possessing these qualities, but a solution apt to the physical situation. On the right of the hill, it must specify a wave-train (I) going from right to left; for we are interested in the adventures of particles coming from the right toward the hill. But on the right of the hill, it must also be capable of specifying a wave-train (II) going from left to right, for some or all of the particles may be reflected from the hillside. On the left of the hill it must be capable of specifying a wave-train (III) going from right to left, for some or all of the particles may traverse the hill and continue on their way.

So far as the region to the right of the hill ($x > x_2$) is concerned, a solution having all of these qualities is the following:

$$\Psi = A_1 e^{-ikx\sqrt{E}} + A_2 e^{+ikx\sqrt{E}}, \qquad k = \sqrt{\frac{8\pi^2 m}{h^2}}, \tag{22}$$

in which the two terms stand for wave-trains I and II, and A_1 and A_2 are adjustable constants. So far as the region to the left of the hill ($x < x_1$) is concerned, a solution having all of the required qualities is the following:

$$\Psi = C_1 e^{-ikx\sqrt{E}}. \tag{23}$$

It stands for wave-train III and C_1 is an adjustable constant. As I have already said, our "intuition" based on notions of what particles should do, expects C_1 to vanish and A_2 to become equal to A_1 when E is less than neV_m, but the solution of (21) does not consent to these limitations. Our intuition also expects that when E is less than neV_m nothing will happen in the region comprised within the hill ($x_1 < x < x_2$), but here again the solution of (21) does not conform with it. For in this region comprised within the hill, the solution must take the form:

$$\Psi = B_1 e^{-kx\sqrt{neV_m - E}} + B_2 e^{+kx\sqrt{neV_m - E}}, \tag{24}$$

which looks at first glance like (22) but is essentially different, since the exponents are real and not imaginary, and the terms do not represent progressive waves. The five coefficients—A_1, A_2, B_1, B_2, C_1—must now be mutually adjusted so that at the sides of the hill ($x = x_1$ and $x = x_2$) the expressions (22) and (23) and (24) flow smoothly each into the next, with no discontinuity either of ordinate or of slope. This imposes four conditions on the five coefficients, and therefore fixes the relative values of all of them—in other words, determines them completely except for a common arbitrary factor which corresponds to the intensity of the incident beam, and is irrelevant to the course of the argument.

In particular, this requirement of continuity imposed by the fundamental principles of quantum mechanics upon the acceptable solution of (21) fixes the ratio of the amplitudes C_1 and A_1 of "transmitted" and "incident" wave-train. From Gurney and Condon I quote an approximate formula [26] for the ratio of the squares of these amplitudes, denoting them on the left by the customary symbols:

$$\begin{aligned} \frac{(\Psi\Psi^*)_{\text{trans.}}}{(\Psi\Psi^*)_{\text{inc.}}} &= \phi(E) \exp.\,[-(4\pi a/h)\sqrt{2m(NeV_m - E)}], \\ \phi(E) &= 16(E/neV_m)(1 - E/neV_m). \end{aligned} \tag{25}$$

[26] Exact formula given by E. U. Condon, *Reviews of Modern Physics*, 3, 57 (1931).

This expression does not vanish suddenly as soon as E drops below neV_m, but falls away continuously—and very rapidly, it must be admitted, owing to the exponential factor—as E diminishes from neV_m on downwards. Its value depends on a, the breadth of the hill (Fig. 14) in such a way that the broader or thicker the hill of given height the less the amplitude of the transmitted waves: the thicker the hill, the more nearly it comes to fulfilling the classical quality of being a perfect obstacle to particles having insufficient energy to climb it!

I rewrite (25) in the equivalent form,

$$\frac{(\Psi\Psi^*)_{\text{trans.}}}{(\Psi\Psi^*)_{\text{inc.}}} = \phi(E) \exp.\left[-(4\pi/h)\int_{x_1}^{x_2}\sqrt{2m(NeV_m - E)}dx, \quad (26)\right.$$

the integral in the exponent being taken "through the hill" from x_1 to x_2. This form is generalizable. Take the case of Fig. 14: the ratio of the squared-amplitudes of transmitted and incident wave-trains is given, according to Fowler and Nordheim, by an expression which is of the type (26), except that $\phi(E)$ is a somewhat different function (it is $4[(E/NeV_m)(1 - E/NeV_m)]^{1/2}$). The distance from x_1 to x_2, over which the integration is carried, obviously depends on E in this and every other case but the particular one of Fig. 14. Take finally the general case of a rounded hump, such as appears in Fig. 12. According to Gamow, a formula of type (26) is approximately—not exactly—valid for every such case, $\phi(E)$ being given by him as simply the number 4 when the hill descends to the same level on both sides as in Fig. 14; while in the general case where the potential-curve approaches different asymptotes at $-\infty$ and $+\infty$—say zero at the latter, V_r at the former—the factor $\phi(E)$ assumes the form $4[E/(E - neV_r)]^{1/2}$. Now E was the kinetic energy of the particles at infinity in the direction whence they come, and $(E - neV_r)$ will be the kinetic energy of the particles at infinity in the direction whither they are going. We have been denoting the first of these quantities by K; denote the second by K_r, and the corresponding velocities by v and v_r. Then $\phi(E)$ can be written as $4v/v_r$.

The question must now be answered: what is the actual relation between the ratio $(\Psi\Psi^*)_{\text{trans.}}/(\Psi\Psi^*)_{\text{inc.}}$, and the probability that a particle will traverse the hill? In associating waves with corpuscles, it is the rule to postulate that the square of the amplitude of the waves at any point is proportional to the number-per-unit-volume of corpuscles in the vicinity of that point. If one prefers to think of a single particle instead of a great multitude, one may say that the square of the amplitude of the waves at any point is proportional to the proba-

bility of the particle being at that point. Let us hold, however, to the picture of a dense stream of corpuscles approaching the potential-hill—say that of Fig. 14—from the right, and a much weaker stream receding from it on the left. If there are s times as many corpuscles per-unit-volume on the right as on the left, then there cannot be a steady flow unless only one out of s incident particles traverses the hill. The reciprocal of s is the fraction of particles getting through the hill, or the probability of a single particle getting through; and it is also the ratio $(\Psi\Psi^*)_{\text{trans.}}/(\Psi\Psi^*)_{\text{inc.}}$. This statement, however, is too narrow, being valid for the case where the speeds v and v_r of the particles on the two sides of the hill are the same. In the general case, we have:

Probability of transmission or penetration

$$= (v_r/v)(\Psi\Psi^*)_{\text{trans.}}/(\Psi\Psi^*)_{\text{inc.}}$$

$$= (v_r/v)\cdot\phi(E)\cdot\exp.\left[-(2\pi/h)\int_{x_1}^{x_2} dx\sqrt{2m(E-neV_m)}\right]. \quad (27)$$

In Gamow's approximation the product of the first two factors has the pleasantly simple constant value of 4. In the approximations of Gurney and Condon and of Fowler and Nordheim for the cases of Figs. 14 and 15, the product is some function of E which the reader can construct from the foregoing equations. In all these cases, however, it is the exponential factor which dominates the trend of either member of (27) considered as function of E.

Now immediately one sees, that *if* transmutation is due to the penetration of a charged particle through a potential-hill or potential-barrier surrounding a nucleus—and *if* this penetration is governed by laws of quantum mechanics as illustrated in the one-dimensional cases —*then* when the number of observed transmutations is plotted against the kinetic energy K of the impinging particles, the curve should be expected to rise with a gradual smooth upward curvature from the axis of K; and there should be no critical minimum value of K for the advent of the phenomenon, but rather the beginning of perceptible transmutation should be observed at progressively lower and lower energy-values, as the sensitiveness of the detecting-apparatus is improved; and it may well be that transmutation can be detected when K is still so low, that the quotient of K by the charge of the particle is far smaller than any reasonable guess that can be made of the height of the barrier. All these are features of such curves as those of Fig. 17, or Figs. 16 and 17 of the Second Part. The adoption of quantum mechanics permits us to accept these features without ascribing them to the hypothetical functions denoted by f_1 and f_2 in equation

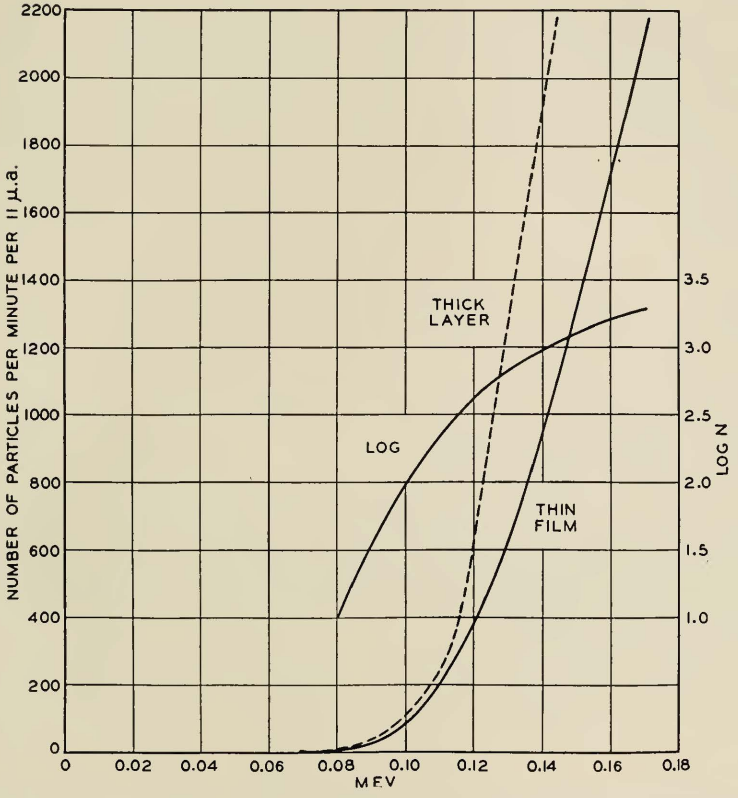


Fig. 17—Transmutation of boron by impact of protons: rate of observed transmutation as function of K, for a very thin film and for a thick layer. (Oliphant & Rutherford, *Proc. Roy. Soc.*).

(20), though it does not rule out the possibility that these functions may have influence upon the curves.

But not all of the curves of probability-of-transmutation versus K are of the simple type of Fig. 17. There are also some which show distinctly-marked peaks superimposed upon the gradual upward sweep: that of Fig. 18 for example, which relates to the transmutation of beryllium by impact of alpha-particles with emission of neutrons, presumably by the process

$$_2He^4 + {}_4Be^9 = {}_6C^{12} + {}_0n^1$$

and that of Fig. 19, which relates to the presumptive process,

$$_2He^4 + {}_{13}Al^{27} = {}_{14}Si^{30} + {}_1H^1.$$

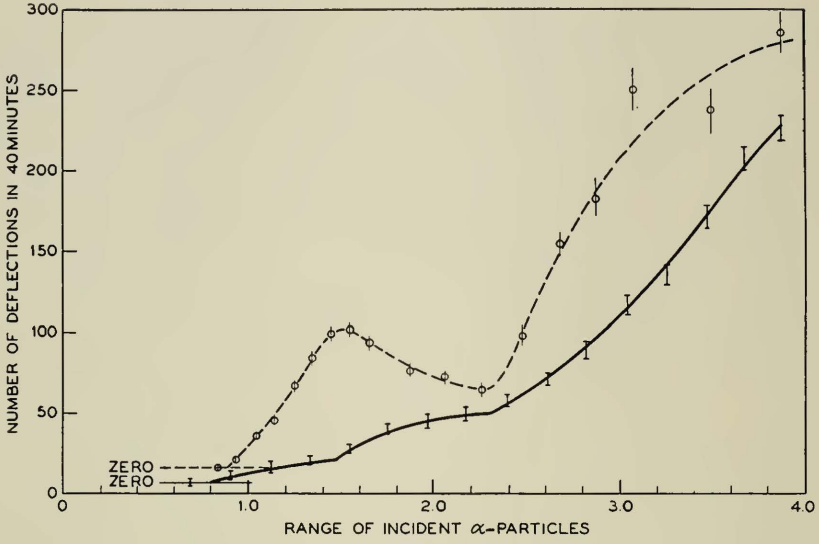


Fig. 18—Transmutation of beryllium by impact of alpha-particles, with production of neutrons; rate of observed transmutation as function of K, for a very thin film and for a thick layer (dashed and full curves respectively), illustrating resonance. (Chadwick; *Proc. Roy. Soc.*).

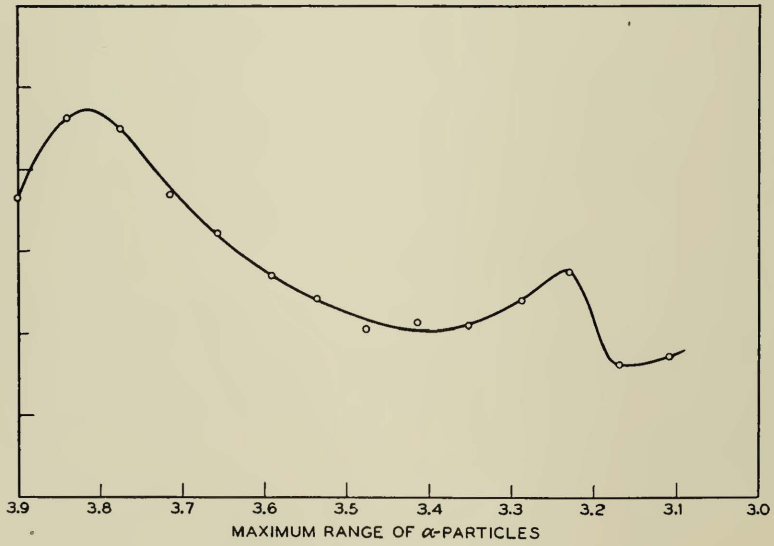


Fig. 19—Transmutation of aluminium by impact of alpha-particles, with production of protons; rate of observed transmutation as function of residual range of alpha-particles, illustrating resonance. (Chadwick & Constable; *Proc. Roy. Soc.*).

In Fig. 19 the abscissa is not K, but a quantity (the range of the impinging alpha-particles) which increases more rapidly than K; but this does not affect the meaning of the peaks. Moreover, there is abundant indication that quantities of such curves are simply waiting for someone to take the data and plot them; for this is the phenomenon of "resonance" to which many pages [27] were devoted in the Second Part, and which has chiefly been observed by the other methods there described, but should always manifest itself in this way when the proper experiments are performed.

If we wish to interpret this without letting go of the classical theory, we must say that either or both of the functions f_1 and f_2 have maxima at certain values of K. But here again, the adoption of quantum mechanics may make this step superfluous. For consider the one-dimensional potential-distribution of Fig. 16, a valley between two hills, with energy-values reckoned from the bottom of the valley. If the wave-equation be solved for this potential-distribution and for any such value of particle-energy E as the dashed line of Fig. 16 indicates—such a value, that according to classical theory a particle possessing it might either be always within the valley or always beyond either hill, but never could pass from one of these three zones to another—a curious result is found. For the solutions which the laws of quantum mechanics demand and accept, the ratio of squared-amplitude $\Psi\Psi^*$ within the valley to squared-amplitude $\Psi\Psi^*$ beyond either hill is usually low, but for certain discrete values of E it attains high maxima!

Now the three-dimensional nucleus-model of which I am speaking resembles this case more than it does the other one-dimensional cases of Figs. 14 and 15, because it consists of a potential-valley surrounded on all sides by a potential-hill. One may therefore expect the probability of entry or penetration to pass through maxima such as are symbolized by the dips in the "fence" of Fig. 13, entailing maxima in the curve of probability-of-transmutation P_t plotted as function of K. Such is the quantum-mechanical explanation of the phenomenon of "resonance," which indeed derives its name from this theory; for the values of K or E at which the maxima occur are those for which the amplitude of the oscillations of the Ψ-function in the valley within the barrier are singularly great.

One wants next to know what quantitative successes have been achieved in predicting or explaining such things as the actual locations of the resonance-maxima, or the precise trend of the curve of P_t-vs-K

[27] "Nucleus, Part II," pp. 148–153; more fully treated in *Rev. Sci. Inst.*, **5**, 66–77 (Feb. 1934).

as it rises away from the axis of abscissæ. Here it must be admitted that almost everything remains to be done. The locations of the resonance-maxima must be expected to depend upon the details of the potential-distribution within the valley, of which there is as yet no notion. The precise trend of P_t as actually observed cannot be the same as that of (27) however good the theory may be, first because not all of the impacts are central (page 596), then because in most (not quite all) of the experiments the bombarded substance is in a thick layer instead of a thin film (page 597), and finally because of the functions f_1 and f_2 of equation (20). Much as we should like for simplicity to put these functions equal to unity and zero respectively, and even though quantum mechanics has removed some of the obstacles to doing so, yet we are obliged to take them into account—the most striking and cogent reason being that with a variety of elements, P_t is not the same for impact of protons as for impact of deutons, though ne is the same for both![28]

The situation being such, one cannot ask as yet for accurate statements about the values of r_m and V_m, the constants of the "crater model" exhibited in Figs. 12 and 13. These must wait upon a thoroughgoing fitting of the theory to the experimental curves of P_t-vs-K, involving a decision as to the magnitude of f_1. The values of V_m for several of the lighter elements have been estimated from the data on transmutation, but the procedure of arriving at the estimates

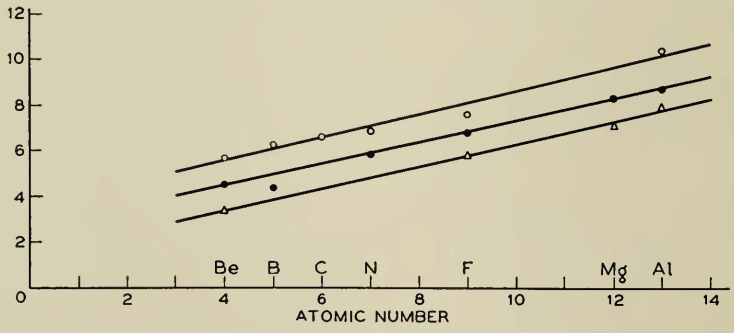


Fig. 20—Resonance-levels and (estimated) heights of potential-barriers for some of the lighter elements, deduced from observations of transmutation. (Pollard; *Phys. Rev.*)

has not (so far as I know) been published. I reproduce as Fig. 20 a graph of Pollard's, the circles along the uppermost line showing the estimated values of eV_m and the crosses along the other two lines

[28] Also it has been said that the shapes of the best experimental curves of P_t-vs-K imply that as K is increased, f_1 increases at first and then becomes constant; but there is a great lack of published theory on these matters.

showing the values of eV at which resonance occurs. The linear trends suggest that these may be properties of the nucleus which are susceptible of simple interpretations. There are also the estimates of V_m made from observations on alpha-particle scattering, most of which are merely minimum-admissible-values below which V_m cannot lie, while a few are more definite. As for r_m, there is at any rate nothing

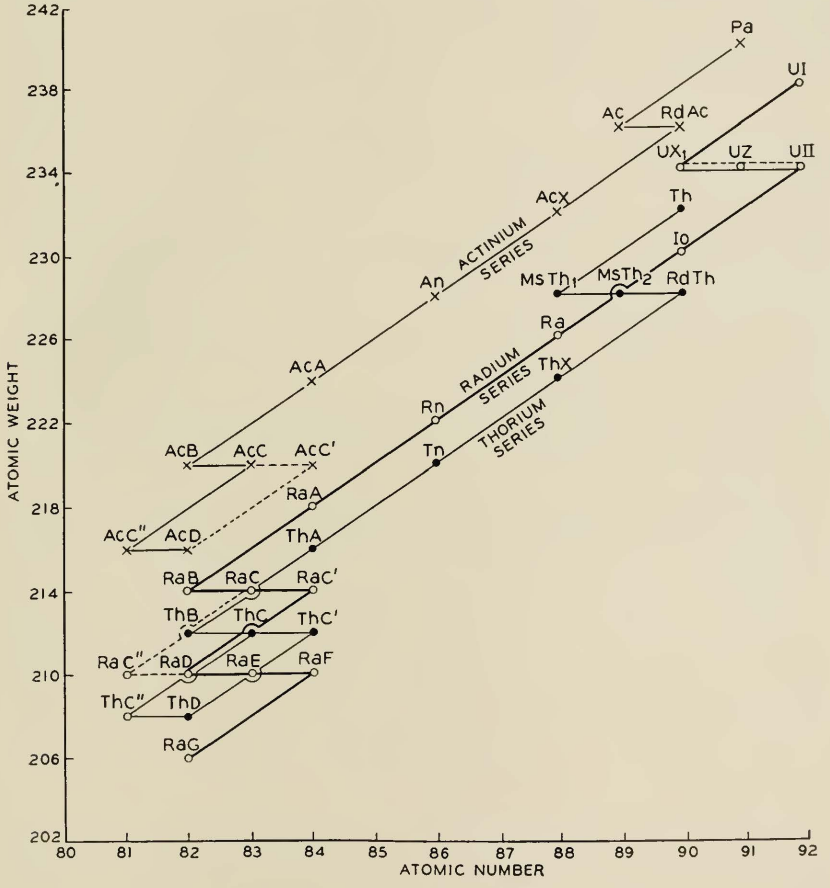


Fig. 21—Genealogies of the radioactive elements. (The actinium series is plotted some distance above the others for legibility, but almost certainly An should lie one unit below Tn, the rest correspondingly).

to indicate that we must make it higher than the values—a few times 10^{-13} cm.—which many reasons impel us to assign to the dimensions of nuclei.

Thus the quantum-mechanical theory of transmutation is as yet

in a primitive state, and indeed not advanced enough (in my estimation) to be considered fully proved by its own successes. Quantum mechanics has, however, many other buttresses, quite sufficiently many to allow us to take it for granted; and in this particular field it has the prestige of prophetic powers. Until the experiments of Cockcroft and Walton, no one had ever effected transmutation except with alpha-particles of charge $2e$ and energy K amounting to several millions of electron-volts. Now Cockcroft and Walton say that they were encouraged to build the elaborate apparatus necessary for trying it with protons of energy much less than one million electron-volts, by Gamow's inference that particles of charge $+e$ should have a very much greater chance of penetrating through a potential-hill and into a nucleus, than particles of equal kinetic energy and only twice the charge—the inference from the fact that ne occurs in the exponent of the exponential function appearing in equation (23) and others like it. Moreover, the phenomenon of resonance was predicted by Gurney (and mentioned by Fowler and Wilson, who, however, apparently did not believe that it could ever be observed) before it was discovered in the experiments of Pose.

The merits of the crater model with the quantum-mechanical theory have, however, not yet been fully presented, for I have left to the last their application to radioactivity.

One of the principal features of radioactivity—both the "induced" variety described in the early part of this article, and the "standard" variety known these thirty-five years—is the exponential decline or decay of the intensity, hence of the quantity of any radioactive substance, as time goes on. This signifies that the average future duration, reckoned from any instant of time, of all the atoms surviving unchanged at that instant, is the same whichever instant be chosen—or, that the probability that an atom, not yet transformed at instant t_0, shall undergo its transformation within (say) a second of time beginning at t_0, has the same value however long the atom may have existed up to this arbitrarily-chosen-moment t_0.

All this is commonly expressed by saying that radioactive transformations obey the laws of chance. I quote (not for the first time) a passage from Poincaré, which illustrates how this had to be interpreted before the advent of quantum mechanics; I take the liberty of writing "nucleus" where he wrote "atom":

". . . If we reflect on the form of the exponential law, we see that it is a statistical law; we recognize the imprint of chance. In this case of radioactivity, the influence of chance is not due to haphazard encounters between atoms or other haphazard external agencies. The

causes of the transmutation, I mean the immediate cause as well as the underlying one (*la cause occasionnelle aussi bien que la cause profonde*) are to be found in the interior of the atom [read, in the nucleus!]; for otherwise, external circumstances would affect the coefficient in the exponent. . . . The chance which governs these transmutations is therefore internal; that is to say, the nucleus of the radioactive substance is a world, and a world subject to chance. But, take heed! to say 'chance' is the same as to say 'large numbers'—a world built of a small number of parts will obey laws which are more or less complicated, but not statistical. Hence the nucleus must be a complicated world." [29]

Well! the advent of quantum mechanics has made unnecessary the conclusion which Poincaré was obliged to draw; for according to this doctrine, the statistical law is characteristic as much of a single particle confronted with a potential-hill, as of the greatest conceivable number of particles mixed up together. It must be admitted that Poincaré's conclusion is probably right enough for the radioactive nuclei of which he knew, all of which must be conceived to comprise several hundreds of particles, protons and electrons and neutrons and the like; but it is not enforced by the reason which he gives, if quantum mechanics is valid. For reversing the argument of previous pages: if in the valley-enclosed-by-hills which is illustrated (for the oversimplified one-dimensional case) by Fig. 16, we postulate a particle and the waves associated with that particle, then the quantum-mechanical boundary-conditions require waves beyond the hills as well, and the coexistence of waves without and within implies a tendency—a tendency governed by the "laws of chance," a probability—for the particle to escape from within to without. As soon as the physicist has successfully made the effort of consenting to quantum mechanics, he is dispensed from the further effort of contriving nuclear models with special features to account for the law of decay of radioactive substances.

Like the probability of entry, the probability of escape of the particle from the confined valley is governed by the ratio of the squared amplitudes $\Psi\Psi^*$ within the valley and beyond the hill (the latter in the numerator). It thus is governed by the exponential function,

$$\exp.\,[-\,(4\pi/h)\int\sqrt{2m(Ne\,V_m - E)}dx],$$

of which we have already made the acquaintance, multiplied by a

[29] H. Poincaré: "Dernières Pensées," pp. 204–205 (he credits Debierne with the idea).

factor $\phi(E)$ which itself may be a function of E the energy of the particle, but is of secondary importance. Such a formula, in the case of penetration from without, represented the probability of entry for a single approach of the particle to the hill. This suggests that we should deem it in this case as representing the probability of escape for a single approach of the particle from the depths of the valley to the inner side of the hill. Suppose the valley to be of breadth a, the particle to be bumping back and forth in it with speed v_i: the number of approaches of the particle per unit time to the hill will be equal to v_i/a. If the bottom of the valley is at the same level as the axis of abscissæ in Fig. 16, v_i is equal to $\sqrt{2E/m}$; if the valley is deeper, v_i is greater. We deduce for the mean sojourn of the particle within the valley, which is the reciprocal of the probability-of-escape-per-unit-time, the expression:

$$T \geqq [(v_i/a)\phi(E)]^{-1} \exp. [+ (4\pi/h) \int \sqrt{2m(NeV_m - E)}dx]. \quad (28)$$

The aspect of this expression is far from encouraging to one who wishes for a striking quantitative test of the theory. Its value depends not merely on the breadth a assumed for the space within the potential-hill and the height V_m of the hillcrest, but on the details of the shape assumed for the potential-curve of Fig. 12 both within and without the crest; and since there is little or no independent knowledge of these qualities of the nucleus, they may be adjusted practically at will to fit any observed value of T whatever. Furthermore it was obtained by making certain crude assumptions and certain not very close approximations.

One essential test, however, can be applied to it, which it must pass; and pass it does. Let values of a and V_m, and a shape for the potential-hill of Fig. 12, be so chosen that for some particular radioactive element, RaA for instance, equation (28) agrees with experiment; which is to say, that when into the right-hand member of (28) is substituted for E the observed kinetic energy of the emerging alpha-particles, the value of this right-hand member becomes equal to the observed mean life of the element. Now let precisely the same values of a and V_m and the same shape of hill[30] be assumed for some other radioactive element, RaC′ for instance; in the right-hand member of (28), let the observed kinetic energy of the (main group of) alpha-particles for RaC′ be substituted for E; and let the value of T be computed. We

[30] Excepting that the two hills should be expected to slope off towards infinity in the manners of the two functions Z_1/r and Z_2/r, where Z_1e and Z_2e stand for the nuclear charges of the nuclei left behind after the alpha-particle departs, and are often (not always) different for two different radioactive substances.

should not expect a perfect agreement, since the two nuclei are not identical; but we should be disconcerted by a sharp disagreement, since both nuclei belong to elements of which the nuclear charges differ at most by only a few per cent and the nuclear masses by little more. A very great disagreement would in fact be gravely injurious to the theory. Making the test, Gurney and Condon found, however that there is no grave disagreement: the theory survives the test.

A very similar test was applied with greater minuteness by Gamow. For each element he assumed a potential-hill having a vertical rise on the inward side, and on the outward side a curved slope conforming exactly to the function $(Z - 2)/r$, where Z stands for the atomic number of the element before the alpha-particle quits it and consequently $(Z - 2)e$ stands for the charge of the residual nucleus. In other words, he postulated a classical inverse-square electrostatic field ("Coulomb field") from infinity inward to a distance r_0 from the centre of the nucleus, and at r_0 a discontinuous potential-fall. This is a potential-distribution distinguished by a single disposable constant, to wit, r_0; for V_m itself is determined by r_0 and Z.

Gamow proceeded to compute what value must be assigned to r_0 in order to achieve agreement between theory and experiment for each of the twenty-three alpha-emitters. Approximations must be made in carrying through the calculations; those which Gamow employed convert equation (28) into this:

$$\log_e T = -\log_e (h/4mr_0^2) + 8\pi^2 e^2 \sqrt{m}\,(Z - 2)/h\sqrt{2E} + (16\pi e\sqrt{m}/h)\sqrt{Z - 2}\sqrt{r_0}.$$

Putting for E the kinetic energy of the alpha-particles and for T the mean lives of the several elements, he evaluated r_0. Had the values proved very different for the various alpha-emitters, it would have spoken ill for the theory; but all the values were comprised between 9.5 and 6.3 times 10^{-13} cm. The order of magnitude is satisfactory; the differences between the several values are by no means disagreeably great; and there are even signs of a systematic upward trend of the values of r_0 with the atomic numbers of the nuclei. The quantum-mechanical theory and the crater model of the nucleus so pass their crucial test.

The Measurement and Reduction of Microphonic Noise in Vacuum Tubes

By D. B. PENICK

The microphonic response of different types of vacuum tubes to the same mechanical agitation covers a 70 db range of levels. Tubes of the same type, on the average, cover a range of about 30 db. These response levels are too sensitive to minute variations in testing conditions to be measurable with any great precision, but values which are reproducible to within 5 db are obtainable with a laboratory test set comprising a vibrating hammer agitator, a calibrated amplifier, and a thermocouple galvanometer indicator. Sputter noise is made measurable by frequency discrimination methods.

Minimum microphonic disturbance under given service conditions is attained by using the less microphonic types of tubes which are available, by selecting the quieter tubes of a given type for use in positions sensitive to mechanical disturbance, and by protecting the tubes from mechanical and acoustic vibration. Examples of quiet triodes are the Western Electric No. 264B (filament) and No. 262A (indirectly heated cathode). Indirectly heated cathode type tubes are intrinsically less microphonic than filamentary types. Further microphonic improvement in the tubes themselves is made difficult by requirements for favorable electrical characteristics. Well designed cushion sockets can reduce microphonic levels by as much as 30 db, and other methods of cushioning, more expensive and less compact, can extend the reduction even farther. Sputter noise can be eliminated almost entirely in most types of tubes by commonly applied design features and manufacturing methods.

A MAJOR problem which has had to be met by every engineer who has designed a high gain amplifier is that of elimination or reduction of noise. Noise of one kind or another, extraneous to the desired signal, is always present in any amplifier, and sets a lower limit on the smallness of the signal which can be amplified without intolerable interference. In many experimental and commercial amplifiers, the technical and economic obstacles to noise reduction necessitate a compromise between inherent noise level and sufficient volume range for ideal reproduction. The possible sources of this noise are numerous and include power supply, faulty contacts, insulation leaks, pick-up from stray fields, and many other disturbing elements. Among the most persistent types of noise, however, requiring particularly careful design for their elimination or satisfactory reduction, are three which originate in the vacuum tubes themselves, namely, fluctuation noise, microphonic noise, and sputter noise.

Fluctuation noise has been treated at some length by Schottky,[1]

[1] W. Schottky, *Ann. der Phys.*, v. 57, p. 541, 1918; v. 68, p. 157, 1922. *Phys. Rev.*, v. 28, p. 74, 1926.

the discoverer of one of its sources and by several other investigators.[2] It is the fundamental noise arising from the circumstance that the electron current is a stream of discrete particles rather than a continuous flow. For ordinary types of low power tubes of good design, over the audio band of frequencies, the root-mean-square amplitude of this noise is equivalent to about 1 microvolt (120 db below 1 volt) of noise voltage applied to the grid of the tube. G. L. Pearson,[2] in a recent paper, has pointed out that for the best signal-to-noise ratio, the input impedance should be so large that the thermal noise arising in this impedance predominates over the fluctuation noise arising in the tube. In many broad-band or high-frequency systems, however, such an ideal condition is practically unattainable, and fluctuation noise remains as a limiting factor.

It is with the second type, microphonic noise, that we are particularly concerned here, though sputter noise is also of interest and will be dealt with briefly in a later paragraph. Microphonic noise, as the name is usually applied, is the familiar gong-like sound which is always produced when a vacuum tube, followed by a sufficiently high-gain amplifier and a sound reproducer, is subjected to a mechanical shock. Its origin is in the vibrations of the various elements of the tube, which make minute, more or less periodic changes in the spacings of the elements and therefore make corresponding changes in the plate current, whose value at every instant depends on these spacings. Its intensity in a given tube depends on the type and intensity of agitation to which the tube is subjected. For a given agitation, microphonic noise may be reduced either by stiffening and damping the tube structure, thereby reducing the amplitude and duration of vibration of the elements, or by cushioning the tube so that it receives only part of the original agitation.

In order to treat the problem of noise reduction intelligently, it is necessary to have a measure of the effectiveness of treatments applied. To this end, the properties of microphonic response in vacuum tubes have been studied, and a test set has been designed and built for laboratory use which affords a quantitative measure of microphonic response in tubes, and of effectiveness of cushioning in cushion sockets. Some of the more important characteristics of microphonic noise will now be considered.

[2] "The Schottky Effect in Low Frequency Circuits," J. B. Johnson, *Phys. Rev.*, v. 26, pp. 71–85, July, 1925.
"A Study of Noise in Vacuum Tubes and Attached Circuits," F. B. Llewellyn, *Proc. I.R.E.*, v. 18, pp. 243–265, Feb., 1930.
"Shot Effect in Space Charge Limited Currents," E. W. Thatcher and N. H. Williams, *Phys. Rev.*, v. 39, pp. 474–496, Feb. 1, 1932.
"Fluctuation Noise in Vacuum Tubes," G. L. Pearson, *Physics*, v. 5, p. 233, September, 1934. Also published in this issue of *Bell Sys. Tech. Jour.*

Factors Affecting Microphonic Noise Levels

In the first place, it may be pointed out that the production of microphonic noise in commercial types of vacuum tubes is an extremely complicated phenomenon. Each individual component of the mechanical structure is a complete vibrating system having several modes of vibration and natural resonant frequencies, and usually very little damping as compared with electrical circuits. These components, all coupled together mechanically in various ways, form a mechanical network much more complex than the electrical networks encountered in communication engineering practice.

The complexity of the mechanical vibration is reflected in the complex character of the noise itself, and is admirably illustrated by the frequency-response characteristics published by Rockwood and Ferris,[3] and by similar characteristics obtained in the course of this work. It is further demonstrated by the experimental fact that when a large group of supposedly identical tubes is tested by applying the same mechanical vibration to each tube in turn, mounted in the same socket, the response levels of individual tubes may differ from each other by as much as 30 db for representative types of tubes. Such a magnitude of variation would not be expected to result from the comparatively small dimensional variations tolerated in manufacture, and must be explained by the exaggerating effect of intercoupled mechanical resonances in a complicated vibrating system. A curve showing a typical distribution of microphonic response levels in a group of tubes of the same type measured under identical conditions of agitation is shown in Fig. 1. The general shape of this curve is characteristic of any function subject to random variations about a mean, and the range of levels included between the quietest and the noisiest tubes, approximately 30 db, is about average for different types of tubes. Tubes of exceptionally firm construction and of fairly wide spacing, may vary over as small a range as 15 or 20 db, while tubes in which there is a possibility of slight looseness of parts, may vary over as great a range as 40 db or even more.

The nature and intensity of the agitation, the vibrational characteristics of the tube mounting, and the type and degree of mechanical coupling between the tube and its mounting also play an important part in the determination of the microphonic response of a particular tube, since the tube and its closely coupled mounting make up a single vibrating system. The reason for considering the coupling apart from the mounting is that it is usually of the pressure-friction variety and is subject to random variation.

[3] "Microphonic Improvement in Vacuum Tubes," Alan C. Rockwood and Warren R. Ferris, *Proc. I.R.E.*, v. 17, pp. 1621–1632, Sept., 1929.

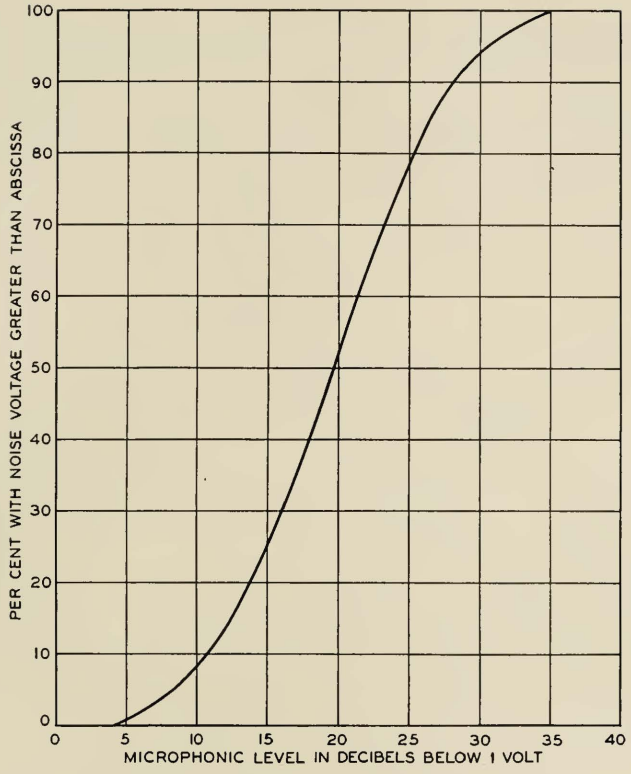


Fig. 1.—Typical distribution of microphonic noise levels produced by a constant, artificial, mechanical stimulus (Western Electric No. 102F Vacuum Tube).

It is found experimentally that with no type of commercial socket which has been tested can a tube be removed and reinserted, or even be left in the socket for a period of time subject to incidental jars and temperature fluctuations, with the expectation of perfectly reproducing a previously measured microphonic level. The sort of random variation which is usually found is illustrated in the reproducibility chart of Fig. 2. In this chart, each point represents two separate observations of microphonic level made on the same tube in the same apparatus, the tube having been removed and reinserted between the two observations. The two levels thus measured are represented by the abscissa and ordinate, respectively, so that if the measurements were perfectly reproducible, all of the points would lie on a straight line making an angle of 45 degrees with the coördinate axes and passing through the origin. The amount of maximum scattering here is about 5 db and may be considered an average value. In some cases, with commercial

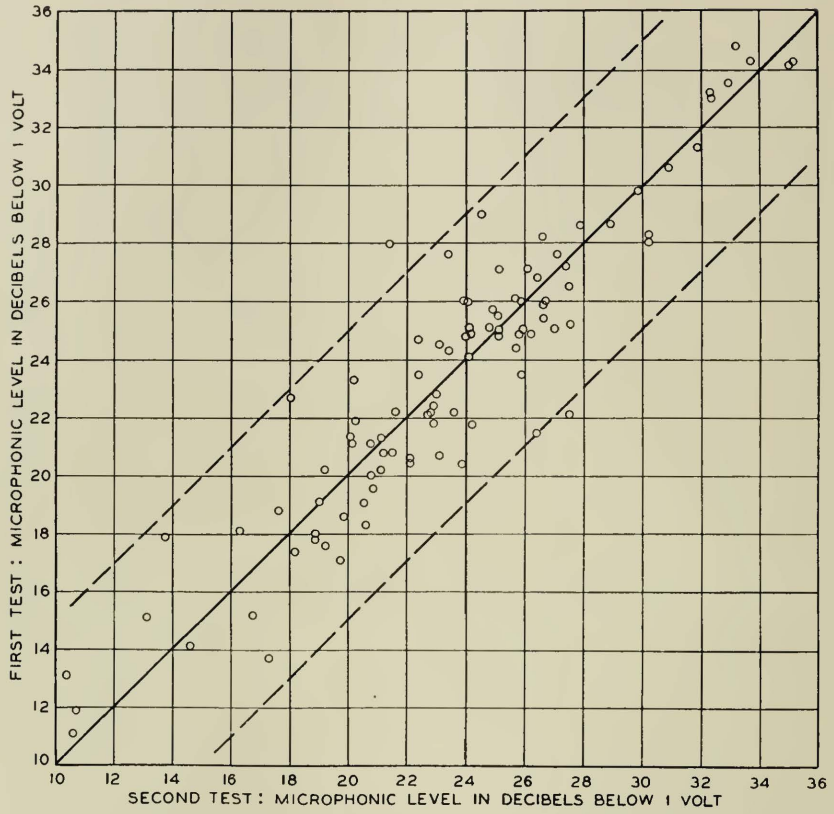


Fig. 2—Reproducibility of microphonic noise level measurements using a commercial socket with a constant, artificial, mechanical stimulus (100 No. 102F Tubes).

sockets, it has been observed to be as low as 3 db and in others as high as 8 db.

In order to show that this random variation is not due to the tube itself, experiments have been made with two forms of suspension which minimize the reaction of the mounting on the vibration of the tube and so reduce as far as possible the effect of variation in coupling. In one set-up, the tube is hung by a single thread of rubber, stretched to its elastic limit, the electrical connections being made by very light, flexible leads fastened with light clips directly to the prongs of the base. In the other, the tube is clamped lightly between two large blocks of very soft sponge rubber, and the electrical connections are made through mercury cups into which the base prongs dip. In both cases, the agitator is a light pendulum striking the base or bulb of the tube. The two mountings give very similar results, and are charac-

terized by very much less scattering than any normal tube mounting, as may be seen in the correlation chart of Fig. 3, which is typical of all of the tests made with these light suspensions. The maximum scattering here is only about 1 db.

Going to the opposite extreme in tube mounting, similar tests have been made with the tube base held tightly in a split metal clamp,

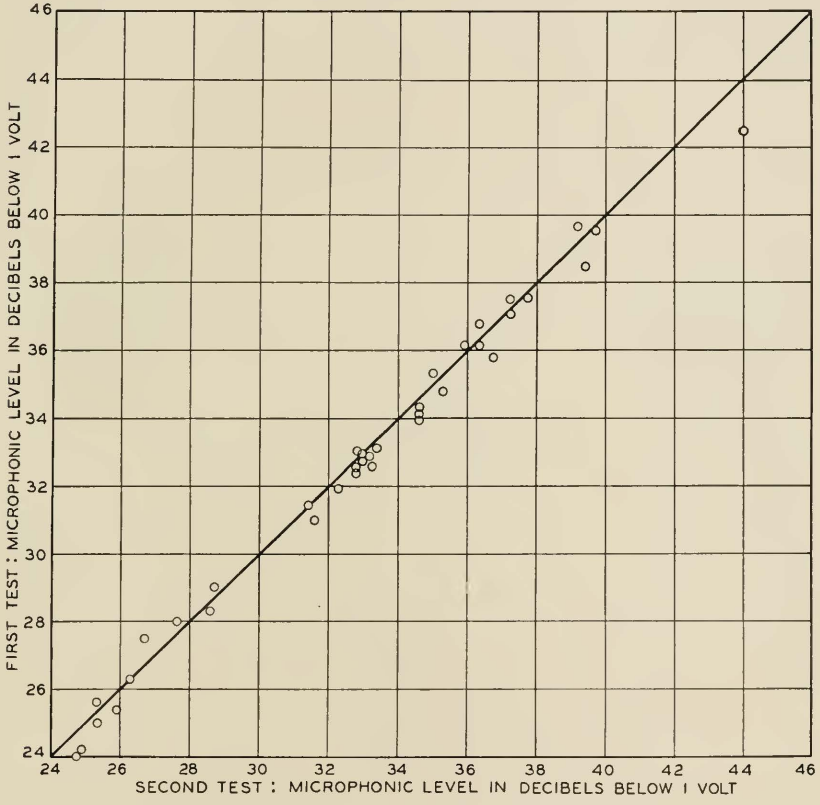


Fig. 3—Reproducibility of microphonic measurements using a rubber clamp tube mounting with a constant, artificial, mechanical stimulus (37 No. 102F Tubes).

which itself is bolted rigidly to a heavy base. As is to be expected, the observed levels vary widely and erratically for successive insertions of the tube, and the mere tightening or loosening of the thumb-screw controlling the pressure of the clamp on the base in some cases changes the level by as much as 10 db.

As for the nature of the applied agitation and the vibrational characteristics of the tube mounting, a countless number of combinations

of these exists, each of which would agitate the tube in a different way. However, from tests made with a variety of mounting arrangements for the tube under test and a variety of degrees of intensity and points of application of forms of impact agitation, it may be concluded that in practical set-ups these factors may be varied widely without changing the general nature of the microphonic level measurements greatly. That is, the form and breadth of the distribution curve and the scattering of the points on the reproducibility chart for any typical group of tubes are likely to be quite similar to Figs. 1 and 2, respectively, for almost any practical impact agitator.

Although the general nature of the results obtained with various combinations of these agitator and mounting arrangements is about the same for all of them, there are certain particular differences, which show up chiefly in two characteristics. One is that the mean noise level of a group of tubes is in general not the same for different mountings and methods of agitation. That this must be true is fairly obvious and needs no comment. The other is illustrated in Fig. 4, which is a correlation chart showing typical results of measurements

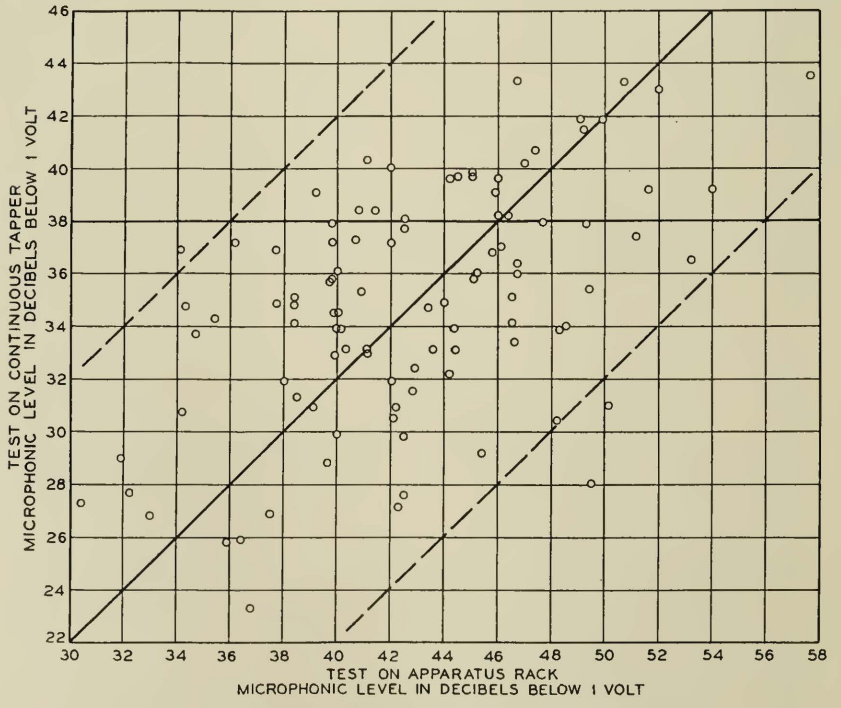


Fig. 4—Comparison of two tube mountings with a constant, artificial, mechanical stimulus (100 No. 102F Tubes).

of the same group of tubes on two different agitating systems. One system in this case consists of a rectangular slate block vibrated by repeated blows of an electrically operated hammer. The other system consists of a steel panel carrying the tube under test, mounted on an apparatus rack which is vibrated by a single blow from a steel ball falling as a pendulum against the rack. The points on this chart scatter about an ideal line over a band about twice as broad as that in Fig. 2 where a test is made and repeated on the same testing unit. It may also be observed that the mean noise levels produced by the two systems are different, about 35 and 43 db below one volt respectively.

The effect of varying the intensity of agitation is shown in Fig. 5.

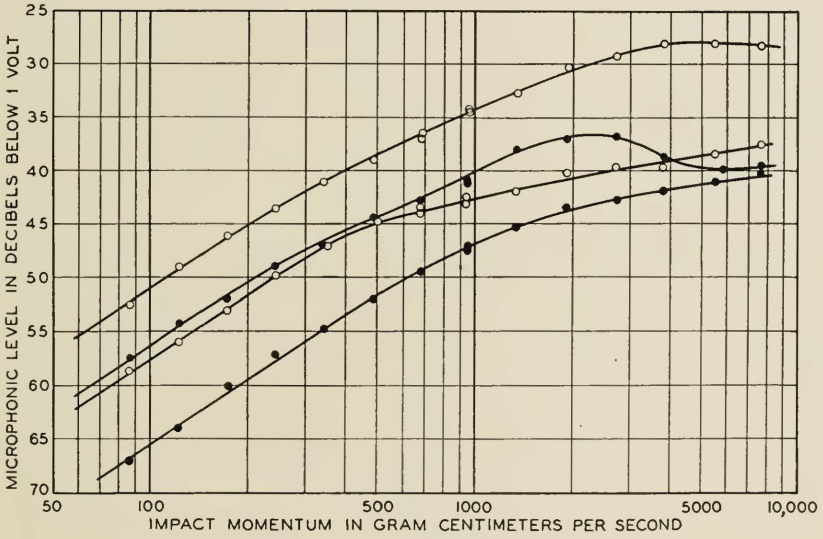


Fig. 5—Effect of intensity of agitation on 4 No. 264B Tubes.

The four curves represent four No. 264B Tubes tested under the same conditions. In making the measurements, the tube under test is mounted in an ordinary socket on a heavy base, which is agitated by means of a pendulum swinging against it and making one rebound. From measurements of the initial swing of the pendulum, its rebound, and its mass, the total momentum imparted to the tube mounting during the impact can be calculated. This quantity is plotted as abscissa in the figure and is proportional to the initial velocity imparted to the tube mounting at the point of impact. Different values of momentum are obtained by varying the initial swing and the mass of the pendulum. At the lower values of momentum, the observed

points lie, within the limits of experimental error, on parallel straight lines so drawn that along them the microphonic noise level expressed in volts is proportional to the initial velocity of the tube mounting. Some such relation as this would be expected to hold as long as the response of the system is linear. The departure from this law at higher values of momentum, then, probably indicates non-elastic motion either of elements of the tube with respect to one another or of the tube with respect to the socket. It may be noted in passing that the No. 264B Tube is exceptionally rigid in structure and that in more loosely constructed tubes, the straight line part of the response curve ends at much lower intensities of agitation.

Since the noise energy is spread over a band of frequencies, the microphonic response observed in any given reproducing system depends also on its frequency-response characteristic. In the usual type of volume indicator, the response is substantially uniform over the audio range of frequencies, but where the final auditory sensation is being considered, the overall characteristic is modified by that of the ear of the listener.[4] The effect of changing the overall response characteristic is illustrated for one particular case in Fig. 6 and Table I.

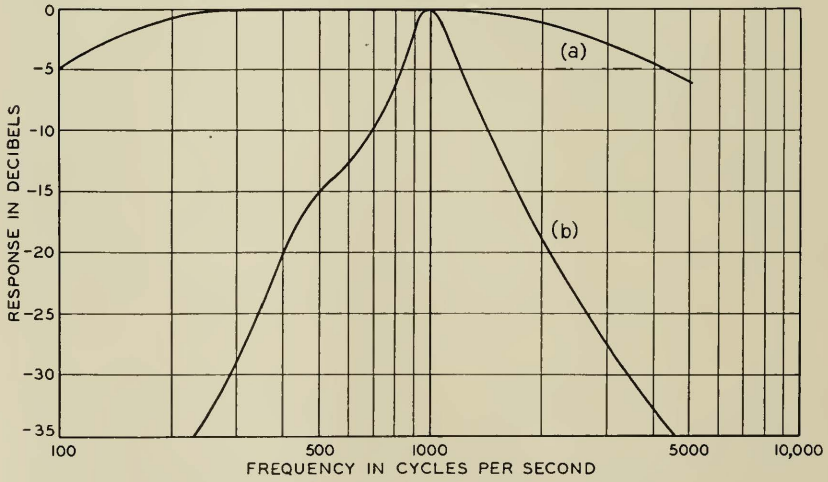


Fig. 6—Microphonic noise amplifier frequency characteristics.

Here, two sets of measurements have been made on each of three types of tubes, one set using an amplifier having a fairly uniform gain characteristic, curve (a), Fig. 6, and the other set using a weighted amplifier, weighted as in curve (b) in the figure. The same agitator and

[4] "Speech and Hearing," H. Fletcher. D. Van Nostrand Co., 1929.

TABLE I

MICROPHONIC LEVELS IN DB BELOW 1 VOLT

Type Tube	Amplifier (*a*)	Amplifier (*b*)	Difference
264B.....	37.9	63.2	25.3
102F.....	29.7	44.7	15.0
231D.....	18.2	38.9	20.7

indicator are used in both sets of measurements. Table I gives the mean noise levels for each of the three types of tubes and the differences between the values obtained for each type with the two amplifiers. The results represent about ten tubes of each type. The weighted amplifier, of course, gives the lower levels for all tubes since the noise components at all frequencies except 1000 c.p.s. are amplified less by this amplifier than by the more uniform amplifier. The magnitude of the difference in level depends on the frequency spectrum of the microphonic noise being measured and in general is different for different types of tubes as in this illustration.

Still another important factor which affects the microphonic response of a given system is the relation between the rate of variation of the noise intensity and the time-response characteristic of the system as a whole, usually determined by the indicator. The indicator may be a meter, oscillograph, or other device, or it may be the ear of a listener. A slow moving indicator would respond less to a pulse of noise, such as might be produced by a single shock to a tube, than a more rapidly responding indicator having the same sensitivity to a steady signal. The time required for the ear to reach its maximum response to a suddenly applied sound is about 0.2 second.[5]

The degree of importance of the time-response characteristic of the indicator in measuring transient pulses may be inferred from Table II. This table gives the results of two sets of measurements made on the same three groups of tubes with a single impact type of agitator. The

TABLE II

MICROPHONIC LEVELS IN DB BELOW 1 VOLT

Type Tube	0.2 Second Indicator	2.0 Second Indicator	Difference
264B.....	37.9	50.5	12.6
102F.....	29.7	37.2	7.5
231D.....	18.2	25.2	7.0

[5] "Theory of Hearing; Vibration of Basilar Membrane; Fatigue Effect," G. V. Bekesy, *Physikalische Zeitschrift*, v. 30, p. 115, March, 1929.

two sets differ only in the indicators used. One requires approximately 2 seconds to reach its maximum deflection with a steady impressed signal, and the other requires about 0.2 second. The differences in level corresponding to different types of tubes do not vary greatly, but are nevertheless appreciable. They are, to some extent, a measure of merit of the tube, for a larger difference indicates higher damping of the microphonic disturbance, and high damping is of course desirable.

A Microphonic Noise Measuring Set

The type of test set which has been built in the course of this study for use in the laboratory, comprises an arbitrary standard of agitation, a calibrated amplifier, and an indicating instrument. The agitator consists of a heavy, rectangular slate base at one end of which are mounted sockets for several types of tubes. At the other end is an electrically driven vibrating armature carrying a hammer which strikes about 9 blows per second against a steel block bolted firmly near the center of the slate base. This unit is set on a thick felt pad in a felt lined copper box which provides electrical shielding and some degree of sound-proofing. The sockets used (except those for the bayonet-pin bases) are of the type in which contact springs push each base prong firmly to one side, against the body of the socket. This type has been found to stand up well under repeated insertions and withdrawals of tubes and gives as good correlations between repeated microphonic measurements as any type which has been tried.

The amplifier is basically a simple resistance-choke coupled unit having a frequency-response characteristic which is essentially flat (within 3 db) between 80 and 30,000 c.p.s. The tube under test, whose plate voltage is supplied through an 80-henry choke, works directly into a 100,000-ohm potentiometer, variable in 10 db steps, whose output is connected to the input of the amplifier. The indicator is a sensitive thermocouple galvanometer whose scale is marked off in db and half db divisions so that the noise level may be read directly from the setting of the input potentiometer and the position of the indicator. It has been found convenient to think of the noise level in terms of the root-mean-square voltage developed by the tube across the 100,000-ohm load resistance and to use 1 volt as the reference level. Accordingly, unless otherwise noted, the noise levels given herein are expressed as db below 1 volt across a 100,000-ohm load resistance.

In order to correct for time shifts in tube characteristics and battery voltages, provision is made for checking the amplifier calibration at any time by throwing a switch which transfers its input circuit from the tube under test to a local oscillator. This oscillator delivers a

small, fixed output voltage which is measured and set at a predetermined value with the aid of another thermocouple galvanometer. The amplifier gain may be adjusted, by means of a small range, continuously variable potentiometer until the indicator gives the proper reading to correspond with the known level of the applied input. With reasonably steady battery voltages, this calibration is necessary only two or three times in the course of a day's testing. The range of noise levels for which the amplifier is calibrated extends from 10 db above 1 volt to 65 db below 1 volt. This range has been found to include practically all tubes which it has been desired to test with the standard agitator.

The flat amplifier characteristic, which has been described, is normally used for general testing in connection with vacuum tube design work since it gives the highest microphonic level readings and therefore the most conservative picture of the performance of the tube from the standpoint of the designer. Provision is made, however, for switching in a specially designed weighted amplifier such as is used in making routine noise measurements in telephone speech circuits.[6] The frequency characteristic including this unit has already been shown in Fig. 6, curve (*b*), and is designed to compensate for the interfering effect of each component of noise on the average ear plus the effect of the frequency characteristic of the telephone subset. A similar weighting network compensating for the non-uniform frequency response of the ear alone would also be useful, but has not yet been provided.

Nature and Measurement of Sputter Noise

By making a slight modification of the amplifier circuit, this test set may also be used to measure sputter noise. Sputter noise is a descriptive name applied to a class of noises characterized by a harsh crackling or sputtering sound easily distinguished from the gong-like quality of microphonic noise or the steady roar of electron noise. It may occur either with or without agitation and is the result of discontinuous changes in electrode potential such as may be produced by imperfect contact between conducting members in a tube or by intermittent electrical leaks across insulation.

Sputter noise due to agitation is always accompanied by microphonic noise, and though it often contains instantaneous peaks of high intensity which constitute a very disagreeable and annoying type of interference, its total energy content is usually so small that it contri-

[6] "Methods for Measuring Interfering Noises," Lloyd Espenschied, *Proc. I.R.E.*, v. 19, pp. 1951–54, Nov., 1931.

butes very little to the ordinary microphonic level reading. Special methods must therefore be used in order to make measurements of sputter noise which are independent of microphonic noise. One method which has been found to be effective and convenient, is that of frequency discrimination. If the audio frequency components of the total noise are cut out, then microphonic noise is completely eliminated. Sputter noise, however, due to its discontinuous character has, theoretically, an infinite frequency spectrum, and, practically, one which extends at least into the broadcast band of radio frequencies.

In the microphonic noise test set, sputter noise measurement is provided for by switching in a high-pass filter cutting off sharply at 16,000 cycles. Greater sensitivity is also provided by additional stages of amplification to permit the measurement of the lower levels found to be characteristic of sputter noise in this frequency range. A schematic diagram of the microphonic and sputter noise test set is shown in Fig. 7. The weighted amplifier and the calibrating oscilla-

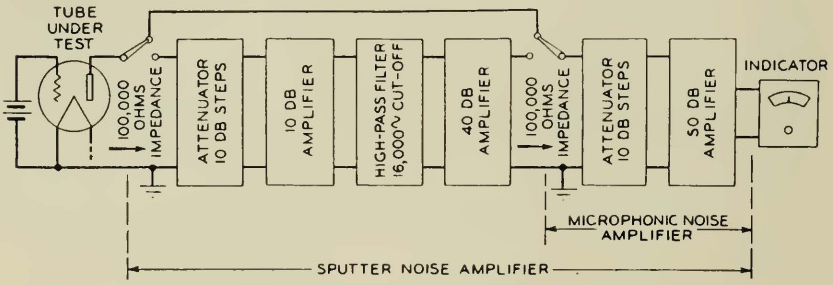


Fig. 7—Microphonic and sputter noise amplifier schematic diagram.

tors (one for microphonic noise and one for sputter noise) have been omitted for the sake of clearness.

Reduction of Microphonic Noise

The reduction of microphonic noise from the view-point of the vacuum tube designer is chiefly a matter of mechanical design and manufacturing technique. The quietest construction is obviously one which has the stiffest electrodes and supporting members, the shortest distances between points of support, and the highest damping of mechanical vibration. The extent to which these features can be incorporated in a practical tube design, however, is limited by the requirements for favorable electrical characteristics. A low filament current, for example, requires that the filament be small in diameter, which renders it more susceptible to vibration than a heavier filament; or if a tube has an indirectly heated cathode, it is a problem to support

the cathode rigidly without conducting away large amounts of heat along the supports. The diameter, length, and spacing of the grid lateral wires are fixed within relatively narrow limits when the desired values of amplification factor and internal impedance are fixed, precluding any important increase in stiffness here; and where high mutual conductance is desired, it is necessary to use relatively close spacings between the elements, under which condition a given amplitude of vibration produces a relatively large per cent change in spacing, and therefore a high microphonic noise level.

The Western Electric No. 264B Vacuum Tube is an example of what has been done in working for a stiff, compact structure. The plate support wires, which also support the whole top of the structure, are short, straight, and as heavy as is practicable, and an extra wire from the press braces the glass bead. One of the most important features of this tube, however, is its filament. In most filament type tubes the vibration of the filament is the chief source of microphonic noise. In the No. 264B Tube, therefore, the filament is made comparatively short and heavy and is mounted in the form of a broad, inverted V to whose apex considerable tension is applied by means of a cantilever spring. The effectiveness of this treatment may be seen from Table III which lists the mean noise levels for a number of types of Western Electric small tubes, and the maximum and minimum

TABLE III*

Class	No. of Samples Tested	Type Number	Microphonic Noise Level in db below 1 volt			
			Measured output Level			Equivalent Mean Input Level
			Max.	Min.	Mean	
Filament type triode	250	101D	23	38	32	47
	833	101F	8	30	19	35
	505	102F	9	36	20	46
	235	215A	12	42	27	41
	1,144	231D	2	28	16	33
	201	239A	4	36	22	37
	715	264B	30	52	42	58
Indirectly heated cathode type triode	99	244A	28	48	39	58
	448	247A	26	52	42	64
	452	262A	36	62	49	71
Screen grid and pentode	24	245A	18	39	29	63
	42	259A	2	36	20	61
	30	283A	4	42	21	62
	30	285A	12	30	23	57

* The microphonic properties of the No. 259A Tube given in this table are identical with those of the 259B discussed by Pearson.[2]

levels which have been observed for each type. The No. 264B, with a mean level of 42 db below 1 volt, is 20 db quieter than the No. 239A which it was designed to replace, and is the quietest of the filament type triodes. The next quietest tube of this structure is the No. 101D, in which the elements are supported from a rigid glass arbor and the filament is quite heavy, requiring one ampere of heating current. The No. 215A is almost identical with the No. 239A except for a firmer supporting structure which results in a 5 db improvement. The most microphonic of the types listed is the No. 231D, which has a very fine wire filament whose diameter is fixed by the requirement that the heating current be 0.060 ampere.

If the filament is the chief source of microphonic noise in filament type tubes, then it is to be expected that tubes having indirectly heated cathodes will be much less microphonic, inasmuch as the cathode is an extremely rigid member. An examination of Table III shows that this is indeed true. The No. 244A and No. 247A types, in which no special precautions have been taken to obtain quietness, are about as quiet as the No. 264B Tube. In the No. 262A Tube, therefore, it has been possible to reduce the microphonic noise still further, to 49 db below 1 volt, by cementing the elements into rigid supporting blocks of ceramic material. This tube is also quiet in other respects, notably in its freedom from AC hum picked up from the cathode heater circuit.[7]

In comparing tubes having widely different electrical characteristics, it is not quite fair to compare their noise output levels alone, for given two tubes having the same noise output, the tube having the higher gain can be used with smaller signal inputs and have no greater noise interference in the output. Accordingly, another column is given in Table III listing the equivalent noise input level which would produce the observed noise output if the tube itself were perfectly quiet. The ratio of this value to the signal input level is directly related to the degree of microphonic noise interference which is effective in the output of the tube. It is computed by adding the voltage gain expressed in db, of the tube in the measuring circuit, to the microphonic output level obtained experimentally. The value of this criterion is illustrated in comparing the noise interference produced by multi-element tubes and triodes. Multi-element tubes as a rule have higher noise output levels than triodes as may be seen by comparing the Nos. 245A, 259A, 283A, and 285A screen-grid and pentode types with the Nos. 244A, 247A, and 262A triodes. When account is taken of the higher voltage amplification of these former types, however, the noise inter-

[7] "Analysis and Reduction of Output Disturbances Resulting from the Alternating Current Operation of the Heaters of Indirectly Heated Cathode Triodes," J. O. McNally, *Proc. I.R.E.*, v. 20, pp. 1263–83, August, 1932.

ference as indicated in the equivalent input noise column of the table, compares quite favorably with that of the triodes.

From the point of view of the user of vacuum tubes, constrained to work with available types, the most effective means of microphonic noise reduction is the use of one of the quieter types of tubes which have been described. In cases where noise difficulties are experienced in existing apparatus not readily convertible to the use of a quieter type of tube, however, some relief may be gained by selecting the quieter tubes from a number of the type to be used. To be fully effective, the selection should be based on measurements made while the tube is in the socket in which it actually works. Under such circumstances, the measurements are reliable to within about 5 db.

Where selection in the field is not feasible, a smaller degree of relief may still be gained by selection at the factory. The degree of effectiveness of this method can be deduced from Fig. 4. Suppose, for example, that quiet tubes for service on the apparatus rack are to be selected by a test made on the continuous tapper. Choosing the best 25 of the group as tested by the continuous tapper (those plotted above the horizontal line in the figure), it is immediately obvious that when these selected tubes are tested on the apparatus rack (compare abscissae in Fig. 4) the worst tubes are somewhat quieter than some of those in the remaining portion of the group. This is more clearly shown in Fig. 8

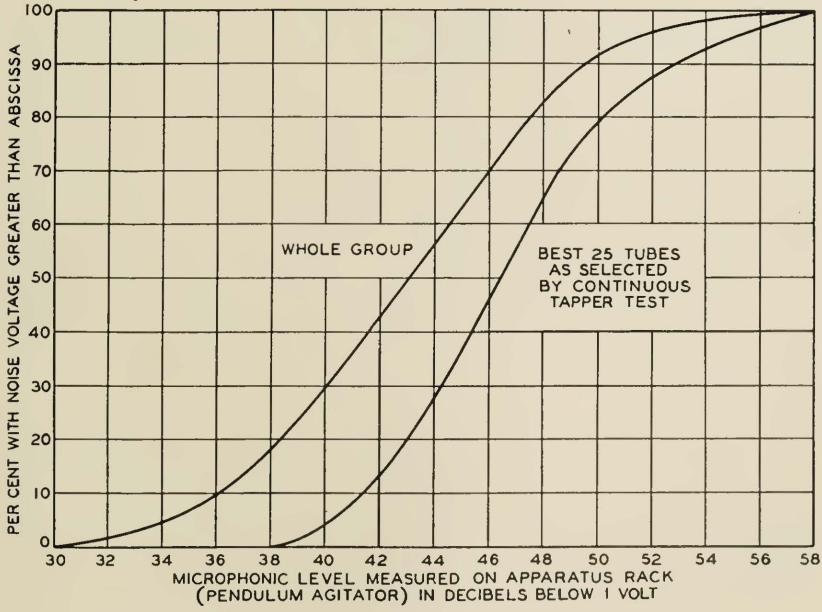


Fig. 8—Effectiveness of selection of quiet tubes.

in which the distribution of levels in this group of 25 tubes is compared with that of the total group as tested on the apparatus rack. The noisiest tubes in the selected group are from 6 to 8 db quieter than the noisiest tubes of the unselected group, and in the selected group, there are none of the tubes which make up the worst 18 per cent of the whole group.

In several commercial situations where microphonic disturbance was at one time troublesome, this type of selection has proved to be of practical value. In these situations, the number of quiet tubes required is only a small percentage of the manufactured output. Furthermore, only a small percentage of the normal output of tubes are found to be prohibitively noisy. Under such circumstances, it is found that when selected tubes from the quietest 25 per cent of the manufacturers stock are used in the positions most sensitive to mechanical shock, the disturbance in these types of equipment either disappears entirely or recedes to such a level that it is no longer troublesome.

Protection from Shock

Where selection of quieter tubes is not feasible or is not sufficiently effective, further reduction of microphonic noise may be achieved by protecting the tube from mechanical and acoustic shock. A very efficient agency for protection from mechanical shock is a well-designed cushion socket. The effectiveness of such a socket depends on its vibrational transmission characteristics considered in relation to the response characteristics of the tubes used. Considerable improvement is usually obtained, however, whatever the combination of tube type and socket type. Figure 9 shows two typical cases of microphonic improvement obtained by using one of several good types of cushion socket which have been tested. The curves drawn in solid lines represent the distributions of microphonic noise levels of a group of No. 102F Tubes tested in one instance in a rigid socket, and in the other in a cushion socket. The mean improvement here due to the cushion socket, is about 30 db. The dotted curves represent similar tests made on a group of No. 262A Tubes and show a mean improvement of about 18 db.

In cases where the noise must be reduced to very low levels, it may not be sufficient to protect the tube from disturbances transmitted mechanically through its base and socket. Except in a perfectly quiet location, there is always some disturbance produced by sound waves impinging directly on the bulb of the tube. Ordinarily this disturbance is negligible, but where the base is sufficiently well cushioned, it may be of controlling importance. It can be reduced only by reducing

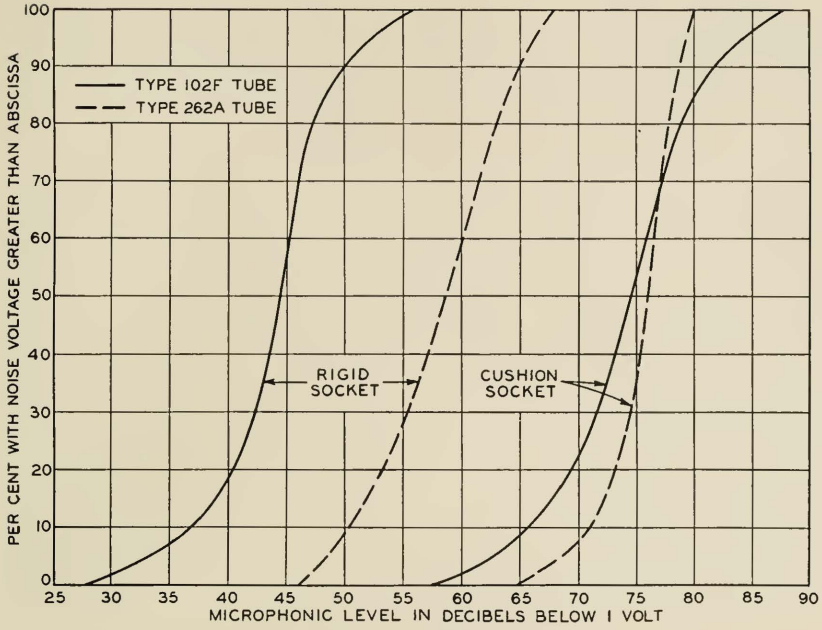


Fig. 9—Effect of cushion socket.

the intensity of the sound wave which is finally allowed to reach the tube, by some such means as enclosing the tube in a heavy, air-tight container.

Reduction of Sputter Noise

The reduction of sputter noise in vacuum tubes is chiefly a problem for the tube manufacturer. Where sputter noise exists in a tube, and exists only with agitation, it is often eliminated by the same cushioning measures which are applied to reduce microphonic noise, but in many cases, satisfactory reduction of sputter would require prohibitive amounts of cushioning. Fortunately, however, the known design features and manufacturing methods, which are now generally applied to tubes of good design, are for the most part quite effective in reducing sputter noise to a negligible level. In the older types of filamentary tubes, for example, sputter noise was often present due to the rattling of the filament at the hook supports at operating temperatures. This source of sputter has been removed in most present day tubes by keeping the filament under tension at all times by means of flexible cantilever spring supports. The effectiveness of this treatment is illustrated in Fig. 10, which shows distributions of sputter noise levels for two

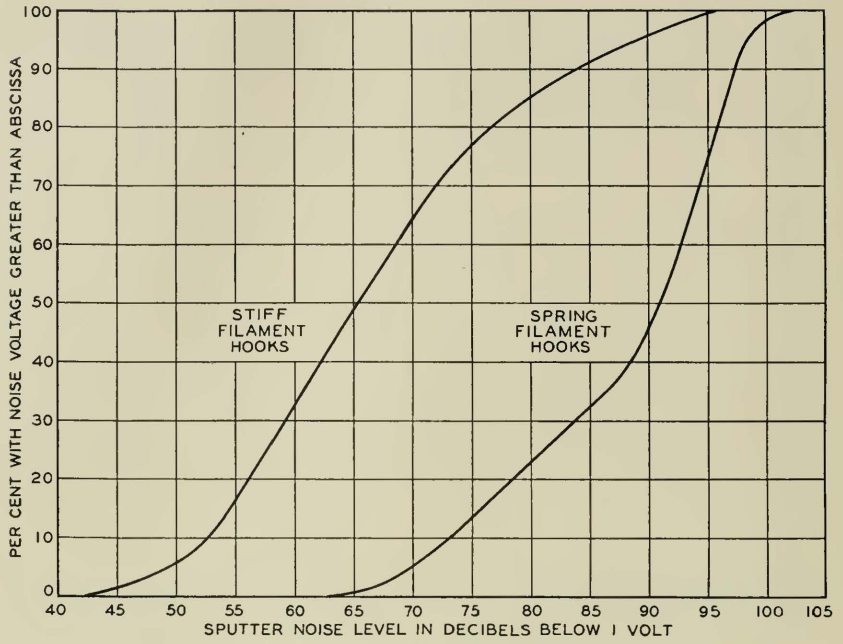


Fig. 10—Effect of filament looseness on sputter noise.

groups of tubes identical in every respect except that one group has spring filament supports while the other has the older rigid supports. In 80 per cent of the tubes, the improvement in the sputter noise is from 20 to 25 db when the spring hook is used.

The source of sputter noise most difficult to control in present day tubes is insulation leaks. These are commonly due to very thin films of conducting material which have been deposited on the surface of the insulating members by sputtering or evaporation during the exhaust or operation of the tube. Experience has shown the conductivity of these films to be intrinsically unstable and discontinuously variable. This condition alone can and does produce sputter noise, but to make matters worse, the metal support wires of the tube are often in only loose contact with the insulating parts and the conducting films covering them so that mechanical agitation breaks and makes the contact and increases the intensity of the noise. The reduction of these insulation leaks is largely a matter of choice of materials, of manufacturing technique to reduce the evaporation of conducting material during exhaust, and of mechanical design to shield important surfaces from contamination during the normal operation of the tube. Great prog-

ress has been made in recent years in effecting an adequate reduction of leaks economically, and in applications where requirements for exceptionally low noise levels warrant slightly increased manufacturing costs, almost any degree of reduction of leaks may be obtained.

Conclusion

The methods which have been outlined for reducing microphonic noise by cushioning and by making use of the quiet tubes which are available are, for the present, adequate to meet all but the most extreme requirements. Should the necessity for further reduction become sufficiently urgent in the future, however, it can probably be obtained either by designing still quieter tubes or by improving the cushioning of sockets. The latter course appears to be the more economical. In either case, however, greatest effectiveness can be attained by considering particular types of tubes and sockets in their relation to one another.

The author is greatly indebted to Drs. M. J. Kelly and H. A. Pidgeon for their kind coöperation and many helpful suggestions in the course of this work.

Fluctuation Noise in Vacuum Tubes *

By G. L. PEARSON

The fluctuation noises originating in vacuum tubes are treated theoretically under the following headings: (1) *thermal agitation* in the internal plate resistance of the tube, (2) *shot effect and flicker effect* from space current in the presence of space charge, (3) *shot effect* from electrons produced by collision ionization and secondary emission, and (4) *space charge fluctuations* due to positive ions. It is shown that thermal agitation in the plate circuit is the most important factor and should fix the noise level in low noise vacuum tubes; shot noise and flicker noise are very small in tubes where complete temperature saturation is approached; shot noise from secondary electrons is negligible under ordinary conditions; and noise from space charge fluctuation due to positive ions is usually responsible for the difference between thermal noise in the plate circuit and total tube noise.

A method is deduced for the *accurate rating of the noise level of tubes* in terms of the input resistance which produces the equivalent thermal noise. *Quantitative noise measurements* by this method are reported on four different types of vacuum tubes which are suitable for use in the initial stage of high gain amplifiers. Under proper operating conditions the noise of these tubes approaches that of thermal agitation in their plate circuits at the higher frequencies and is 0.54 to 2.18 $\times 10^{-16}$ mean square volts per cycle band width in the frequency range from 200 to 15,000 cycles per second. Below 200 cycles per second the noise is somewhat larger.

The minimum noise in different types of vacuum tube circuits is discussed. These include input circuits for high gain amplifiers, ionization chamber and linear amplifier for detecting corpuscular or electromagnetic radiation, and photoelectric cell and linear amplifier for measuring light signals.

With the aid of these results it is possible to design circuits having the maximum signal-to-noise ratio obtainable with the best vacuum tubes now available.

Introduction

IT is well known that the noise inherent in the first stage of a high gain amplifier is a barrier to the amplification of indefinitely small signals. Even when fluctuations in battery voltages, induction, microphonic effects, poor insulation, and other obvious causes are entirely eliminated, there are two sources of noise which remain, namely, thermal agitation of electricity in the circuits and voltage fluctuations arising from conditions within the vacuum tubes of the amplifier. The effect of thermal agitation in circuits outside the vacuum tube is well understood, but in the case of tube noise there is considerable confusion. In order to clarify the whole subject, the present paper analyzes the various sources of noise in vacuum tubes and their attached circuits, points out a new method for the measurement of tube noise, reports the results of such measurements on four different types of vacuum tubes, and discusses the minimum noise in different types of vacuum tube circuits.

* Published in *Physics*, September, 1934.

Often, in the use of high-gain amplifiers, the impedance of the input circuit is naturally high or may effectively be made high by the use of a transformer. In this case the contribution of noise from the vacuum tube is small compared with the noise arising from thermal agitation in the input circuit. This is a desirable condition since it furnishes the largest ratio of signal to noise for a given input power. Sometimes, however, the input impedance is perforce so small that the tube noise may be comparable with or greater than the thermal agitation noise. Such conditions may arise, for example, in amplifiers where the frequency dealt with is high or the frequency range is wide. It is, therefore, desirable to know the noise level to be expected from different types of tubes that may be used in the first stage of high-gain amplifiers as well as to be able to calculate the thermal noise level of the input circuit.

The noise of thermal agitation [1] arises from the fact that the electric charge in a metallic conductor shares the thermal agitation of the molecules of the substance so that minute variations of potential difference are produced between the terminals of the conductor. The mean square potential fluctuation is proportional to the absolute temperature and to the resistive component of the impedance of the conductor, but is independent of the material. The thermal noise power is distributed equally over all frequencies although the apparent magnitude depends on the electrical characteristics of the measuring system as well as on those of the conductor itself. From purely theoretical considerations the following equation has been derived [2] to give the thermal noise voltage at the output of an amplifier due to the thermal agitation of electric charge in an impedance at the input:

$$\overline{E_T^2} = 4kT \int_F R(f) |G_1(f)|^2 df. \tag{1}$$

$\overline{E_T^2}$ is here the mean square thermal noise voltage across the measuring device, k is Boltzmann's constant (1.37×10^{-23} watt second per degree), T the temperature of the impedance expressed in degrees Kelvin, $R(f)$ the resistive component of the impedance at the frequency f, $G_1(f)$ the voltage amplification between the input impedance and the measuring device at the frequency f, and F the frequency band within which the amplification is appreciable.

While the thermal noise in the circuit is accurately predictable, the noise originating within the vacuum tube is not completely under-

[1] J. B. Johnson, *Phys. Rev.*, **32**, 97 (1928).

[2] H. Nyquist, *Phys. Rev.*, **32**, 110 (1928).

stood and cannot be calculated accurately. It is known, however, that tube noise arises from a number of different causes, chief among which are: (1) thermal agitation in the internal plate resistance of the tube, (2) shot effect and flicker effect from space current in the presence of space charge, (3) shot effect from electrons produced by collision ionization and secondary emission, and (4) space charge fluctuations due to positive ions. Each of these sources of noise will be discussed in the following section:

Origin of Noise in Thermionic Amplifier Tubes

Thermal Agitation in the Internal Plate Resistance of the Tube [3]

Just as voltage fluctuations are produced by thermal agitation in resistances comprising the input circuit, so the resistance component of the impedance between plate and cathode is a source of thermal noise. This impedance consists of the internal plate impedance of the vacuum tube in parallel with the external load impedance. Llewellyn [4] has shown that the resistive component of the internal plate impedance produces thermal noise as if it were at the temperature of the cathode. The following formula has been developed by him to cover the case where the tube impedance and load impedance are pure resistances:

$$\overline{E_T^2} = 4k[(r_p r_0)/(r_p + r_0)^2](T_0 r_p + T_f r_0) \int_F |G_2(f)|^2 df. \qquad (2)$$

Here r_p is the internal plate resistance of the tube, r_0 the external load resistance in the plate circuit, $G_2(f)$ the voltage amplification between the load resistance r_0 and the measuring device, and T_0 and T_f respectively the temperatures of the external load resistance and cathode expressed in degrees Kelvin. The relationship between $G_1(f)$ and $G_2(f)$ is given by

$$G_1(f) = G_2(f)(\mu r_0)/(r_0 + r_p), \qquad (3)$$

where μ is the voltage amplification factor of the tube. By assuming $G_2(f)$, in equation (2), to be constant over the frequency range F and substituting for it the value given by equation (3), it is found on integrating that the thermal noise in the plate circuit of the tube produces the same effect in the measuring device as a signal applied

[3] During the preparation of this paper a paper by E. B. Moullin and H. D. Ellis entitled "Spontaneous Background Noise in Amplifiers Due to Thermal Agitation and Shot Effects" appeared in *I. E. E. Jour.*, **74**, 323 (1934). The authors there contend that no thermal noise is produced in the plate impedance of a thermionic vacuum tube and that shot noise is not altered by the presence of space charge. With these contentions I cannot agree and I hope to state my definite reasons therefor at a later date.

[4] F. B. Llewellyn, *Proc. I. R. E.*, **18**, 243 (1930).

to the input circuit whose magnitude at the grid expressed in mean square volts is given by

$$\overline{V^2} = 4kT_0(r_p/\mu)^2[T_f/(T_0 r_p) + 1/r_0]F. \quad (4)$$

Since the noise of thermal agitation is always present, this equation gives the absolute minimum to which fluctuation noise in an amplifying tube can be reduced after all other causes have been eliminated. It shows that for the ideal low noise tube in which thermal noise in the plate circuit is the limiting factor, the noise level may be reduced by a decrease in the cathode temperature, a decrease in the effective frequency band, or by an independent decrease in the plate resistance or increase in the amplification factor. In order to operate at a minimum noise level the tube should work into a load resistance which is large in comparison with $r_p T_0/T_F$. Under this circuit condition the noise level is inversely proportional to μ^2/r_p, a quantity often defined as the "figure of merit" of an amplifying tube.

Shot Effect and Flicker Effect in the Presence of Space Charge

The theory of the shot effect in the absence of space charge has been studied quite completely both theoretically and experimentally by many investigators.[5] The results, however, are not applicable to the study of noise in thermionic vacuum tubes used in high-gain amplifiers, since a high degree of space charge is required in tubes used for this purpose. Llewellyn has extended the theory of the shot effect to cases where partial temperature saturation exists, and obtained a general equation to cover all conditions.[4] This equation reduces to the following form when the load impedance is a pure resistance:

$$\overline{E_S^2} = 2ej(\partial i/\partial j)^2[r_p r_0/(r_p + r_0)]^2 \int_F |G_2(f)|^2 df. \quad (5)$$

$\overline{E_S^2}$ is here the mean square shot voltage across the measuring device, i the total space current, j the total current emitted by the cathode, and e the electronic charge (1.59×10^{-19} coulomb).

A precise experimental verification of this equation is very difficult because of the difficulty in determining $\partial i/\partial j$ accurately. Thatcher,[6] however, has made shot measurements in the presence of space charge ($1 \gtrless \partial i/\partial j \gtrless 0.66$) which verify the theory within the experimental error of the determination of $\partial i/\partial j$.

[5] W. Schottky, *Ann. d. Physik*, **57**, 541 (1918); T. C. Fry, *Jour. Franklin Inst.*, **199**, 203 (1925); A. W. Hull and N. H. Williams, *Phys. Rev.*, **25**, 147 (1925).

[6] Everett W. Thatcher, *Phys. Rev.*, **40**, 114 (1932).

Equation (5) shows that as long as space charge is too small to affect the flow of current, that is when i is equal to j, the mean square shot voltage is directly proportional to the space current. As emission is increased, however, space charge begins to control and finally limits the space current so that the value of $\partial i/\partial j$ approaches zero. Thus the shot voltage increases less rapidly as space charge becomes effective and then finally decreases rapidly toward zero as complete space charge control is reached.

Experimental curves showing the effect of space charge on tube noise are shown in Fig. 1 where abscissæ represent space current in

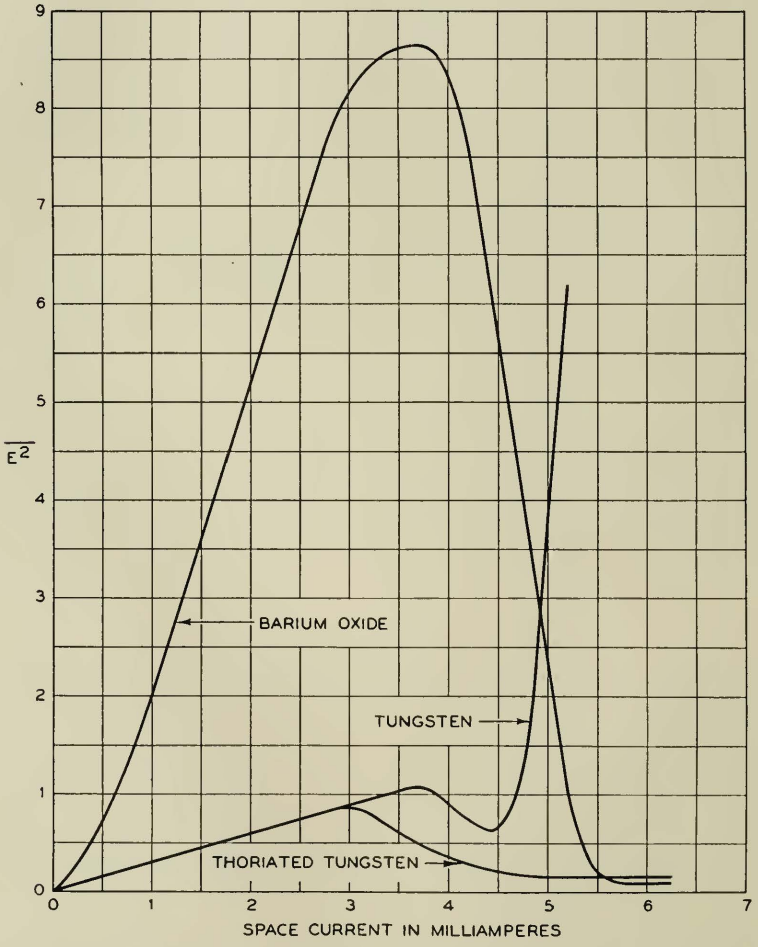


Fig. 1—The effect of space charge on fluctuation noise. Three tubes having filaments composed of tungsten, thoriated tungsten, and barium oxide. $\overline{E^2}$ is the mean square noise voltage across the output measuring device expressed in arbitrary units. The variation in space current was obtained by changing the cathode temperature, the plate voltage remaining constant.

milliamperes, and ordinates represent mean square noise voltage across the output measuring device expressed in arbitrary units. The change in space current was obtained by varying the filament heating current while the plate voltage remained constant. Tubes having thoriated tungsten, tungsten, and barium oxide cathodes were used.

At low space currents where no space charge is present the thoriated tungsten and tungsten filaments each give a pure shot effect, the mean square voltage increasing linearly with the space current. As the space current is increased further and space charge sets in, the shot voltage in each tube goes through a maximum and decreases with oncoming temperature saturation as suggested by equation (5). With the approach of complete temperature saturation the noise, however, does not decrease to zero in accordance with this equation. If it were possible to reach complete temperature saturation the residual noise would not be due to the shot effect, but rather to thermal noise in the plate circuit of the tube, positive ions and secondary emission within the tube, and other contributing causes. Usually this condition is approached in the better commercial tubes so that the contribution of true shot noise is a small part of the total noise.

If the methods used in obtaining equation (4) are applied to equation (5), it is found that the shot noise in the plate circuit of the tube produces the same effect in the output measuring device as a signal applied to the input circuit whose magnitude at the grid expressed in mean square volts is

$$\overline{V^2} = 2ej(\partial i/\partial j)^2(r_p/\mu)^2F. \tag{6}$$

This equation shows that the level of shot noise at the input is lowered by an increase in the cathode temperature, which increases the degree of temperature saturation, and by an increase in the ratio μ/r_p, which by definition is the transconductance of the tube, but is independent of the external load resistance. It should be remembered, however, that shot noise in the plate circuit should not fix the noise level in low noise vacuum tubes and that never, as is sometimes done in the literature, can the noise of an amplifier be calculated as pure shot noise in the plate circuit, for in the absence of space charge the tube would not be an amplifier.

Although space charge can counteract the effect of random electron emission from the cathode so that shot noise is reduced, other factors can alter the flow of current in such a way that the noise is increased. This is the case when changes in emission occur over small areas of the cathode, giving rise to an additional fluctuation which has been termed flicker effect.[7] This type of noise is particularly noticeable

[7] J. B. Johnson, *Phys. Rev.*, **26**, 71 (1925); W. Schottky, *Phys. Rev.*, **28**, 74 (1926).

with oxide coated cathodes. Since the flicker effect is due to localized variations in the emission of the cathode, one would expect it to disappear in the presence of a complete space charge condition.

The experimental curve for the barium oxide coated filament, Fig. 1, shows a flicker effect many times larger than the shot effect on which it is superimposed. At low space currents the mean square flicker effect voltage increases faster than the pure shot noise, a square law rather than a linear relationship being followed. As space charge sets in, the flicker effect voltage goes through a maximum and then decreases with increased space current in the same manner as does the shot effect voltage. In spite of the large flicker effect, as complete temperature saturation is approached the total noise is even less than that found with the thoriated tungsten filament which has no flicker effect. This illustrates clearly the effectiveness of space charge in smoothing the space current.

When the control grid of a vacuum tube is floating at its equilibrium potential, the noise level is much higher than when the grid is connected through an input circuit to the cathode. This increase in noise is primarily due to thermal noise in the extremely high input resistance of the tube and to shot noise arising from small grid currents.[8] The magnitude of the thermal noise may be calculated, knowing that the input impedance of the tube consists of its input resistance, r_g, in parallel with its dynamic grid-to-ground capacitance. In such a combination the real resistance component, $R(f)$, is related to the pure resistance, r_g, and the dynamic capacitance, c, according to the equation

$$R(f) = r_g/(1 + 4\pi^2 c^2 r_g^2 f^2). \tag{7}$$

According to equation (1) the mean square thermal noise input voltage is then

$$\overline{V_T^2} = 4kTr_g \int_F df/(1 + 4\pi^2 c^2 r_g^2 f^2). \tag{8}$$

With the grid floating at its equilibrium position (usually slightly negative with respect to the cathode) the grid current is composed of two components equal in magnitude but opposite in sign. The one component consists of electrons reaching the grid, while the other consists of positive ions reaching and electrons leaving the grid. The electrons are liberated from the grid by secondary emission, the photoelectric effect, thermionic emission, and soft X-rays. It should be pointed out that space charge does not reduce the noise produced by

[8] L. R. Hafstad, *Phys. Rev.*, **44**, 201 (1933).

the shot effect in any of these currents. The general shot effect equations [9] show that the magnitude of shot noise from these grid currents is

$$\overline{V_S^2} = 2ei_g r_g^2 \int_F df/(1 + 4\pi^2 c^2 r_g^2 f^2)^2, \qquad (9)$$

where i_g is the sum of the grid currents regardless of sign.

Noise Produced by Secondary Effects

In this classification are grouped several sources of disturbance whose individual effects are very difficult to calculate and measure under the operating conditions of the vacuum tube. For this reason the following discussion will include only a general consideration of the more obvious contributing causes.

Although the cathode is the principal source of electrons which reach the plate, in actual practice electrons are produced by ionization of the gas molecules within the tube or by secondary emission resulting from bombardment of the tube elements. Electrons produced in this manner are drawn to the plate and generate noise which is not much affected by the space charge. Assuming a reasonable magnitude for the current produced in this manner it can be shown by the shot equations that noise from this source is usually negligible. In cases where the gas pressure within a tube is above normal, or in screen-grid and multi-grid tubes having high plate resistances and considerable secondary emission, the shot noise from secondary and ionization electrons may be of the same order of magnitude as thermal noise in the plate circuit.

Positive ions formed from ionized gas molecules or emitted from the tube elements are much more effective in producing noise since, instead of being drawn off to the plate, they are attracted into the space charge region where small disturbances in equilibrium produce large momentary fluctuations in space current. Due to their large mass the motions of the ions are relatively slow, so that they are very effective in this respect. This type of noise is quite disturbing in amplifying tubes for it tends to become a maximum at complete temperature saturation. This is illustrated very clearly in the noise measurements on the tungsten filament shown in Fig. 1. Here positive ions from the filament begin to show their effect as space charge sets in, the number of ions and the amount of noise increasing as temperature saturation is approached. As heard in the loud speaker, this noise consists of sharp crackling sounds which can easily be dis-

[9] E.g. Ref. 4 or 5.

tinguished from the steady rustling noise of the shot and thermal effects.

Ballantine [10] has recently made calculations and measurements on the noise due to positive ions from collision ionization in which he has shown that the mean square noise voltage is roughly proportional to the gas pressure within the tube and to the 3/2 power of the plate current. Comparing his results with equation (2), it appears that under ordinary working conditions the noise due to collision ionization in a vacuum tube may be of the same order of magnitude as noise from thermal agitation in its plate circuit. The noise level of tubes having a poor vacuum, however, may be much higher.

Measurement of Tube Noise

The performance, as regards freedom from noise, of a vacuum tube used in an amplifier may be indicated by a comparison between the noise and a signal applied to the grid. Usually we say that the noise is equivalent to a signal which gives the same power dissipation in the output measuring instrument as the noise, the frequency of the signal being suitably chosen with respect to the frequency characteristics of the amplifier. Since tube noise is distributed over all frequencies and the noise power increases with the effective band width, it will be advantageous to express this input signal in equivalent mean square volts per unit frequency band width, effective over a given frequency range.

From these considerations it can be seen that the most convenient standard signal for measuring the equivalent input noise over any given frequency range is one in which the mean square signal voltage is distributed equally over all frequencies. With such a signal the equivalent input noise over any frequency range can be measured directly, while if an oscillator is used a number of measurements are required and the result must be computed by graphical integration. A signal which meets these frequency requirements perfectly is the noise of thermal agitation. Accordingly, in the measurements to be described here the standard input signal will be the thermal agitation voltage of a resistance R, connected between the control grid and cathode of the tube under test.[1]

The thermal noise voltage of the grid circuit, referred to the output measuring device, is given by equation (1), where $R(f)$ is the real resistance component of an input impedance consisting of the pure resistance R in parallel with its shunt capacity and that of its leads

[10] Stuart Ballantine, *Physics*, **4**, 294 (1933).

and of the vacuum tube. In such a combination $R(f)$ is related to the pure resistance R and the total capacitance c according to equation (7). In all the measurements described here the factor $4\pi^2c^2R^2f^2$ is so small in comparison with unity that it may be neglected without appreciable error. Under these conditions equation (1) reduces to

$$\overline{E_T^2} = 4kTR \int_F |G_1(f)|^2 df, \tag{10}$$

where R is the direct current value of the resistance between control grid and cathode of the tube under test.

The voltage fluctuations arising from conditions within the tube produce a mean square voltage output $\overline{E_N^2}$ according to the equation

$$\overline{E_N^2} = \int_F |V(f)|^2 |G_1(f)|^2 df, \tag{11}$$

where $|V(f)|^2$ is the tube noise at the frequency f for unit frequency band width, expressed in volts squared and referred to the input circuit. Letting V_F^2 be the effective value of $|V(f)|^2$ over the band width of the amplifier we obtain

$$\overline{E_N^2} = V_F^2 \int_F |G_1(f)|^2 df. \tag{12}$$

Since the integrals in equations (10) and (12) are identical it is found on dividing one equation by the other and solving for V_F^2 that:

$$V_F^2 = 4kTR(\overline{E_N^2}/\overline{E_T^2}). \tag{13}$$

Equation (13) enables one to calculate the magnitude of tube noise in the frequency range F, per unit cycle band width, in terms of the thermal noise generated in a resistance R placed in the input circuit.[11] Since this equation contains no integral the measurements are simplified in that neither standard signal generator nor calibrated amplifier is required.

Apparatus

The experimental arrangement used in the measurements to be reported here is given in schematic form in Fig. 2. The system includes the tube under test, a high gain amplifier, appropriate filters, an attenuator, and an output measuring device.

[11] It is assumed that tube noise does not vary with frequency, or that the band width of the amplifier is so narrow that no appreciable error is introduced in applying the result.

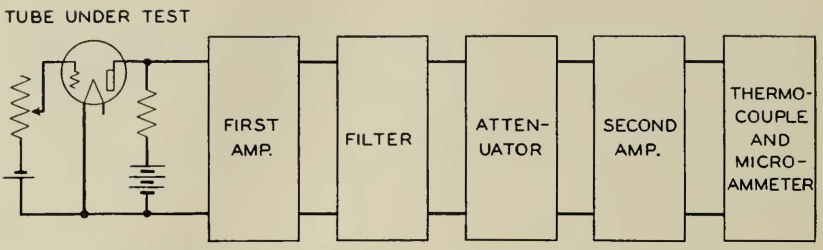


Fig. 2—Schematic amplifier circuit for measuring fluctuation noise in vacuum tubes.

The input circuit consists of the tube under test together with the variable grid resistor, external load resistor, and batteries for furnishing the required filament, grid, and plate voltages. Because of the high value of amplification required and the wide frequency range covered by the amplifier, this circuit required shielding from external disturbances arising from electrical, mechanical, and acoustical shock. Accordingly the tube under test was suspended by means of rubber bands, the whole circuit with the exception of batteries placed inside a tightly sealed lead lined box, and this box in turn suspended by means of a system of damped springs. The box with its cover removed and the tube in place is shown in Fig. 3. This shielding was sufficient to reduce the noise from outside disturbances to such a low level that no correction had to be made for it at any time.

The high gain amplifier [12] consists of two separate resistance coupled units each containing three stages. Each unit is so designed and shielded that the effect of external disturbances is eliminated. The total gain obtainable is about 165 db (constant to within 2 db from 10 cycles to 15,000 cycles). Since this gain is in excess of that required for the study of thermal and tube noises, an attenuator having a range of 63 db was inserted between the two units. In order to limit amplification to certain desired frequency bands, specially designed electric filters were inserted between the first amplifier unit and the attenuator. Three such filters were used of which one is a low-pass filter with cut-off around 205 cycles, and the other two are band-pass filters with mid-frequencies at 1750 and 11,000 cycles respectively. The frequency characteristic of the amplifier with no filter and with each filter inserted is shown in Fig. 4.

The recording instrument is a 600-ohm vacuum thermocouple and microammeter. Conveniently, the deflection of the microammeter is closely proportional to the mean square voltage applied to the couple. The procedure in making a measurement of tube noise is as follows:

[12] The essential parts of this amplifier were designed by Mr. E. T. Burton.

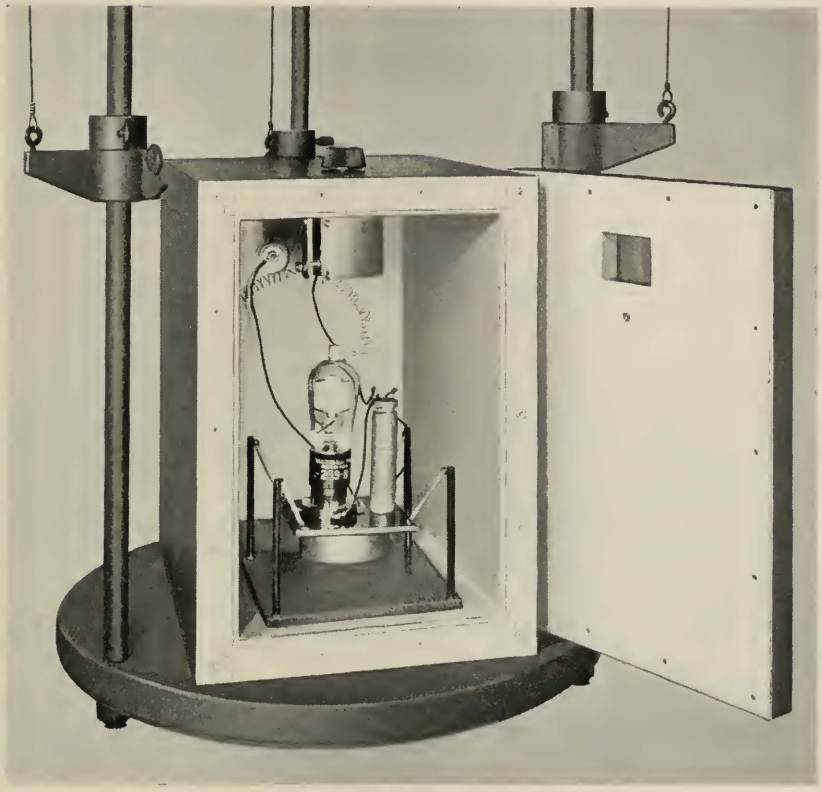


Fig. 3—Tube under test mounted in the shielding box

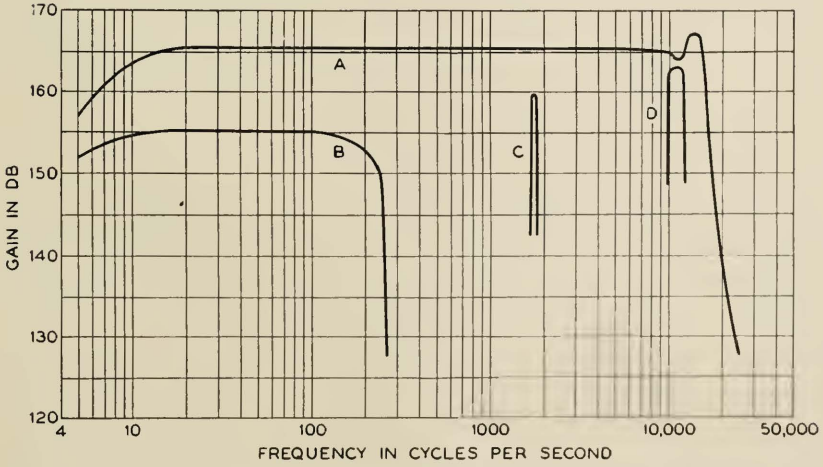


Fig. 4—Frequency characteristic of amplifier circuit. Curve *A* with no filter, Curve *B* with low pass filter, and Curves *C* and *D* with band pass filters.

With the tube under test operating at zero grid resistance, the attenuator is adjusted to give a convenient deflection of the microammeter (due to noise in the tube under test). Grid resistance is now added until this deflection is exactly doubled, thus making $\overline{E_N^2}$ equal to $\overline{E_T^2}$. This value of input resistance, designated by R_G, is a measure of the inherent noise of the tube. Substituting R_G in equation (13) the tube noise is calculated from the relation

$$V_F^2 = 4kTR_G = 1.64 \times 10^{-20} R_G \text{ volt}^2, \tag{14}$$

where R_G is expressed in ohms and T is 300° K. (approximate room temperature).

Noise in Certain Vacuum Tubes[13]

Quantitative measurements of tube noise were made on four different types of standard Western Electric vacuum tubes, namely: Nos. 102G, 264B, 262A and 259B. These tubes have as low a noise as any tube obtainable at the present time.

In order to obtain the best signal to noise ratios it was found that operating conditions different from those normally recommended must be used. In general, the cathode must be operated at as high a temperature as possible without impairing the life of the tube, the negative bias of the control grid must be reduced to as near zero as possible without causing excessive grid current, and the plate voltage must be reduced below the value normally recommended. In all the measurements described here the tube under test was coupled to the first amplifier unit through a 50,000-ohm load resistance. It was found that the signal-to-noise ratio could be improved a fraction of a db by increasing the load resistance (in accordance with equation (4)); this, however, necessitated a large plate voltage which was inconvenient. Six tubes of each type were tested and the noise data given below were obtained by averaging the six measurements for each type. Individual tubes may differ from these average values by as much as ± 1 db.

No. 102G Tube

This is a three-element, filament-type tube. Its long life, exceptionally high stability of operation, and good temperature saturation make it a desirable tube to use in the input stage of certain high-gain amplifiers. This tube also has a comparatively small microphonic response to mechanical and acoustical shock although it is not as good as the No. 262A and the No. 264B tubes in this respect.[14]

[13] Noise in other types of vacuum tubes has been reported by G. F. Metcalf and T. M. Dickinson, *Physics, 3,* 11 (1932); E. A. Johnson and C. Neitzert, *Rev. Sc. Inst., 5,* 196 (1934); E. B. Moullin and H. D. M. Ellis, *I. E. E. Jour., 74,* 323 (1934); W. Brentzinger and H. Viehmann, *Arch. f. Hochfr. und Elektroauk, 39,* 199 (1932).

[14] The microphonic response of several types of Western Electric vacuum tubes to mechanical agitation is reported by D. B. Penick in this issue of the *Bell Sys. Tech. Jour.*

The conditions found most suitable for quiet operation of the No. 102G tube and the corresponding average tube characteristics are given in the first two columns of Table I. Under these conditions the

TABLE I

WESTERN ELECTRIC NO. 102G TUBE

Operating Conditions	Tube Characteristics	Noise Data		
		Frequency Range Cycles per Sec.	V_F^2 Volt²	R_G Ohms
Filament Voltage, 2.0 volts Current, 1.0 ampere	Type of Tube, 3 Element Type of Cathode, Oxide Coated Filament			
Grid Voltage, − 0.5 volt		10–15,000	0.64×10^{-16}	3,900
Plate Voltage, 130 volts Current, 1.2 milliamperes	Amplification Factor, 30	5–205	2.2	13,600
	Plate Resistance, 45,000 ohms	1,750–1,850	0.58	3,550
Load Resistance, 50,000 ohms	Approx. Dynamic Input Capacitance, 80 $\mu\mu$f	10,000–12,000	0.54	3,300

average equivalent tube noise voltage, referred to the grid circuit, is given in the last column of the same table. These noise data are given in terms of R_G, the experimentally determined equivalent noise resistance of the tube, and in terms of V_F^2, calculated by means of equation (14), for each of the four frequency ranges shown in Fig. 4.

The No. 102G has the lowest noise of all the tubes tested and was found suitable for use in the first stage of high-gain amplifiers where tube noise is the limiting factor, provided it is not required that the input capacitance and microphonic response to mechanical and acoustical shock be extremely low.

No. 264B Tube

This is a three-element filament-type tube. Due to the rigid construction and the short filament which is designed to reduce vibration to a minimum, the microphonic response of the tube to mechanical and acoustical shock is exceptionally low.[15] The extensive system of spring suspensions and the heavy sound-proof chamber usually required for shielding low noise tubes may be simplified when using the No. 264B. In addition, this tube has good temperature saturation, low power consumption, and high stability of operation.

The operating conditions and noise data for this tube are given in Table II. Although the noise of this tube is slightly higher than that

[15] M. J. Kelly, *S. M. P. E. Jour.*, **18**, 761 (1932).

TABLE II

WESTERN ELECTRIC NO. 264B TUBE

Operating Conditions	Tube Characteristics	Noise Data		
		Frequency Range Cycles per Sec.	V_F^2 Volt²	R_G Ohms
Filament Voltage, 1.5 volts Current, .30 ampere	Type of Tube, 3 Element Type of Cathode, Oxide Coated Filament			
Grid Voltage, − 0.5 volt	Amplification Factor, 7	10–15,000 5–205	1.3×10^{-16} 6.6	7,650 40,000
Plate Voltage, 26 volts Current, 0.6 milliampere	Plate Resistance, 18,500 ohms	1,750–1,850	1.1	6,800
Load Resistance, 50,000 ohms	Approx. Dynamic Input Capacitance, 30 $\mu\mu$f	10,000–12,000	1.0	6,200

of the No. 102G, the lower microphonic response and the lower power consumption make it a more desirable tube to use in input stages of certain high gain amplifiers.

No. 262A Tube

This is a three-element tube having an indirectly heated cathode. It is designed to give a microphonic response to mechanical and acoustical shock [15] still lower than that of the 264B. Except for frequencies below 200 cycles per second it was found that no acoustic shield was necessary for this tube even when working at extremely low levels. Although this tube is designed to have a low hum disturbance resulting from alternating current for heating the cathode (the interference from this effect can be held to less than 7×10^{-6} equivalent input volt), direct current power was used in the measurements here described.

The operating conditions and noise data for the No. 262A tube are given in Table III.

No. 259B Tube

This is a four-element, screen-grid tube having an indirectly heated cathode. Its comparatively high amplification factor makes possible a relatively large gain per stage so that when it is used in the first stage of a high-gain amplifier succeeding stages contribute nothing to the total noise.

Noise measurements on the No. 259B tube show that the signal-to-noise ratio is approximately independent of the plate voltage over a

TABLE III

WESTERN ELECTRIC NO. 262A TUBE

Operating Conditions	Tube Characteristics	Noise Data		
		Frequency Range Cycles per Sec.	V_F^2 Volt²	R_G Ohms
Heater Voltage, 10 volts Current, 0.32 ampere	Type of Tube, 3 Element Type of Cathode, Oxide Coated, Indirectly Heated.....	10–15,000	1.3×10^{-16}	7,700
Grid Voltage, − 1.0 volt	Amplification Factor, 15.7.....	5–205	17.	100,000
Plate Voltage, 44 volts Current, 1.0 milliampere	Plate Resistance, 22,000 ohms.....	1,750–1,850	1.0	6,400
Load Resistance, 50,000 ohms	Approx. Dynamic Input Capacitance, 23 $\mu\mu$f.....	10,000–12,000	0.84	5,100

wide operating range, but is closely dependent on the plate current as affected by the control and screen grid voltages. Table IV contains the operating conditions and noise data for this tube.

Noise measurements were also made on the No. 259B tube with its control grid floating at equilibrium potential. Using the operating voltages specified above, the noise level was about 20 db higher than those given in Table IV. The level can be greatly reduced by operating the tube at a lower cathode temperature and with lower screen

TABLE IV

WESTERN ELECTRIC NO. 259B TUBE

Operating Conditions	Tube Characteristics	Noise Data		
		Frequency Range Cycles per Sec.	V_F^2 Volt²	R_G Ohms
Heater Voltage, 2.0 volts Current, 1.7 amperes	Type of Tube, 4 Element Screen Grid Type of Cathode, Oxide Coated, Indirectly Heated.....	10–15,000	3.2×10^{-16}	19,800
Grid Control Voltage, − 1.5 volts	Amplification Factor, 1,500.....	5–205	7.7	47,000
Screen Voltage, 22.5 volts	Plate Resistance, 2.75 megohms.....	1,750–1,850	2.8	17,100
Plate Voltage, 100 volts Current, 0.6 milliampere	Approx. Dynamic Input Capacitance, 6.0 $\mu\mu$f.....	10,000–12,000	2.8	17,000
Load Resistance, 50,000 ohms				

and plate voltages.[16] This reduction in noise is due to a decrease in current to the floating grid. Using a heater current of 1.3 amperes, a plate current of 0.1 milliampere, a screen potential of 16.5 volts and a plate potential of 30 volts the equivalent input noise was 1.4×10^{-5} volt for the entire frequency range from 10 cycles to 15,000 cycles. Under these operating conditions the floating grid potential was 1.0 volt negative with respect to the cathode, the input resistance 1.4×10^{10} ohms, the dynamic grid-to-cathode capacitance 6×10^{-12} farad, and each component of grid current about 4.5×10^{-12} ampere.

Discussion of Results

From the noise data in the preceding tables one can estimate quite accurately the equivalent input noise voltage of each of the four types of tubes at any frequency between 5 and 15,000 cycles, and for any band width within these limits. For example, using the noise data given in Table I the equivalent input noise voltage of the No. 102G tube working over a band having sharp cut-offs at 5 cycles and 205 cycles is computed to be

$$(\overline{V^2})^{1/2} = (V_F^2 F)^{1/2} = 2.1 \times 10^{-7} \text{ volt.} \qquad (15)$$

For a band width of 200 cycles with mid-frequency at 10,000 cycles this noise is reduced to 1.0×10^{-7} volt. It can be seen that for each type of tube the noise voltage over equal band widths is between 1.5 and 4.5 times greater at frequencies below 200 cycles than at the higher frequencies.[17]

Even at high frequencies the noise voltage is above that expected from thermal noise in the plate circuit which, as stated above, is the absolute minimum to which fluctuation noise in a thermionic vacuum tube may be reduced after all other causes are eliminated. In the case of the No. 102G tubes for instance, using the operating conditions of Table I, and assuming 1100° K. as the temperature of the barium oxide filament, it is found by means of equation (4) that the equivalent input noise voltage produced by thermal agitation in the plate circuit is 2.7×10^{-8} volt for a band width of 200 cycles. The total input noise voltage obtained experimentally at the higher frequencies is greater than this by a factor of 3.8. In like manner it is found that the total input noise voltages found experimentally for the Nos. 264B, 262A and 259B tubes are greater than the equivalent input thermal

[16] I am indebted to Dr. J. R. Dunning of Columbia University for pointing out this fact.

[17] Other investigators have also found an increase in tube noise energy at the lower frequencies. G. F. Metcalf and T. M. Dickinson, *Physics*, 3, 11 (1932).

noise voltages produced in the plate circuit by factors 2.1, 3.7, and 16 respectively. These calculations show that each of these four types of tubes approaches the requirements of an ideal low noise amplifying tube although none of them is perfect in this respect.

As stated above, the best signal-to-noise ratio in a high-gain amplifier is obtained when thermal agitation in the input resistance is responsible for most of the noise in the amplifier. This condition is met when the resistance of the input circuit is higher than the value of R_G for the input tube. In case the resistance in the input circuit is less than R_G the input signal and the thermal noise from the input circuit can be raised above the noise of the tube by using an input transformer having a sufficiently high voltage step-up. The voltage ratio of the transformer, and in turn the possible ratio of input circuit thermal noise to tube noise, is limited, especially at the higher frequencies, by the dynamic grid-to-ground capacitance of the input tube and its leads. In such a circuit the No. 259B tube with its lower inter-electrode capacities and higher tube noise is often more desirable than even the quietest three-element tubes.

In those high-gain amplifiers in which unavoidably the resistance of the input circuit is low, the tube rather than thermal agitation in the grid circuit is responsible for most of the noise. Here the best signal-to-noise ratio can be obtained by choosing a tube for the initial stage having the lowest possible noise level. The above measurements show that one of the three-element tubes, particularly the No. 102G tube if sufficient shielding is used, is best suited for this purpose.

The lower limits of noise obtainable with high gain amplifiers may be estimated by means of Fig. 5, which shows the noise as a function of input resistance and frequency band width when thermal agitation in the input circuit is responsible for all the noise. The data for this figure are obtained from the thermal noise relationship

$$\overline{V_T^2} = 1.64 \times 10^{-20} RF \text{ volt}^2. \tag{16}$$

R is expressed in ohms and the temperature has been taken at 300° K., which is approximately room temperature. It must be remembered that the attainment of these noise levels at low input resistances is limited by the input transformer.

The results of the noise measurements on the No. 259B tube with floating grid may be compared with the value predicted by equations (8) and (9). Inserting the tube characteristics obtained by experiment ($r_g = 1.4 \times 10^{10}$ ohms, $i_g = 9 \times 10^{-12}$ ampere, and $c = 6 \times 10^{-12}$ farad), and integrating between the frequency limits 10 cycles

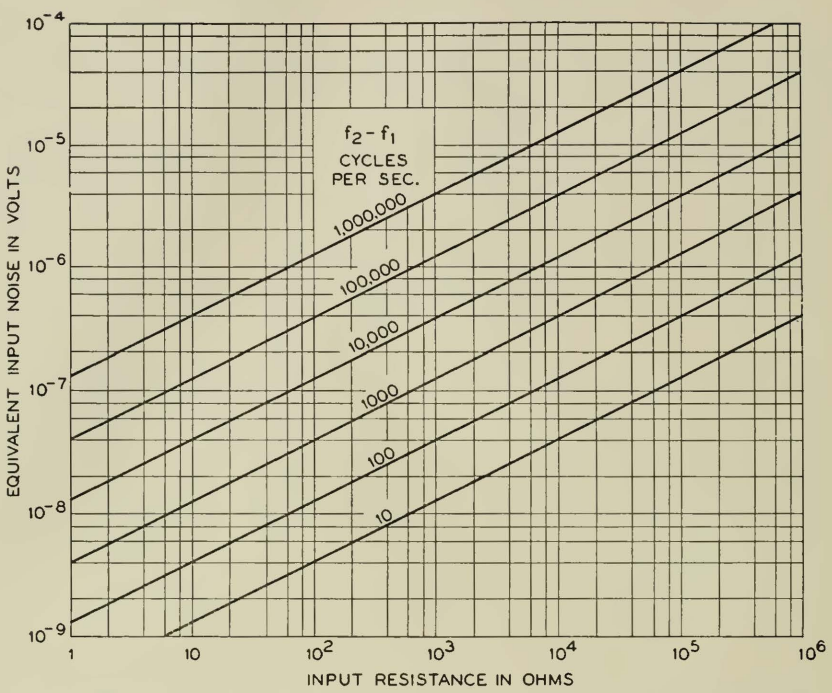


Fig. 5—Thermal noise level as a function of input resistance and frequency range.

and 15,000 cycles, it is found that the equivalent thermal noise input is 0.9×10^{-5} volt, while the shot noise input is 1.4×10^{-5} volt. The total noise is the square root of the sum of the squares of these values or 1.7×10^{-5} volt. This agrees with the measured value of 1.4×10^{-5} volt within an error of 20 per cent, which is as accurate as the determination of the grid currents. These equations may also be used to calculate the noise originating in the grid circuit when external resistance or capacitance is connected between grid and cathode, provided r_g and c are now calculated from the internal and external impedances in parallel.

A common method of detecting corpuscular or electromagnetic radiation makes use of an ionization chamber and linear amplifier. In this circuit the control grid in the first tube of the amplifier is connected to the collecting electrode of the ionization chamber and both allowed to float at equilibrium potential.[18] The shot and thermal noise in this grid circuit sets a limit to the measurement of extremely weak radiation. Knowing the value of input capacitance, input

[18] H. Greinacher, *Zeits. f. Physik*, **36**, 364 (1926).

resistance, floating grid current, and the frequency limits of the amplifier, equations (8) and (9) may be used to calculate this limiting noise level. For example, if one uses a No. 259B tube with the operating voltages specified for floating grid, an ionization chamber having a capacitance of 15×10^{-12} farad, and an amplifier having a frequency range from 200 to 5000 cycles per second the limiting noise level is 1×10^{-6} root mean square volt.

The limiting noise level in a system consisting of a photoelectric cell and thermionic amplifier is determined by thermal agitation in the coupling circuit between the photoelectric cell and amplifier, and by shot noise in the photoelectric current (in circuits where the photoelectric current is very small and the coupling resistance is very high, shot noise from grid current in the vacuum tube becomes appreciable). The noise of thermal agitation may be calculated by means of equation (8) provided r_g is now replaced by R, the coupling resistance. If vacuum cells are used, the photoelectric current produces a pure shot noise which can be calculated by equation (9) provided i_g is replaced by I, the photoelectric current. In gas filled photocells where collision ionization occurs, the noise is in excess of the value calculated in this manner.[19] The relative magnitude of shot noise and thermal noise depends on the values of I and R, and by combining equations (8) and (9) it is found that

$$\overline{V_S^2}/\overline{V_T^2} = eIR/2kT = 19.4IR, \tag{17}$$

where I is expressed in amperes, R in ohms, and T is 300° K. Thus an increase in either I or R will tend to make shot noise exceed thermal noise. This is the desirable condition since it furnishes the largest ratio of signal-to-noise for a given light signal on the photoelectric cell.

In conclusion I wish to acknowledge my indebtedness to Dr. J. B. Johnson for the helpful criticism he has given during the course of this work.

[19] B. A. Kingsbury, *Phys. Rev.*, **38**, 1458 (1931).

Systems for Wide-Band Transmission Over Coaxial Lines

By L. ESPENSCHIED and M. E. STRIEBY

In this paper systems are described whereby frequency band widths of the order of 1000 kc. or more may be transmitted for long distances over coaxial lines and utilized for purposes of multiplex telephony or television. A coaxial line is a metal tube surrounding a central conductor and separated from it by insulating supports.

IT appears from recent development work that under some conditions it will be economically advantageous to make use of considerably wider frequency ranges for telephone and telegraph transmission than are now in use [1, 2] or than are covered in the recent paper on carrier in cable.[3] Furthermore, the possibilities of television have come into active consideration and it is realized that a band of the order of one million cycles or more in width would be essential for television of reasonably high definition if that art were to come into practical use.[4, 5]

This paper describes certain apparatus and structures which have been developed to employ such wide frequency ranges. The future commercial application of these systems will depend upon a great many factors, including the demand for additional large groups of communication facilities or of facilities for television. Their practical introduction is, therefore, not immediately contemplated and, in any event, will necessarily be a very gradual process.

Types of High-Frequency Circuits

The existing types of wire circuits can be worked to frequencies of tens of thousands of cycles, as is evidenced by the widespread application of carrier systems to the open-wire telephone plant and by the development of carrier systems for telephone cable circuits.[2, 3] Further development may lead to the operation of still higher frequencies over the existing types of plant. However, for protection against external interference these circuits rely upon balance, and as the frequency band is widened, it becomes more and more difficult to maintain a sufficiently high degree of balance. The balance requirements may be made less severe by using an individual shield around

[1] For references, see end of paper.

* Published in *Electrical Engineering*, October, 1934. Scheduled for presentation at Winter Convention of A. I. E. E., New York, N. Y., January, 1935.

each circuit, and with sufficient shielding balance may be entirely dispensed with.

A form of circuit which differs from existing types in that it is unbalanced (one of the conductors being grounded), is the coaxial or concentric circuit. This consists essentially of an outer conducting tube which envelops a centrally-disposed conductor. The high-frequency transmission circuit is formed between the inner surface of the outer conductor and the outer surface of the inner conductor. Unduly large losses at the higher frequencies are prevented by the nature of the construction, the inner conductor being so supported within the tube that the intervening dielectric is largely gaseous, the separation between the conductors being substantial, and the outer conductor presenting a relatively large surface. By virtue of skin effect, the outer tube serves both as a conductor and a shield, the desired currents concentrating on its inner surface and the undesired interfering currents on the outer surface. Thus, the same skin effect which increases the losses within the conductors provides the shielding which protects the transmission path from outside influences, this protection being more effective the higher the frequency.

The system which this paper outlines has been based primarily upon the use of the coaxial line. The repeater and terminal apparatus described, however, are generally applicable to any type of line, either balanced or unbalanced, which is capable of transmitting the frequency range desired.

The Coaxial System

A general picture of the type of wide band transmission system which is to be discussed is briefly as follows: A coaxial line about 1/2 inch in outside diameter is used to transmit a frequency band of about 1,000,000 cycles, with repeaters capable of handling the entire band placed at intervals of about 10 miles. Terminal apparatus may be provided which will enable this band either to be subdivided into more than 200 telephone circuits or to be used *en bloc* for television.

Such a wide-band system is illustrated in Fig. 1. It is shown to comprise several portions, namely, the line sections, the repeaters, and the terminal apparatus, the latter being indicated in this case as for multiplex telephony. Two-way operation is secured by using two lines, one for either direction. It would be possible, however, to divide the frequency band and use different parts for transmission in opposite directions.

A form of flexible line which has been found convenient in the experimental work is illustrated in Fig. 2 and will be described more fully

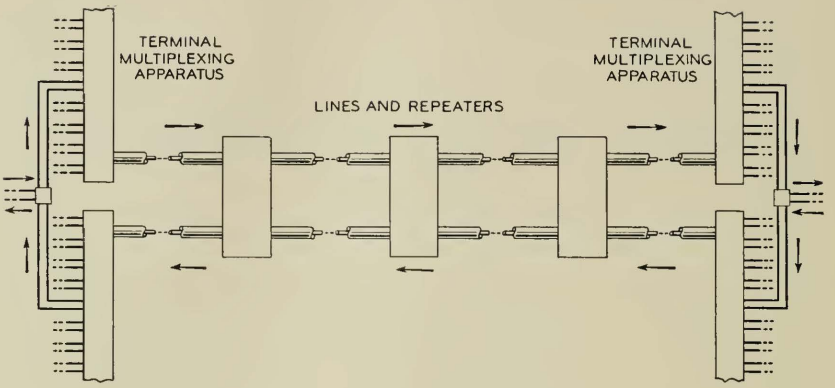


Fig. 1—Diagram of coaxial system.

subsequently. Such a coaxial line can be constructed to have the same degree of mechanical flexibility as the familiar telephone cable. While this line has a relatively high loss at high frequencies, the transmission path is particularly well adapted to the frequent application of repeaters, since the shielding permits the transmission currents to fall to low power levels at the high frequencies.

Of no little importance also is the fact that the attenuation-frequency characteristic is smooth throughout the entire band and obeys a simple law of change with temperature. (This is due to the fact that the dielectric is largely gaseous and that insulation material of good dielectric properties is employed.) This smooth relation is extremely

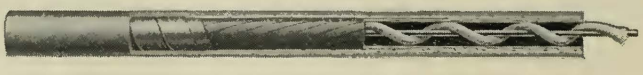


Fig. 2—Small flexible coaxial structure.

helpful in the provision of means in the repeaters for automatically compensating for the variations which occur in the line attenuation with changes of temperature. This type of system is featured by large transmission losses which are offset by large amplification, and it is necessary that the two effects match each other accurately at all times throughout the frequency range.

It will be evident that the repeater is of outstanding importance in this type of system, for it must not only transmit the wide band of frequencies with a transmission characteristic inverse to that of the line, with automatic regulation to care for temperature changes, but must also have sufficient freedom from inter-modulation effects to permit the use of large numbers of repeaters in tandem without objec-

tionable interference. Fortunately, recent advances in repeater technique have made this result possible, as will be appreciated from the subsequent description.

An interesting characteristic of this type of system is the way in which the width of the transmitted band is controlled by the repeater spacing and line size, as follows:

1. *For a given size of conductor and given length of line, the band width increases nearly as the square of the number of the repeater points.* Thus, for a coaxial circuit with about .3-inch inner diameter of outer conductor, a 20-mile repeater spacing will enable a band up to about 250,000 cycles to be transmitted, a 10-mile spacing will increase the band to about 1,000,000 cycles, and a 5-mile spacing to about 4,000,000 cycles.
2. *For a given repeater spacing, the band width increases approximately as the square of the conductor diameter.* Thus, whereas a tube of .3-inch inner diameter will transmit a band of about 1,000,000 cycles, .6-inch diameter will transmit about 4,000,000 cycles, while a diameter corresponding to a full-sized telephone cable might transmit something of the order of 50,000,000 cycles, depending upon the dielectric employed and upon the ability to provide suitable repeaters.

Earlier Work

It may be of interest to note that as a structure, the coaxial form of line is old—in fact, classical. During the latter half of the last century it was the object of theoretical study, in respect to skin effect and other problems, by some of the most prominent mathematical physicists of the time. Reference to some of this work is made in a paper by Schelkunoff, dealing with the theory of the coaxial circuit.[6]

On the practical side, it is found on looking back over the art that the coaxial form of line structure has been used in two rather widely different applications: first, as a long line for the transmission of low frequencies, examples of which are usage for submarine cables,[7, 8] and for power distribution purposes, and second as a short-distance, high-frequency line serving as an antenna lead-in.[9, 10]

The coaxial conductor system herein described may be regarded as an extension of these earlier applications to the long-distance transmission of a very wide range of frequencies suitable for multiplex telephony or television.[11] Although dealing with radio frequencies, this system represents an extreme departure from radio systems in that a relatively broad band of waves is transmitted, this band being con-

fined to a small physical channel which is shielded from outside disturbances. The system, in effect, comprehends a frequency spectrum of its own and shuts it off from its surroundings so that it may be used again and again in different systems without interference.

This new type of facility has not yet been commercially applied. It is, in fact, still in the development stage. Sufficient progress has already been made, however, to give reasonable assurance of a satisfactory solution of the technical problems involved. This progress is outlined below under three general headings: (1) the coaxial line and its transmission properties, (2) the wide band repeaters, and (3) the terminal apparatus.

The Coaxial Line

An Experimental Verification

One of the first steps taken in the present development was in the nature of an experimental check of the coaxial conductor line, designed primarily to determine whether the desirable transmission properties which had been disclosed by a theoretical study could be fully realized under practical conditions. For this purpose a length of coaxial structure capable of accurate computation was installed near Phoenixville, Pa. Figure 3 shows a sketch of the structure used and gives its dimensions. It comprised a copper tube of 2.5 inches outside diameter, within which was mounted a smaller tube which, in turn, contained a small copper wire. Two coaxial circuits of different sizes were thus made available, one between the outer and the inner tubes, and the other between the inner tube and the central wire. The instal ation comprised two 2600-foot lengths of this structure.

The diameters of these coaxial conductors were so chosen as to obtain for each of the two transmission paths a diameter ratio which approximates the optimum value, as discussed later. The conductors were separated by small insulators of isolantite. The rigid construction and the substantial clearances between conductors made it possible to space the insulators at fairly wide intervals, so that the dielectric between conductors was almost entirely air. The outer conductor was made gas-tight, and the structure was dried out by circulating dry nitrogen gas through it. The two triple conductor lines were suspended on wooden fixtures and the ends brought into a test house, as shown in Fig. 4.

The attenuation was measured by different methods over the frequency range from about 100 kilocycles to 10,000 kilocycles. Investigation showed that the departures from ideal construction occasioned by the joints, the lack of perfect concentricity, etc., had remark-

ably little effect on the attenuation. In order to study the effect of eccentricity upon the attenuation, tests were made in which this effect was much exaggerated, and the results substantiated theoretical predictions. The impedance of the circuits was measured over the same range as the attenuation. A few measurements on a short length were made at frequencies as high as 20,000 kilocycles.

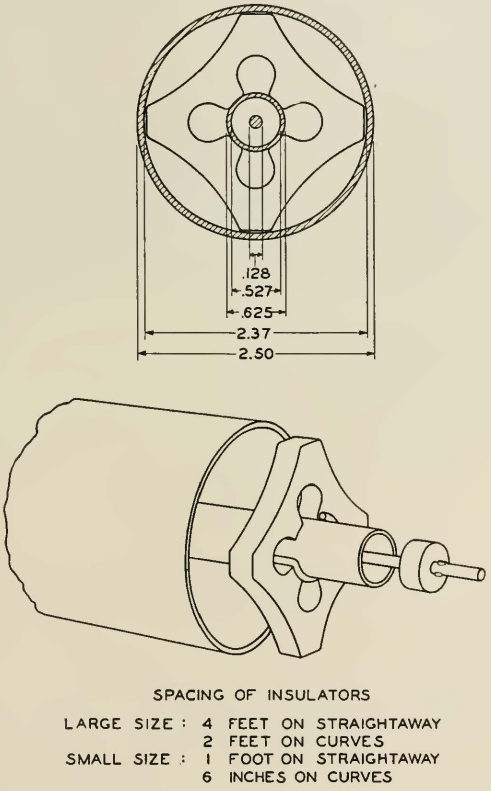


Fig. 3—Structure used in Phoenixville installation.

Measurements were secured of the shielding effect of the outer conductor of the coaxial circuit up to frequencies in the order of 100 to 150 kilocycles, the results agreeing closely with the theoretical values. Above these frequencies, even with interfering sources much more powerful than would be encountered in practice, the induced currents dropped below the level of the noise due to thermal agitation of electricity in the conductors (resistance noise) and could not be measured.

The preliminary tests at Phoenixville, therefore, demonstrated that



Fig. 4—Phoenixville installation showing conductors entering test house.

a practical coaxial circuit, with its inevitable mechanical departures from the ideal, showed transmission properties substantially in agreement with the theoretical predictions.

Small Flexible Structures

Development work on wide-band amplifiers, as discussed later, indicated the practicability of employing repeaters at fairly close intervals. This pointed toward the desirability of using sizes of coaxial circuit somewhat smaller than the smaller of those used in the preliminary experiments, and having correspondingly greater attenuation. Furthermore, it was desired to secure flexible structures which could be handled on reels after the fashion of ordinary cable. Accordingly, several types of flexible construction, ranging in outer diameter from about .3 inch to .6 inch, have been experimented with. Structures were desired which would be mechanically and electrically satisfactory, and which could be manufactured economically, preferably with a continuous process of fabrication.

One type of small flexible structure which has been developed is shown in Fig. 2. The outer conductor is formed of overlapping copper strips held in place with a binding of iron or brass tape. The insulation consists of a cotton string wound spirally around the inner conductor, which is a solid copper wire. This structure has been made in several sizes of the order of 1/2 inch diameter or less. When it is to be used as an individual cable, the outer conductor is surrounded by a

lead sheath, as shown, to prevent the entrance of moisture. One or more of the copper tape structures without individual lead sheath may be placed with balanced pairs inside a common cable sheath.

Another flexible structure is shown in Fig. 5. The outer conductor in this case is a lead sheath which directly surrounds the inner conductor with its insulation. Since lead is a poorer conductor than copper, it is necessary to use a somewhat larger diameter with this construction in order to obtain the same transmission efficiency. Lead is also inferior to copper in its shielding properties and to obtain the same degree of shielding the lead tube of Fig. 5 must be made correspondingly thicker than is necessary for a copper tube.

The insulation used in the structure shown in Fig. 5 consists of hard

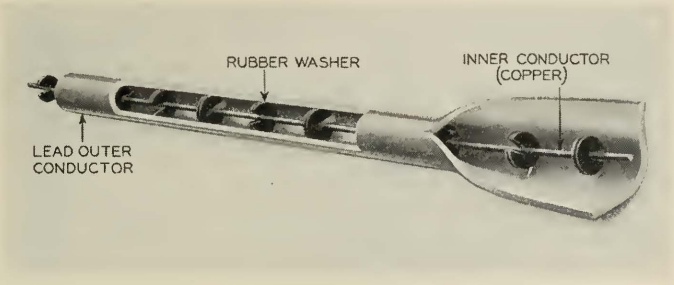


Fig. 5—Coaxial structure with rubber disc insulators.

rubber discs spaced at intervals along the inner wire. Cotton string or rubber disc insulation may be used with either form of outer tube. The hard rubber gives somewhat lower attenuation, particularly at the higher frequencies.

Another simple form of structure employs commercial copper tubing into which the inner wire with its insulation is pulled. Although this form does not lend itself readily to a continuous manufacturing process, it may be advantageous in some cases.

Transmission Characteristics

Attenuation

At high frequencies the attenuation of the coaxial circuit is given closely by the well-known formula:

$$\alpha = \frac{R}{2}\sqrt{\frac{C}{L}} + \frac{G}{2}\sqrt{\frac{L}{C}}, \tag{1}$$

where R, L, C and G are the four so-called "primary constants" of the line, namely, the resistance, inductance, capacitance and conductance

per unit of length. The first term of (1) represents the losses in the conductors, while the second term represents those in the dielectric.

When the dielectric losses are small, the attenuation of a coaxial circuit increases, due to skin effect in the conductors, about in accordance with the square root of the frequency. With a fixed diameter ratio, the attenuation varies inversely with the diameter of the circuit. By combining these relations there are obtained the laws of variation of band width in accordance with the repeater spacing and the size of circuit, as stated previously.

The attenuation-frequency characteristic of the flexible structure illustrated in Fig. 2, with about .3 inch diameter, is given in Fig. 6.

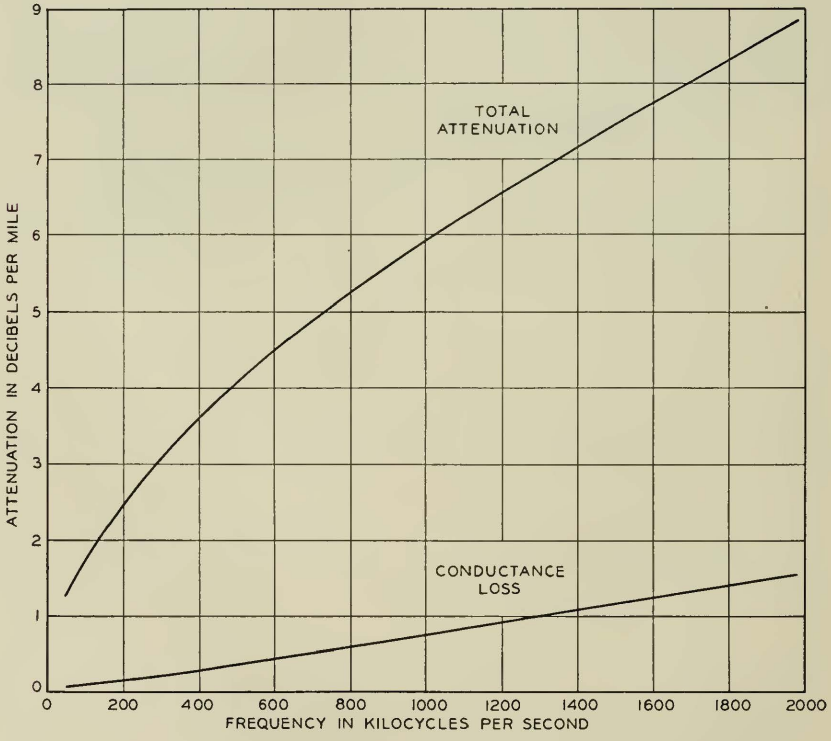


Fig. 6—Attenuation of small flexible coaxial structure (Fig. 2).

The figure shows also that the conductance loss due to the insulation is a small part of the total.

It is interesting to compare the curves of the transmission characteristics of the coaxial circuit with those of other types of circuits. Figure

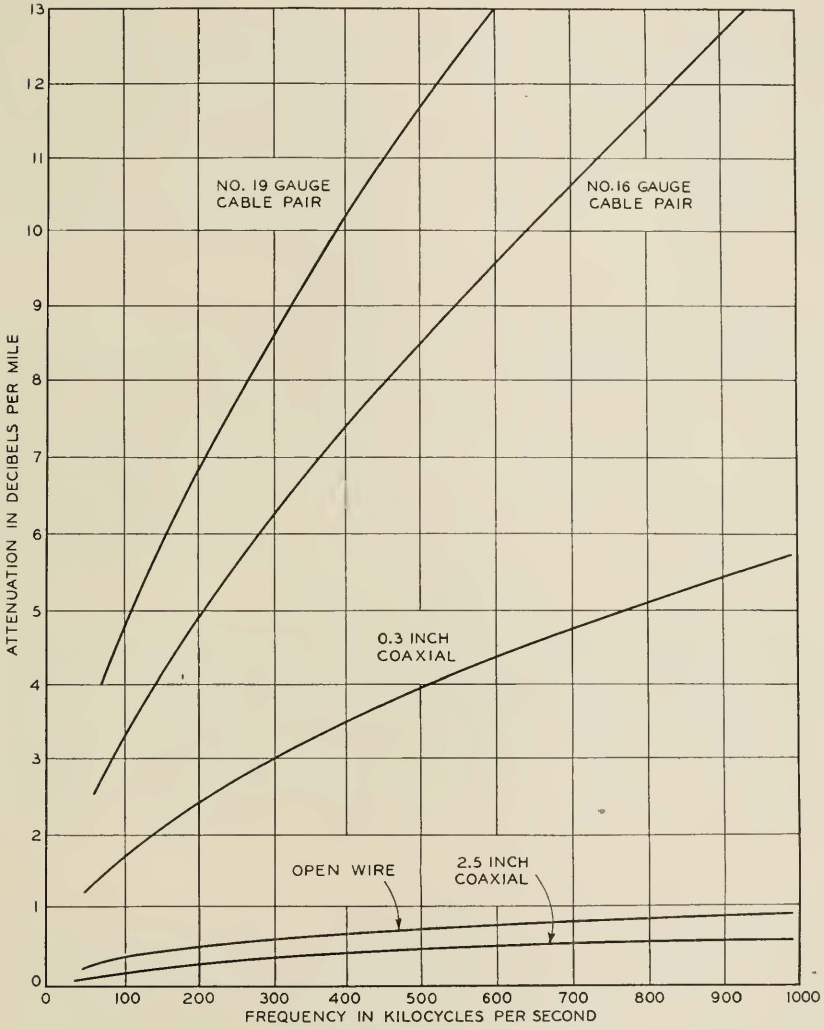


Fig. 7—Attenuation frequency characteristics of coaxial and other circuits.

7 shows the high-frequency attenuation of two sizes of coaxial circuit using copper tube outer conductors, of .3 inch and 2.5 inch inner diameter, and that of cable and open-wire pairs in the same frequency range.

Effect of Eccentricity

The small effect of lack of perfect coaxiality upon the attenuation of a coaxial circuit is illustrated by the curve of Fig. 8, which shows

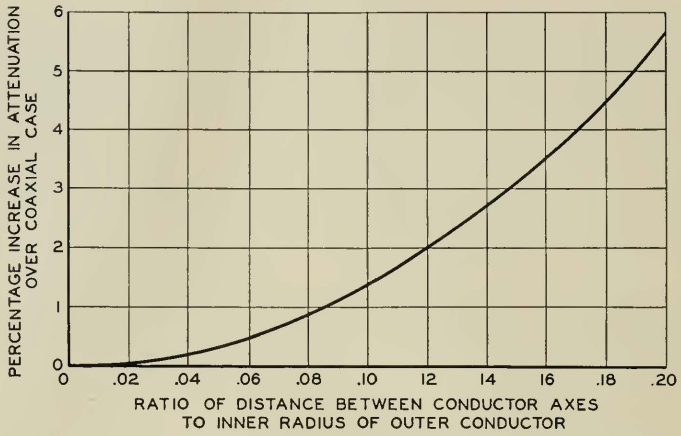


Fig. 8—Increase in attenuation of coaxial circuit due to eccentricity.

attenuation ratios plotted as a function of eccentricity, assuming a fixed ratio of conductor diameters and substantially air insulation.

Temperature Coefficient

With a coaxial circuit, as with other types of circuits, the temperature coefficient of resistance decreases as the frequency is increased, due to the action of skin effect, and approaches a value of one-half the d.-c. temperature coefficient.[12] Thus, for conductors of copper the a.-c. coefficient at high frequencies is approximately .002 per degree Centigrade. When the dielectric losses are small, the temperature coefficient of attenuation at high frequencies is the same as the temperature coefficient of resistance.

Diameter Ratio

An interesting condition exists with regard to the relative sizes of the two conductors. For a given size of outer conductor there is a unique ratio of inner diameter of outer conductor to outer diameter of inner conductor which gives a minimum attenuation. At high frequencies, this optimum ratio of diameters (or radii) is practically independent of frequency. When the conductivity is the same for both conductors, and either the dielectric losses are small or the insulation is distributed so that the dielectric flux follows radial lines, the value of the optimum diameter ratio is approximately 3.6. When the outer and inner conductors do not have the same conductivity, the optimum diameter ratio differs from this value. For a lead outer conductor and copper inner conductor, for example, the ratio should be about 5.3.

Stranding

Inasmuch as the resistance of the inner conductor contributes a large part of the high frequency attenuation of a coaxial circuit, it is natural to consider the possibility of reducing this resistance by employing a conductor composed of insulated strands suitably twisted or interwoven.[13] Experiments along this line showed that this method is impractical at frequencies above about 500 kilocycles, owing to the fineness of stranding required.

Characteristic Impedance

The high-frequency characteristic impedance of a coaxial circuit varies inversely with the square root of the effective dielectric constant, i.e., the ratio of the actual capacitance to the capacitance that would be obtained with air insulation. The impedance of a circuit having a given dielectric constant depends merely upon the ratio of conductor diameters and not upon the absolute dimensions. For a diameter ratio of 3.6, the impedance of a coaxial circuit with gaseous insulation is about 75 ohms.

Velocity of Propagation

For a coaxial circuit with substantially gaseous insulation, the velocity of propagation at high frequencies approaches the speed of light. Hence the circuit is capable of providing high velocity telephone channels with their well-recognized advantages. The fact that the velocity at high frequencies is substantially constant minimizes the correction required to bring the delay distortion within the limits required for a high quality television band.

Shielding and Crosstalk

The shielding effect of the outer conductor of a coaxial circuit is illustrated in Fig. 9, where the transfer impedance between the outer and inner surfaces of the outer conductor is plotted as a function of frequency. There will be observed the sharp decrease in inductive susceptibility as the frequency rises. On this account, the crosstalk between adjacent coaxial circuits falls off very rapidly with increasing frequency. The trend is, therefore, markedly different from that for ordinary non-shielded circuits which rely upon balance to limit the inductive coupling. As a practical matter, less shielding is ordinarily required to avoid crosstalk than to avoid external interference.

With suitable design the shielding effect of the outer conductor renders the coaxial circuit substantially immune to external interference at frequencies above the lower end of the spectrum. Hence the signals transmitted over the circuit may be permitted to drop

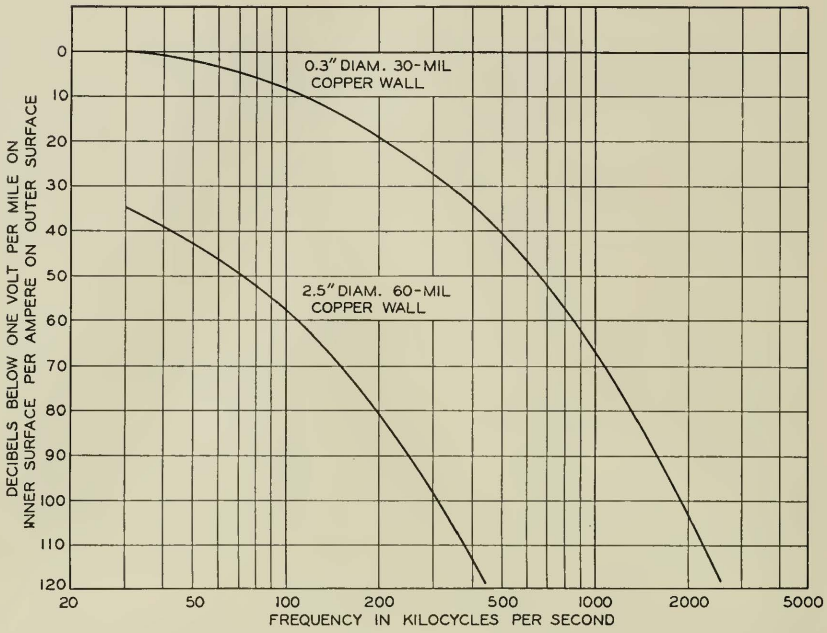


Fig. 9—Transfer impedance of coaxial circuit.

down to a level determined largely by the noise due to thermal agitation of electricity in the conductors and tube noise in the associated amplifiers. It appears uneconomical to make the outer conductor sufficiently thick to provide adequate shielding for the very low frequencies. Also it seems impractical to design the repeaters to transmit very low frequencies. Hence the best system design appears to be one in which the lowest five or ten per cent of the frequency range is not used for signal transmission. The coaxial circuit is, however, well suited to the transmission of 60-cycle current for operating repeaters, a matter which will be referred to later.

Broad-Band Amplifiers

In order to realize the full advantage of broad-band transmission, the repeater for this type of system should be capable of amplifying the entire frequency band *en bloc*. Furthermore, it should be so stable and free from distortion that a large number of repeaters may be operated in tandem. Although high-gain radio frequency amplifiers are in everyday use, these are generally arranged to amplify at any one time only a relatively narrow band of frequencies, a variable tuning device being provided so that the amplification may be obtained at

any point in a fairly wide frequency range. The high gain is usually obtained by presenting a high impedance to the input circuits of the various tubes through tuning the input and interstage coupling circuits to approximate anti-resonance.

In amplifying a broad band of frequencies, it is difficult to maintain a very high impedance facing the grid circuits. The inherent capacitances between the tube elements and in the mounting result in a rather low impedance shunt which can not be resonated over the desired frequency band. It is, therefore, necessary to use relatively low impedance coupling circuits and to obtain as high gain as possible from the tubes themselves. The amount of gain which can be obtained without regeneration depends, of course, upon the type of tube, the number of amplification stages, the band width, and also upon the ratio of highest to lowest frequency transmitted.

Repeater Gain

The total net gain desired in a line amplifier is such as to raise the level of an incoming signal from its minimum permissible value, which is limited by interference, up to the maximum value which the amplifier can handle.

As pointed out above, the noise in a well shielded system is that due to resistance noise in the line conductors and tube noise in the amplifiers. In some of the repeaters which have been built, the amplifier noise has been kept down to about 2 db above resistance noise, corresponding to about 7×10^{-17} watt per voice channel. In a long line with many repeaters the noise voltages add at random, or in other words, the noise powers add directly. Assuming, for example, a line with 200 repeaters, the noise power at the far end would be 200 times that for a single repeater section. In general, the line and amplifier noise will not be objectionable in a long telephone channel if the speech sideband level at any amplifier input is not permitted to drop more than about 55 db below the level of the voice frequency band at the transmitting toll switchboard.

The determination of the volume which a tube can handle in transmitting a wide band of frequencies involves a knowledge of the distribution in time and frequency of the signaling energy and of the requirements as to distortion of the various components of the signal. The distribution of the energy in telephone signals has been the subject of much study. This distribution is known to vary over very wide limits, depending upon the voice of the talker and many other factors. It is, therefore, obvious that the problem of summing up the energy of some hundreds of simultaneous telephone conversations is a difficult one.

Enough work has been done, however, to indicate fairly well what the result of such addition will be.

As to distortion in telephone transmission, the most serious problem has been to limit the intermodulation between various signals which are transmitted simultaneously through the repeater and appear as noise in the telephone channel. The requirement for such noise is similar to that for line and tube noise, and similarly it will add up in successive repeater sections for a long line. With present types of tubes operating with a moderate plate potential, the modulation requirement can be met only at relatively low output levels. To improve this situation and also to obtain advantages in amplifier stability, the reversed feedback principle employed for cable carrier amplifiers, as described in a paper by H. S. Black,[14] has been extended to higher frequency ranges. It has been found that amplifiers of this type having 30 db feedback reduce the distortion to such an extent that each amplifier of a long system carrying several hundred telephone channels will handle satisfactorily a channel output signal level about 5 db above that at the input of the toll line.

The maximum gain which can be used in the repeater, therefore, is, in the illustrative case given above of a long system carrying several hundred telephone channels, the difference between the minimum and maximum levels of 55 db below and 5 db above the point of reference, respectively, or a total gain of 60 db. (With a .3-inch coaxial line of the type shown in Fig. 2, this corresponds to a repeater spacing of about 10 miles.) If a repeater is to have 60 db net gain and at the

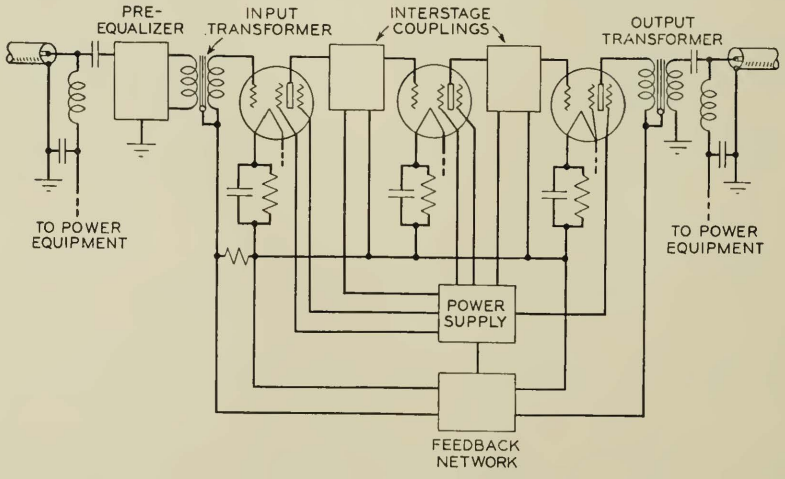


Fig. 10—Circuit of 1000-kilocycle three-stage feedback repeater.

same time about 30 db feedback, it is obvious that the total forward gain through the amplifying stages must be about 90 db. The circuit of an experimental amplifier meeting the gain requirements for a frequency band from 50 to 1000 kilocycles is shown schematically in Fig. 10.

Gain-Frequency Characteristic

As pointed out above, the line attenuation is not uniform with frequency. For a repeater section which has a loss of, say, 60 db at 1000 kilocycles, the loss at 50 kilocycles would be only about 15 db. Such a sloping characteristic can be taken care of either by designing the repeater to have an equivalent slope in its gain-frequency characteristic or by designing it for constant gain and supplementing it with an equalizer which gives the desired overall characteristic. Both methods have been tried out, as well as intermediate ones. Figure 11

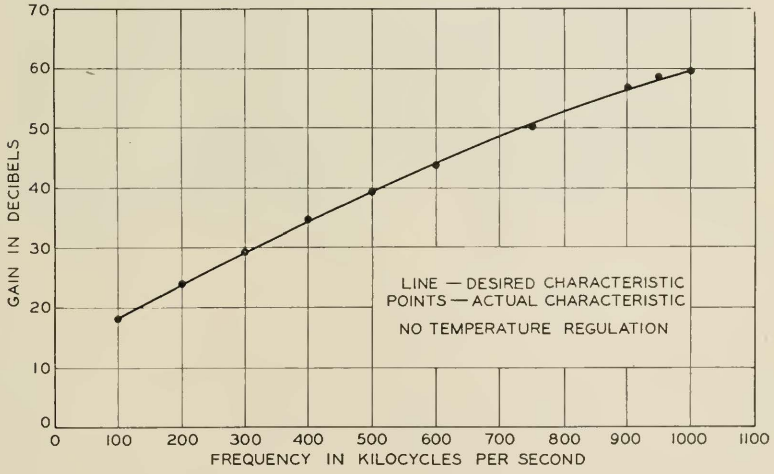


Fig. 11—Gain of 1000-kilocycle repeater compared with line characteristic.

illustrates such a sloping characteristic obtained by adjusting the coupling impedances in a three-tube repeater, designed in this case for 60 db gain at 1000 kilocycles. The accompanying photograph, Fig. 12, gives an idea of the apparatus required in such a repeater, apart from the power supply equipment.

Regulation for Temperature Changes

It is necessary that the repeater provide compensation for variations in the line attenuation due to changes of temperature. In the case of aerial construction such variations might amount to as much as 8 per cent in a day or 16 per cent in a year. If the line is under-

ground the annual variation is only about one-third of the above value and the changes occur much more slowly. On a transcontinental line the annual variation might total about 1500 db. Inasmuch as it is desirable to hold the transmission on a long circuit constant within about ± 2 db, it is obvious that the regulation problem is a serious one.

In a single repeater section of aerial line the variation might amount to ± 2.5 db per day or ± 5 db per year. Such variations, if allowed

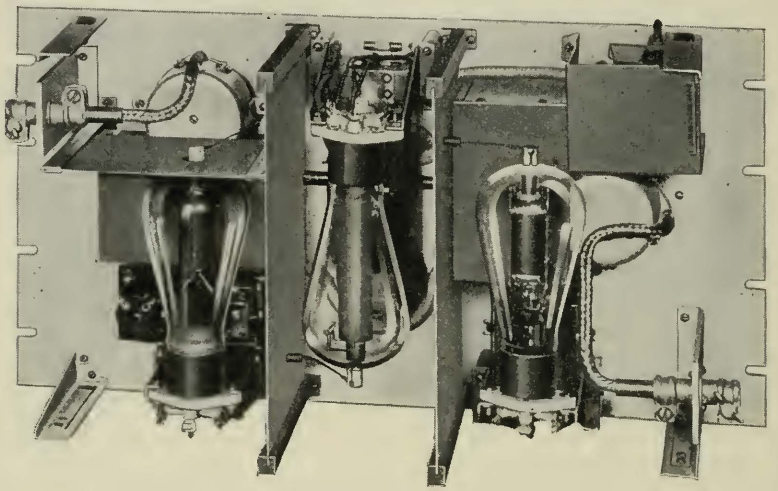


Fig. 12—Photograph of 1000-kilocycle repeater.

to accumulate over several repeater sections, will drop the signal down into the noise or raise it so as to overload the tubes. It is, therefore, advisable to provide some regulation at every repeater in an aerial line so as to maintain the transmission levels at approximately their correct position. For underground installations the regulating mechanism may be omitted on two out of every three repeaters.

In choosing a type of regulator system the necessity for avoiding cumulative errors in the large number of repeater sections has been borne in mind. In view of the wide band available, a pilot channel regulator system was naturally suggested. Such a scheme employing two pilot frequencies has been used experimentally to adjust the gain characteristic in such a way as to maintain the desired levels throughout the band. The accuracy with which this has been accomplished for a single repeater section is illustrated in Fig. 13. Over the entire band of frequencies and the extreme ranges in temperature which may be encountered, the desired regulation is obtained within a few tenths of a db.

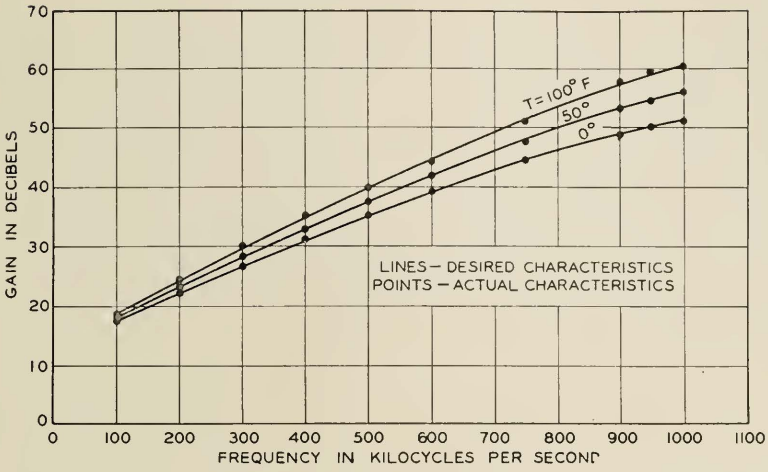


Fig. 13—Temperature regulation—line and repeater characteristics.

Repeater Operation, Power Supply, Housing, Etc.

In view of the large number of repeaters required in a broad-band transmission system it is essential that the repeater stations be simple and involve a minimum of maintenance. With the repeater design as described it is expected that most of the repeaters may be operated on an unattended basis, requiring maintenance visits at infrequent intervals.

An important factor in this connection is the possibility of supplying current to unattended repeaters over the transmission line itself. The coaxial line is well adapted to transmit 60-cycle current to repeaters without extreme losses and without hazard. The repeaters with regulating arrangements as built experimentally for a million-cycle system are designed to use 60-cycle current, which in this case appears to have the usual advantages over d.-c. supply. One repeater requires a supply of about 150 watts. The number of repeaters which can be supplied with current transmitted over the line from any one point depends upon the voltage limitation which may be imposed on the circuit from considerations of safety.

For a repeater of the type described with current supplied over the line, only a very modest housing arrangement will be required. For the great majority of stations, it appears possible to accommodate the repeaters in weatherproof containers mounted on poles, in small huts, or in manholes.

Higher Frequency Repeaters

Most of what has been said above applies particularly to repeaters transmitting frequencies up to about 1000 kilocycles. However, study has been given also to repeaters, both of the feedback and the non-feedback type, for transmitting higher frequencies. Experimental repeaters covering the range from 500 to 5000 kilocycles have been built and tested. These were capable of handling simultaneously the full complement of over 1000 channels which such a broad band will permit. The frequency characteristic of one of these repeaters, and the measured attenuation of a section of line of the type tested at Phoenixville are shown in Fig. 14.

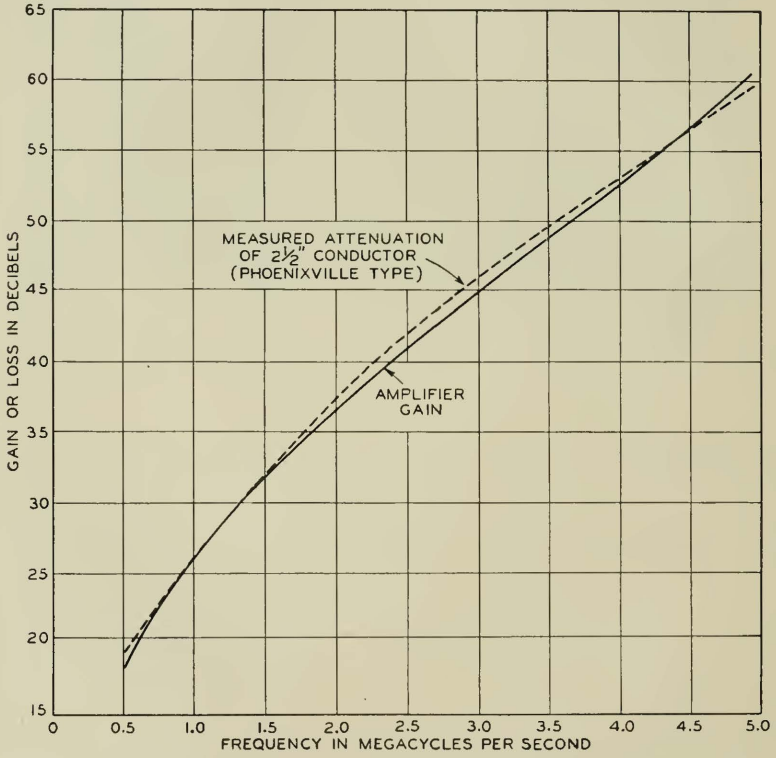


Fig. 14—Frequency characteristic of coaxial line and 5000-kilocycle repeater.

Terminal Arrangements

In order to utilize a broad band effectively for telephone purposes, the speech channels must be placed as close together in frequency as practicable. The factors which limit this spacing are: (1) The width of

speech band to be transmitted and (2), the sharpness of available selecting networks.

As to the width of speech band, the present requirement for commercial telephone circuits is an effective transmission band width of at least 2500 cycles, extending from 250 to 2750 cycles. It has been found that a band of this width or more may be obtained with channels spaced at 4000-cycle intervals. Band filters using ordinary electrical elements are available,[3] for selecting such channels in the range from zero to about 50 kilocycles. Channel selecting filters using quartz crystal elements [15, 16] have been developed in the range from about 30 to 500 kilocycles. The selectivity of a typical filter employing quartz crystal elements is shown on Fig. 15.

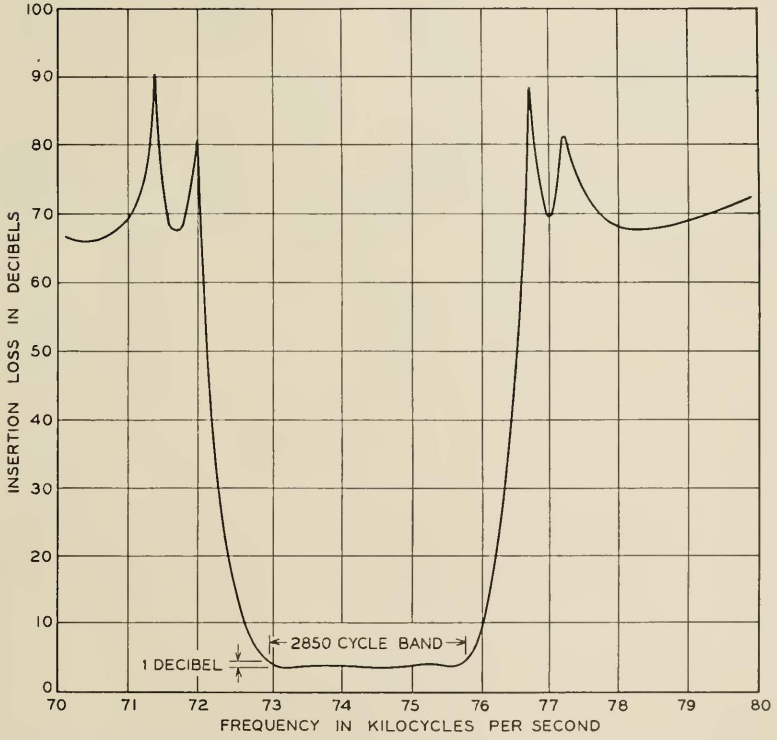


Fig. 15—Frequency characteristic of quartz crystal channel band filter.

Initial Step of Modulation

The initial modulation (from the voice range) may be carried out in an ordinary vacuum tube modulator or one of a number of other non-linear devices. The method chosen for the present experimental work

employs a single sideband with suppressed carrier, using a copper-oxide modulator associated with a quartz crystal channel filter. The terminal apparatus required for two-way transmission over a two-path circuit is shown diagrammatically on the left-hand side of Fig. 16.

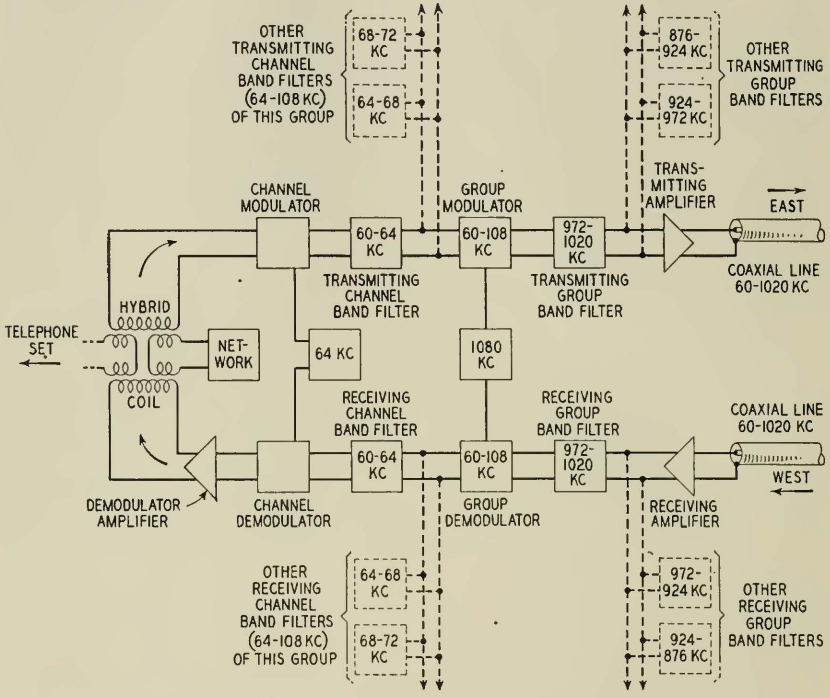


Fig. 16—Schematic of four-wire circuit employing two steps of modulation.

A frequency allocation which has been used for experimental purposes employs carriers from 64 to 108 kilocycles for the initial step of modulation. The lower sidebands are selected and placed side by side in the range from 60 to 108 kilocycles, as illustrated in Fig. 17, forming a group of 12 channels.

Double Modulation

In order to extend the frequency range of a system to accommodate a very large number of channels, it appears to be more economical to add a second step of modulation rather than carry the individual channel modulation up to higher frequencies. Such a second step of modulation has been used experimentally to translate the initial group of 12 channels en bloc from the range 60 to 108 kilocycles up to higher frequencies. It is possible to place such groups of channels one above another as illustrated in the upper part of the diagram of Fig. 18, up

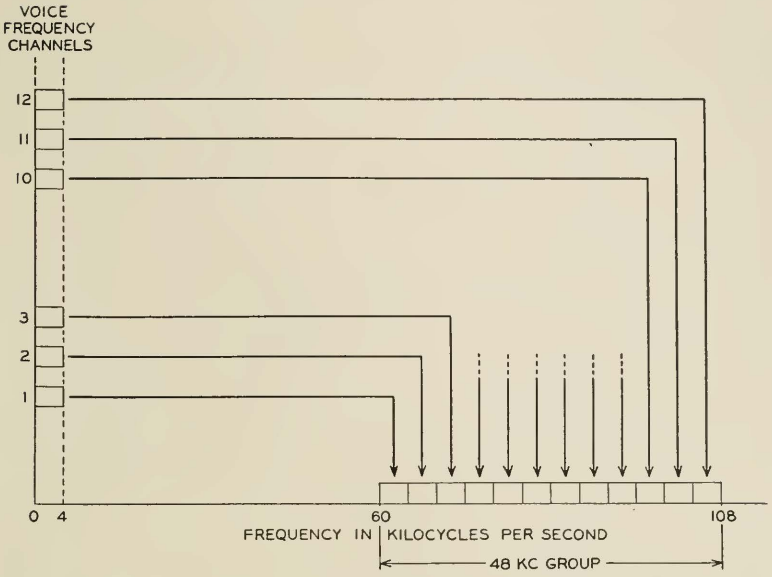


Fig. 17—Diagram illustrating frequency allocation for first step of modulation.

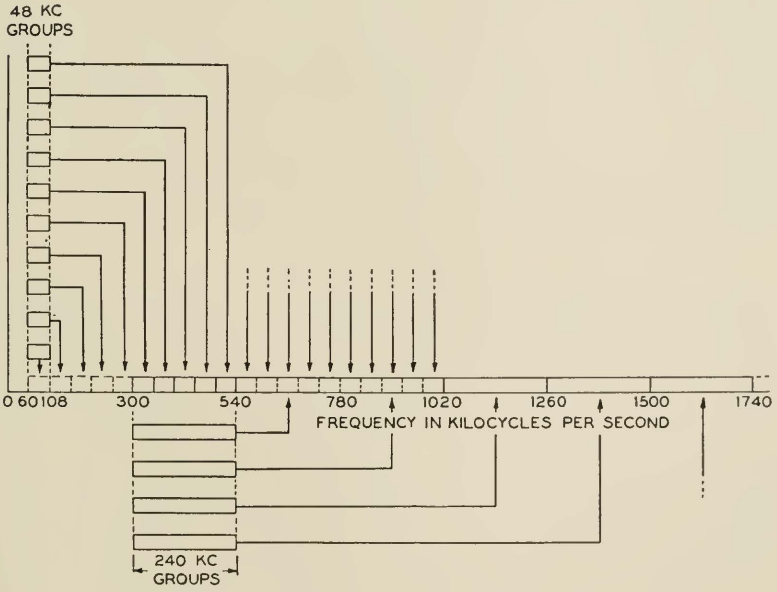


Fig. 18—Diagram illustrating frequency allocation for two or three steps of modulation.

to about 1000 kilocycles, wasting no frequency space between groups and thus keeping the channels spaced at intervals of 4 kilocycles throughout the entire range.

The apparatus required for this purpose is shown schematically in Fig. 16, which illustrates the complete terminal arrangements for a single channel employing double modulation. The figure indicates by dottled lines where the other channels and groups of channels are connected to the system.

A modulator for shifting the frequency position of a group of channels inherently yields many different modulation products as a result of the intermodulation of the signal frequencies with the carrier frequency and/or with one another. Out of these products only the "group sideband" is desired. The number of the modulation products resulting merely from the lower ordered terms of the modulator response characteristic is extremely large. All such products must be considered from the standpoint of interference either with the group which is wanted in the output or with other groups to be transmitted over the system. Various expedients may be used to avoid interference as follows: (1) A proper choice of frequency allocation will place the undesired modulation products in the least objectionable location with respect to the wanted signal bands; (2) a high ratio of carrier to signal will minimize all products involving only the signal frequencies; (3) the use of a balanced modulator will materially reduce all products involving the second order of the signal; (4) selectivity in the group filters will tend to eliminate all products removed some distance from the wanted signal group. Giving due regard to these factors, balanced vacuum tube group modulators have been developed which are satisfactory for the frequency allocations employed.

Triple Modulation

For systems involving frequencies higher than about 1000 kilocycles it may be desirable to introduce a third step of modulation. In some experiments along this line a "super-group" of 60 channels, or five 12-channel groups, has been chosen. The lower part of Fig. 18 illustrates, for a triple modulation system, the shifting of supergroups of 60 channels each to the line frequency position. This method has been employed experimentally up to about 5,000 kilocycles. It is of interest to note that even in extending these systems to such high frequencies, channels are placed side by side at intervals of 4000 cycles to form a practically continuous useful band for transmission over the line.

Demodulation

On the receiving side the modulation process is reversed. The apparatus units are similar to those used on the transmitting side, and are similarly arranged. Figure 16 illustrates this for the case of double modulation.

Carrier Frequency Supply

In systems operating at higher frequencies it is necessary that the carrier frequencies be maintained within a few cycles of their theoretical position in order to avoid beat tones or distortion of the speech band. Separate oscillators of high stability could, of course, be used for the carrier supply but it appears more economical to provide carriers by means of harmonic generation from a fundamental basic frequency. Such a base frequency may be transmitted from one end of the circuit to the other, or may be supplied separately at each end.

Television

The broad band made available by the line and repeaters may be used for the transmission of signals for high-quality television. Such signals may contain frequency components extending over the entire range from zero or a very low frequency up to a million or more cycles.[4] The amplifying and transmitting of these frequencies, particularly the lower ones, presents a serious problem. The difficulty can be overcome by translating the entire band upward in frequency to a range which can be satisfactorily transmitted. To effect such a shift, the television band may first be modulated up to a position considerably higher than its highest frequency and then with a second step of modulation be stepped down to the position desired for line transmission.

This method is illustrated in Fig. 19 for a 500-kc. television signal band. The original television signal is first modulated with a relatively high frequency, two million cycles in this case (C_1). The lower sideband, extending to 1500 kilocycles, is selected and is modulated again with a frequency of 2100 kilocycles (C_2). The lower sideband of 100 to 600 kilocycles is selected with a special filter so designed that the low frequency end is accurately reproduced. The television signal then occupies the frequency range of 100 to 600 kilocycles as shown on the diagram and may be transmitted over a coaxial or other high frequency line. At the receiving end a reverse process is employed. The same method using correspondingly higher frequencies may be used for wider bands of television signals.

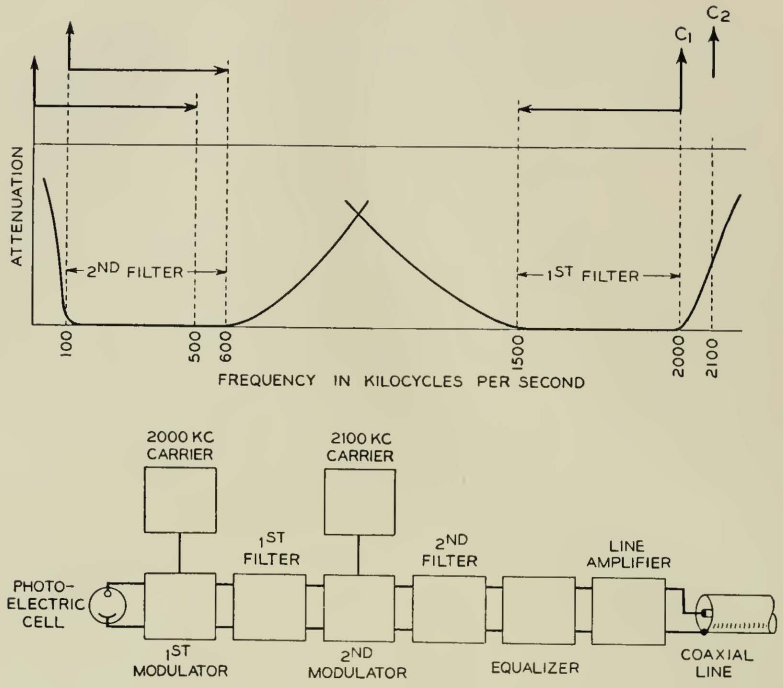


Fig. 19—Double modulation method for translating television signals for wire line transmission.

Other Communication Facilities

The telephone channels provided by the system may be used for other types of communication services, such as multi-channel telegraph, teletype, picture transmission, etc. For the transmission of a high-quality musical program, which requires a wider band than does commercial telephony, two or more adjacent telephone channels may be merged. The adaptability of the broad-band system to different types of transmission thus will be evident.

As already noted, the commercial application of these systems for wide-band transmission over coaxial lines must await a demand for large groups of communication facilities or for television. The results which have been outlined are based upon development work in the laboratory and the field, and it is probable that the systems when used commercially will differ considerably from the arrangements described.

References

1. E. H. Colpitts and O. B. Blackwell, "Carrier Current Telephone and Telegraphy," *A. I. E. E. Trans.*, Vol. 40, February 1921, p. 205–300.
2. H. A. Affel, C. S. Demarest, and C. W. Green, "Carrier Systems on Long Distance Telephone Lines," *A. I. E. E. Trans.*, Vol. 47, October 1928, 1360–1367. *Bell Sys. Tech. Jour.*, Vol. VII, July 1928, p. 564–629.
3. A. B. Clark and B. W. Kendall, "Communication by Carrier in Cable," *Elec. Engg.*, Vol. 52, July 1933, p. 477–481, *Bell Sys. Tech. Jour.*, Vol. XII, July 1933, p. 251–263.
4. P. Mertz and F. Gray, "Theory of Scanning and Its Relation to the Characteristics of the Transmitted Signal in Telephotography and Television," *Bell Sys. Tech. Jour.*, Vol. XIII, July 1934, p. 464.
5. E. W. Engstrom, "A Study of Television Image Characteristics," *Proc. I. R. E.*, Vol. 21, December 1933, p. 1631–1651.
6. S. A. Schelkunoff, "The Electromagnetic Theory of Coaxial Transmission Lines and Cylindrical Shields." *Bell Sys. Tech. Jour.*, Vol. XIII, October 1934.
7. J. R. Carson and J. J. Gilbert, "Transmission Characteristics of the Submarine Cable," *Jour. Franklin Institute*, Vol. 192, December 1921, p. 705–735.
8. W. H. Martin, G. A. Anderegg and B. W. Kendall, "The Key West-Havana Submarine Telephone Cable System," *Trans. A. I. E. E.*, Vol. 41, 1922, p. 1–19.
9. British Patent No. 284,005, C. S. Franklin, January 17, 1928.
10. E. J. Sterba and C. B. Feldman, "Transmission Lines for Short-Wave Radio Systems," *I. R. E. Proc.*, *Vol.* 20, July 1932, p. 1163–1202; also *Bell Sys. Tech. Jour.*, Vol. II, July 1932, p. 411–450.
11. U. S. Patents No. 1,835,031, L. Espenschied and H. A. Affel, December 9, 1931, and No. 1,941,116, M. E. Strieby, December 26, 1933.
12. E. I. Green, "Transmission Characteristics of Open-Wire Lines at Carrier Frequencies," *A. I. E. E. Trans.*, Vol. 49, October 1930, p. 1524–1535; *Bell Sys. Tech. Jour.*, Vol. IX, October 1930, p. 730–759.
13. H. A. Affel and E. I. Green, U. S. Patent No. 1,818,027, Aug. 11, 1931.
14. H. S. Black, "Stabilized Feed-Back Amplifiers," Electrical Engineering, Vol. 53, January 1934, p. 114–120. *Bell Sys. Tech. Jour.*, Vol. XIII, January 1934, p. 1–18.
15. L. Espenschield, U. S. Patent No. 1,795,204, March 3, 1931.
16. W. P. Mason, "Electrical Wave Filters Employing Quartz Crystals as Elements," *Bell Sys. Tech. Jour.*, Vol. XIII, July 1934, p. 405.

Regeneration Theory and Experiment *

By

E. PETERSON, J. G. KREER, AND L. A. WARE

A comprehensive criterion for the stability of linear feed-back circuits has recently been formulated by H. Nyquist, in terms of the transfer factor around the feed-back loop. The importance of any such general criterion lends interest to an experimental verification, with which the paper is primarily concerned.

The subject is dealt with under five principal headings. The first section reviews some of the criteria for oscillation to be found in the literature of vacuum tube oscillators. The second describes the derivation of Nyquist's criterion somewhat along the lines followed by Routh in one of his investigations of the stability of dynamical systems. The third part deals with two experimental methods used in measuring the transfer factor. The fourth is concerned with the particular amplifier circuit used in the test of Nyquist's criterion. The last section applies the criterion to a nonlinear case, and to circuits including two-terminal negative impedance elements.

IN a comparatively recent paper on "Regeneration Theory," [1] Dr. Nyquist presented a mathematical investigation of the conditions under which instability [2] exists in a system made up of a linear amplifier and a transmission path connected between its input and output circuits. The results of the investigation are of interest because of their obvious application to amplifiers provided with feed-back paths,[3] as well as to the starting conditions in oscillators. As a result of his general analysis, Dr. Nyquist arrived at a criterion for stability, expressed in particularly simple and convenient form, which is not restricted in its range of application to particular amplifier and circuit configurations.

The great value attached to a criterion as precise and as general as Nyquist's makes it desirable to submit the criterion to an experimental test. One particularly striking conclusion drawn from this criterion is that under certain conditions a feed-back amplifier may sing within certain limits of gain, but either reduction or increase of gain beyond these limits may stop singing. A feed-back amplifier satisfying these conditions was set up, and the experimental results were found to be in agreement with this conclusion.

* Published in *Proc. I. R. E.*, October, 1934.

[1] *Bell. Sys. Tech. Jour.*, vol. XI, p. 126.

[2] Instability is used in the sense that a small impressed force, which dies out in course of time, gives rise to a response which does not die out.

[3] *Electrical Engineering*, July, 1933; *Bell Sys. Tech. Jour.*, p. 258, July, 1933.

It is interesting to compare the criterion with those derived for the mechanical systems of classical dynamics. In his Adams Prize Paper on "The Stability of Motion," [4] and again in his "Advanced Rigid Dynamics," [5] Routh investigated the general problem of dynamic stability and established a number of criteria based upon various properties of dynamical systems. When applied to the problem of feed-back amplifiers, keeping Nyquist's result in mind, one of them is found to be equivalent to Nyquist's criterion, although expressed in different terms and derived in a different way.

To provide a background for the experiments, we propose to state some of the criteria for stability which are to be found in the literature of vacuum tube oscillators, and to compare them with Nyquist's or Routh's criterion, the development of which is most conveniently described somewhat along the lines followed by Routh. Following this we shall deal with the experimental methods and apparatus which were used in testing the criterion, and conclude with some extensions of the criterion.

Circuit Analysis and Stability

Conditions required for the starting of oscillations in linear feedback circuits, corresponding to instability, are to be found in the literature of vacuum tube oscillator circuits, expressed in a number of ostensibly different forms. These are usually based upon the familiar mesh differential equations for the system which involve differentiations and integrations of the mesh amplitudes with respect to time. Using the symbol p to denote differentiation with respect to time, each mesh equation becomes formally an algebraic one in p, involving the circuit constants and the mesh amplitudes. The solution of this system of equations is known to be expressible as the sum of steady state and transient terms. The transient terms are each of the form $B_k \epsilon^{p_k t}$, the B_k's being fixed by initial conditions, and the p_k's being determined from the circuit equations. If we set up the determinant of the system of equations—the discriminant—and equate it to zero, the roots of the resulting equation are the p_k's above. In general each mesh equation involves p to the second degree at most, and with n meshes the discriminant is of degree $2n$ at most. Accordingly we may express the determinantal equation as

$$F(p) = 0 = K(p - p_1)(p - p_2) \cdots (p - p_{2n}). \qquad (1)$$

As for the steady state term, in the simplest case in which a sinusoidal wave of frequency $\omega/2\pi$ is impressed, it is equal to the impressed

[4] Macmillan, 1877.
[5] Macmillan, 6th edition, 1905.

voltage divided by the discriminant and multiplied by the appropriate minor of the determinant, in which p is replaced by $j\omega$. The character of the response due to a slight disturbance and in the absence of any periodic force is determined by the exponentials. In general, p_k is a complex quantity which may be written as $a_k + j\omega_k$. It is apparent that in the critical case for which a_k is zero, the corresponding term becomes $\epsilon^{j\omega_k t}$, corresponding to an oscillation invariable in amplitude, of frequency $\omega_k/2\pi$. If a_k is negative, as is ordinarily the case when the system is passive (containing no amplifier or negative impedance), then the oscillation diminishes in course of time. When a_k is positive, however, the oscillation increases with time, and the system is said to be unstable. Evidently the stability of a system is determined by the signs of the a_k's.

Several criteria which have previously been enunciated for the maintenance of free oscillations are deducible from the above. One states that the discriminant must vanish when p takes on the value $j\omega$. Another states that the damping (a_k) must be zero at the frequency of oscillation. These are clearly equivalent. Two derived criteria may also be mentioned, based upon the properties of the system when the circuit is broken. The first of these states that if the impedance is measured looking into the two terminals provided by the break. the impedance must be zero at the frequency of steady oscillation.

The second criterion involving the transfer factor has become fairly widespread, perhaps because it leads to a simple and plausible physical picture. To determine the transfer factor around the feed-back loop, the loop is broken at a convenient point, and the two sets of terminals formed by the break are each terminated in a passive impedance equal to that which is connected in the normal (unbroken) condition. Then when a voltage of frequency $\omega/2\pi$ is applied to one of the pairs of terminals so provided—the input terminals [6]—and the corresponding voltage is measured across the other pair, the transfer factor $A(j\omega)$ is obtained as the vector ratio of the output voltage to the input voltage.

The manner in which the transfer factor enters into the problem may be demonstrated directly by comparing the voltages at any point of the main amplifier circuit under the two conditions in which the feed-back path is opened and closed respectively. If with the feed-back path open the voltage at any such point is $E\epsilon^{pt}$, then when the feed-back path is closed the voltage will be changed [7] to

$$E\epsilon^{pt}/[1 - A(p)].$$

[6] Input terminals are those across which an impressed potential leads to propagation in the normal direction of amplifier transmission.

[7] *Bell Sys. Tech. Jour.*, Vol. XI, p. 128.

This may be shown as follows with reference to the particular circuit of Fig. 1: If the feed-back circuit is broken and then properly terminated, the voltage existing across the input is taken as e. Now suppose the feed-back path to be restored. Designating the voltage existing across the input in the presence of feed-back as e_1 we have $e_1 = e + Ae_1$, from which the above equation follows.

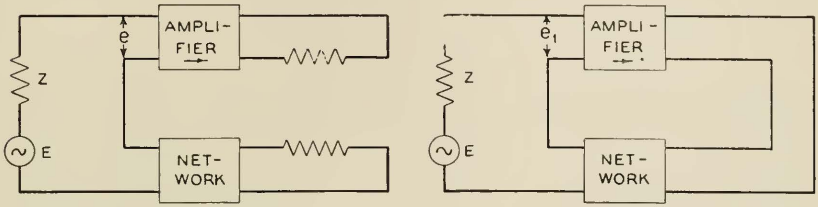


Fig. 1—Series type feed-back; loop broken and terminated at left, normal feed-back circuit at right.

If we let $F_1(p)$ represent the discriminant of the system when the loop is broken and terminated, then the roots of the equation formed by setting the discriminant equal to zero are assumed to have positive real parts. Now for the corresponding discriminant when the loop is restored, we have in accordance with the above considerations

$$F(p) = [(1 - A(p)]F_1(p).$$

In setting this discriminant equal to zero to obtain the roots, the only ones which have nonnegative real parts are those corresponding to the feed-back term

$$f(p) = 1 - A(p). \tag{2}$$

The above-mentioned criterion may be deduced from this expression. For steady oscillations to exist the output potential must be identical in amplitude and in phase with that existing across the input at the frequency of oscillation ($p = j\omega$), in which case the transfer factor is unity. This seems reasonable on the basis that when the input and output terminals are connected through, the oscillation will neither increase nor decrease with time. It may be demonstrated by direct analysis that these several criteria, framed for the critical case of undamped oscillations, all lead to the same correct conclusion.

Of course in any actual oscillating circuit it is practically impossible to get these conditions fulfilled exactly, and what is ordinarily done in the practical design of oscillating circuits is to ensure that the voltage fed back will be greater than that required to produce oscillation. This evidently goes a step further than the above criteria, and

reliance is placed upon the nonlinear properties of the circuit to fulfill the criteria automatically. The procedure is known by experiment to be effective in the usual type of oscillating circuit. In particular forms of feed-back circuits, however, it may be demonstrated that *the transfer factor may be made greater than unity without giving rise to oscillations*. This situation was investigated experimentally, and found to be in accord with the stability criterion stated by Nyquist.

Nyquist's Criterion

The explicit solution of (1) for the p_k's demands an exact knowledge of the configuration of the amplifier and feed-back circuits. When the number of meshes is large, the solution involves much labor. If we wish simply to observe whether or not the system is stable, however, we need not obtain explicit solutions for the roots; in fact, all we need to know is whether or not any one of the p_k's has its real part positive. It turns out that when we know the transfer factor as a function of frequency, by calculation or by measurement, a simple inspection of the transfer factor polar diagram suffices for this purpose. This diagram is constructed by plotting the imaginary part of the transfer factor against the real part for all frequencies from minus to plus infinity.[8]

To obtain Nyquist's criterion we consider the vector drawn from the point (1, 0) to a point moving along the polar diagram; if the net angle which the vector swings through in traversing the curve is zero, the system is stable; if not, it is unstable. To express it in the terms used by Routh, if we set $1 - A(j\omega) = P + jQ$, and observe the changes of sign which the ratio P/Q makes when P goes through zero as the frequency steadily increases, the system is stable when there are the same number of changes from plus to minus as from minus to plus. It may be demonstrated that these two statements are equivalent.

The way in which the above procedures may be shown to reveal the existence of a root with positive real part may be outlined somewhat along the lines followed by Routh in his analysis.[9] Since p is a com-

[8] The transfer factor for negative frequencies $A(-j\omega)$ is the complex conjugate of that for positive frequencies $A(j\omega)$. Thus, if

$$A(j\omega) = X + jY,$$

then

$$A(-j\omega) = X - jY.$$

[9] A number of restrictions on the generality of the analysis may be noted. It is assumed that $A(p)$ has no purely imaginary roots, although the result in this case is otherwise evident. Further it is assumed that $A(p)$ goes to zero as $|p|$ becomes infinite, and that no negative resistance elements are included in the amplifier. Another point which should be mentioned is that the analysis does not apply to the stability in any conjugate paths that may exist. This point may be exemplified by the balanced tube or push-pull amplifier, in the normal transmission path of which the tubes of a stage act in series. When the series output is connected back to the series

plex quantity in general, any value which it may take is representable as a point on a plane—the p-plane of Fig. 2. Since only values of p with positive real parts concern us, attention may be confined to the right-hand half of the p-plane. Now draw a closed contour C in the right-hand half of the p-plane which encloses the root p_k. It is evident

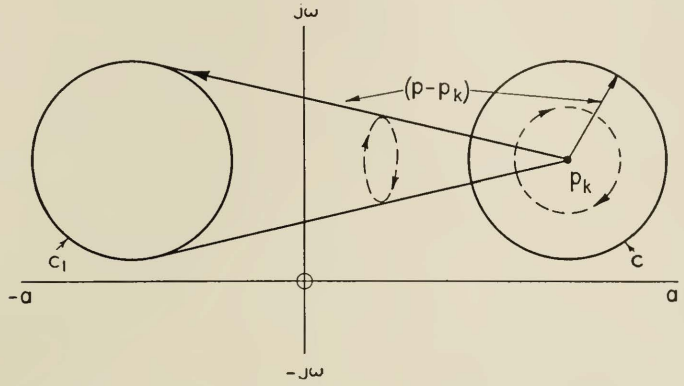


Fig. 2—Plot of two contours C and C_1 in the p-plane. C encloses the root p_k while C_1 does not enclose p_k. The vector $p - p_k$ covers 360 degrees as p traverses C, and covers the net angle of zero as p traverses C_1.

upon inspection of Fig. 2 that the vector extending from the root p_k to the contour makes a complete revolution (360 degrees) in following the closed path. If the contour does not enclose the root, however, as for C_1, then it is clear that when the vector from the root to the contour traverses the whole contour, the net angle turned through by the vector is zero. In the region under consideration we may write

$$f(p) = (p - p_k)\phi(p),$$

where $\phi(p)$ has no zeros within the contour. Hence, when p traverses a closed path and $(p - p_k)$ turns through 360 degrees or through zero the same angle is covered by $f(p)$. If for some different contour several roots are enclosed, it may be shown that $f(p)$ turns through one complete revolution for each of the enclosed roots when p traverses the contour.

In the form in which these considerations are stated, they are not suitable to practical application since complex values of p are involved. Ordinarily, of course, only imaginary values ($p = j\omega$) are conveniently accessible to us since it is a comparatively simple matter to

input, stability of the resultant loop has in general no bearing upon the stability of the path formed with the two tubes of each stage in parallel, since the series and shunt paths are conjugate to one another. To establish the stability of the shunt or parallel path, the transfer factor for that path must be separately determined. In general, the stability criterion applies only to the particular loop investigated, and not to any other existent loop.

measure the response with a sinusoidal impressed wave, but it would involve great difficulties of experiment as well as of interpretation to determine the response with negatively damped waves corresponding to values of p in the right-hand half of the plane. However, these results may be brought within the field of practical experience by a procedure widely used for the purpose.

To include all roots in the right-hand half of the p-plane, the contour must be taken of infinite extent. The path ordinarily followed for this purpose extends from the value $+ R$ to $- R$ on the imaginary axis, and is closed by a semicircle of radius R, where R is assumed to expand without limit. It may be noted that in actual amplifier circuits the transfer factor becomes zero when $|p|$ becomes infinite, so that $A(p)$ is zero along the semicircular part of the closed contour. Consequently, the only values of $A(p)$ which differ from zero are those corresponding to finite values of p, along the imaginary axis. In other words, the plot of $A(p)$ under these conditions comes down to the plot of $A(j\omega)$ where ω is finite. Hence, if we plot $A(j\omega)$ for all values of ω from minus to plus infinity, there will be no roots with positive real parts and the system will be stable when the vector from (1, 0) to the curve sweeps through a net angle of zero. The system will be unstable when the vector sweeps through 360 degrees, or an integral multiple thereof.

Two types of transfer factor curves may be considered as illustrations. The first of these shown in Fig. 3 corresponds to that for a re-

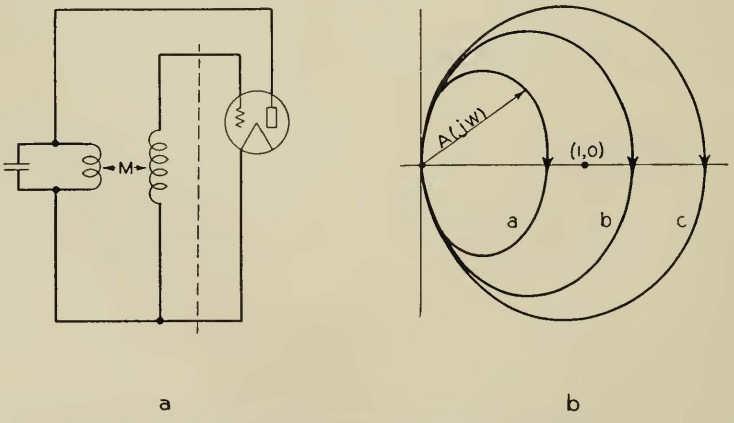


Fig. 3—Schematic of a reversed feed-back oscillator circuit at the left. At the right plot of the transfer factor $A(j\omega)$ around the feed-back loop of Fig. 3a over the frequency range from zero to very high frequencies. The imaginary part of the transfer factor is plotted as ordinate against the real part as abscissa for the three curves *a*, *b*, *c*, which correspond to increasing gains around the loop. Condition *a* is stable, while *b* and *c* are unstable.

versed feed-back oscillator circuit, the three curves marked *a*, *b*, *c*, corresponding to progressively increasing gains around the loop. It will be observed that after the maximum gain has reached and exceeded unity, that the circuit is unstable, since the point (1, 0) is then enclosed. This state of affairs may be contrasted with that existing in the particular form of feed-back circuit to which Fig. 4 applies. Again the

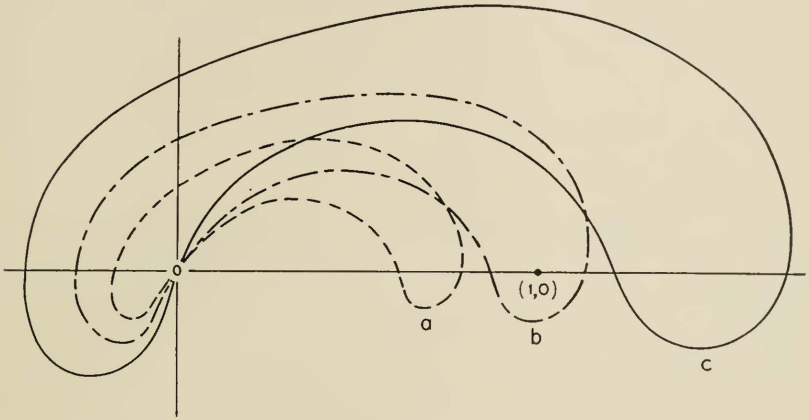


Fig. 4—Transfer factor diagram for a particular form of feed-back circuit, curves *a*, *b*, *c*, corresponding to increasing gains around the feed-back loop. Conditions *a* and *c* are stable, *b* is unstable.

three curves *a*, *b*, *c*, correspond to progressively increasing gains around the feed-back loop. As the gain is increased the system is first stable (*a*), then unstable (*b*), and finally stable (*c*), since it is only within curve (*b*) that the point (1, 0) is enclosed. This striking example is the one which was investigated experimentally. The methods used in determining the transfer factor diagram form the subject of the next section.

Measuring Methods

Application of the Nyquist stability criterion requires the determination of the vector transfer factor around the feed-back loop at all frequencies. This is usually effected by opening the circuit at any point which provides convenient impedances looking in both directions from the break. These points are then connected to an oscillator and to suitable measuring circuits, which are to be described. Care must be taken to ensure that the oscillator and measuring circuit impedances are equal to the output and input impedances respectively of the circuit under test. This precaution is necessary in order that the transfer factor in the measuring condition may not differ significantly from that existing in the operating condition.

Two methods of measurement have been found useful. The first is a null method capable of good precision over a wide frequency range. The second is a visual method in which the transfer factor polar diagram is traced on the screen of a cathode ray oscillograph. This method is not capable of very great precision and, in the model used, the frequency range is somewhat restricted. However, it permits of a rapid survey of the situation for which its precision is adequate, before proceeding with the slower and more precise measurements of the null method, where the latter are required. By making such a preliminary survey the critical frequency ranges can be mapped out for precise measurement, thereby eliminating a large amount of unnecessary labor.

Null Method

In the more precise measurements extending over a wide frequency range, special care is required to ensure freedom from errors in the measurement of phase angles and amplitudes. Much of the difficulty associated with direct measurement over wide frequency ranges is avoided by the use of a simple demodulation scheme. In this scheme, the potentials to be compared are modulated down to a fixed frequency (in actual use 1000 cycles) regardless of the frequency at which the test is being made. In this way a minimum portion of the circuit carries the high frequency. Further this permits the use of voice frequency attenuators, phase shifters, and amplifiers which in fact require calibration at only a single frequency.

In this arrangement, as shown in Fig. 5, demodulators are shunted across the input and output terminals of the circuit under test. A single oscillator supplies the carrier to both demodulators, its frequency differing by 1000 cycles from the frequency supplied to the circuit under test. The demodulated outputs are connected through attenuators and phase shifters to a common amplifier detector. The attenuators and phase shifters are adjusted until the detector gives a null reading. When this condition obtains the difference in the attenuator settings in the two branches is equal to the gain or loss of the circuit under test, and the difference in the phase shifter settings is either equal to or the negative of the phase shift of the circuit under test. To show this, denote the amplifier output voltage by $P_0 \cos (2\pi ft - \phi)$, and the beat frequency voltage supplied to the demodulators by $P \cos 2\pi(f \pm 1000)t$. The demodulated output, proportional to the product of the two applied waves, is then

$$PP_0 \cos (2\pi \cdot 1000t \mp \phi).$$

Correspondingly, the demodulated output from the other demodulator

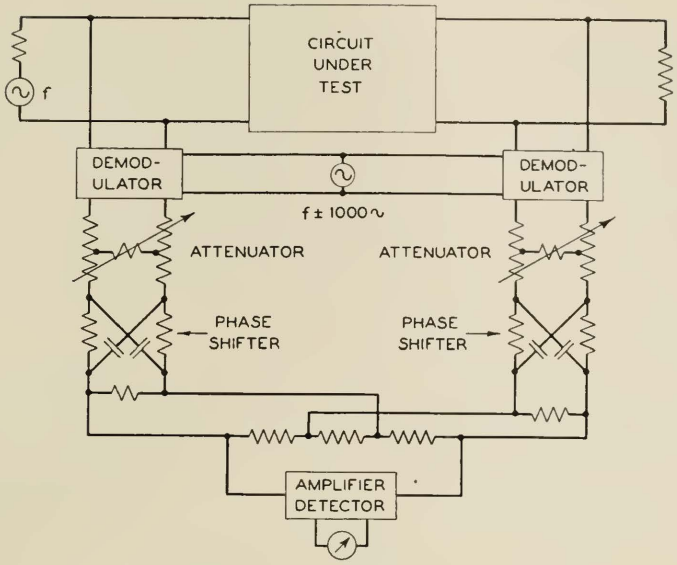


Fig. 5—Schematic diagram of the null method used to measure the transfer factor.

connected across the input is given by

$$PP_i \cos (2\pi \cdot 1000t).$$

If now these two waves are to be made to cancel, there must be a difference in the attenuation of the two branches equal to the ratio P_0/P_i, and a difference in the phase shift equal to $\mp \phi$. The change in sign of the phase angle introduced by setting the beat oscillator above or below the test frequency is most conveniently handled by setting the carrier oscillator consistently on the same side of the test frequency in making a run over the frequency range.

By using a high gain amplifier preceding the detector, the precision may be made great, limited only by circuit noise and by interference. The attenuators and phase shifters are calibrated separately. It should be noted that any difference in the transfer constants of the two demodulator circuits may be compensated by an initial adjustment which is carried out by paralleling the input terminals of the two demodulators across a source of electromotive force. With the particular type of phase shifter used the phase shift may be changed without altering the attenuation, so that the two settings for amplitude and phase may be made independently.

Visual Method

In the visual method of observation, a steady potential proportional to the inphase component of the transfer factor is impressed across one pair of plates of a cathode ray oscillograph and another steady potential proportional to the quadrature component is impressed across the other pair of plates, the constant of proportionality being the same for the two components. In this way the transfer factor at any frequency appears as a single point, the vector from the origin to the displaced beam constituting the transfer factor. The locus of all these points, i.e., vector tips, over the frequency range constitutes the transfer factor polar diagram.

To provide rectified potentials proportional to inphase and to quadrature components respectively, use is made of the properties of the so-called vacuum tube wattmeter.[10] As used in practice, this device consists of two triodes in push-pull connection (Fig. 6), the series arm of the grid circuit being connected to the unknown potential, and the

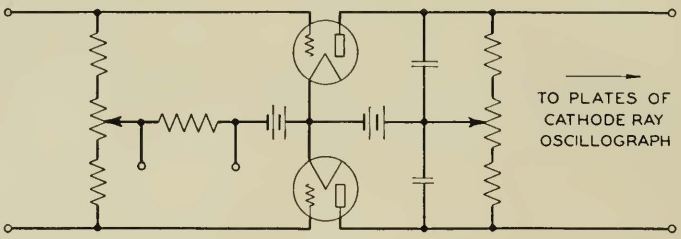


Fig. 6—Circuit of a vacuum tube wattmeter used to provide a rectified potential proportional to the product of the two impressed grid potentials (both of the same frequency) multiplied by the cosine of the phase angle between them.

shunt arm of the grid circuit being connected to a source of the same frequency but of standard phase. Under these conditions the rectified output in the plate circuit flowing in series with the two plates is proportional to the product of the two impressed voltages multipled by the cosine of the angle between them.

As shown in Fig. 7, two separate wattmeters are employed, one for each phase, their series input terminals being connected together across the output of the circuit under test. To the common branch of one of these wattmeters is supplied the same potential as is fed to the input of the circuit under test. The rectified output of this wattmeter therefore is proportional to the product of the input and output voltages multiplied by the cosine of the transfer factor phase angle. This po-

[10] U. S. Patent 1,586,533; Turner and McNamara, *Proc. I. R. E.*, vol. 18, p. 1743; October (1930).

tential is supplied to those plates of the oscillograph which produce a horizontal deflection. To the common branch of the other wattmeter is applied a potential equal in amplitude to the input voltage but lagging behind it by 90 degrees. The rectified output of this wattmeter is proportional to the product of input and output voltages multiplied by the cosine of the transfer factor phase angle minus 90 degrees, or in other words proportional to the sine of the transfer factor phase angle. This voltage is supplied to those plates of the oscillograph which produce a vertical deflection. We have then across one pair of plates of the oscillograph a steady potential proportional to the real component

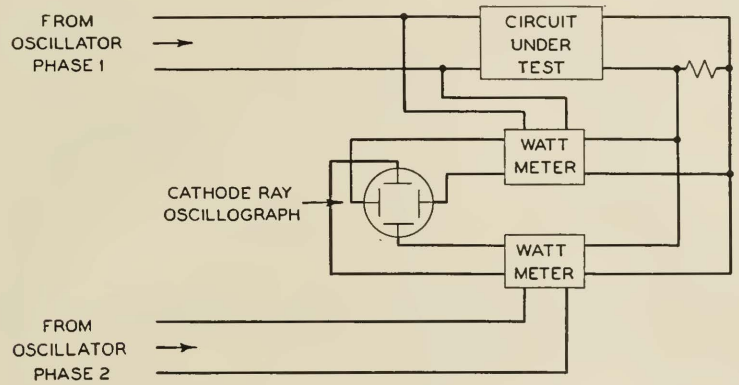


Fig. 7—Schematic diagram of the circuit used to plot the transfer factor diagram on the screen of a cathode ray oscillograph.

of the transfer factor, and across the other pair of plates we have impressed a steady potential proportional to the imaginary component of the transfer factor. These two components act upon the beam of the oscillograph to produce a deflection which in amplitude and in phase is the resultant of the two component deflections and so corresponds to the transfer factor.

It will be observed that the above procedure requires a two-phase source of constant amplitude, the frequency of which is variable over the range necessary to establish the properties of the amplifier. In the present instance the frequency range extends from 0.5 to 30 kilocycles, and the accuracy required is of the order of five per cent.

A schematic of the two-phase oscillator used is shown in Fig. 8. This oscillator is of the heterodyne type. Two independent sources are used, one of constant frequency (100 kilocycles), the other variable in frequency and practically constant in amplitude over the range of 100 to 130 kilocycles. As indicated in the figure, the variable fre-

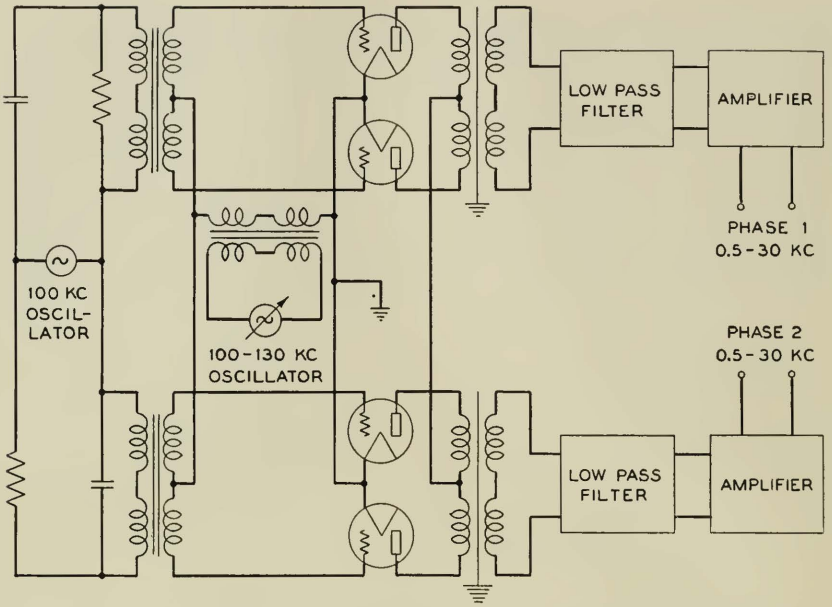


Fig. 8—Circuit diagram of a heterodyne type two-phase oscillator, the output frequency of which is continuously variable from 0.5 to 30 kilocycles. The output of each phase and the 90-degree difference between the two phases are practically constant over the frequency range.

quency oscillator is connected to the common branches of the two push-pull modulators. The fixed frequency oscillator is connected in series with the grid circuits of the two modulators. The resistance-capacity networks shown in the circuits of the fixed frequency oscillator are provided to produce phase shifts of 90 degrees between the two series voltages of the two modulators. In the same manner as that discussed before in connection with the null method measuring circuit, the phase shift introduced to the fixed frequency is maintained in the beat frequency output, so that the phase difference of 90 degrees is preserved in the outputs of the two modulators when the variable frequency oscillator goes from about 100.5 to 130 kilocycles. The outputs of the two phases are connected to the test amplifier and to the wattmeters as shown in the preceding Fig. 7.

Comparison of the Methods

Measurements of transfer factors by the two methods outlined above were found to be in agreement within the error of measurement. The visual method as developed was capable of use over only a very re-

stricted frequency range as compared to the null method, but it covered the region of particular interest in the experiments conducted for the purpose of testing the stability criterion. Through its use, measurements over its frequency range could be made in a few minutes time, whereas corresponding measurements by the more precise null method required three to six hours. Of course the time intervals cited do not include time occupied in setting up and adjusting the apparatus.

Test Amplifier and Experimental Results

Test Amplifier

The stability criterion indicates three distinct conditions of interest, one of which is unstable, the other two being stable. The unstable condition (1) is that in which the transfer factor curve encloses the point (1, 0). Two stable conditions are those in which (1, 0) is not enclosed by the curve, but in which (2) the curve crosses the zero phase shift axis at points greater than unity, and (3) the curve does not cross the zero phase shift axis at points greater than unity. Condition (2) is of particular interest because while it is judged stable on the basis of Nyquist's criterion it would appear to be unstable on the basis of the older transfer criterion discussed in the first and second sections.

For test purposes an amplifier was designed which, upon variation of an attenuator in the feed-back path, would satisfy each of the three above conditions in turn. The amplifier schematic is shown in Fig. 9.

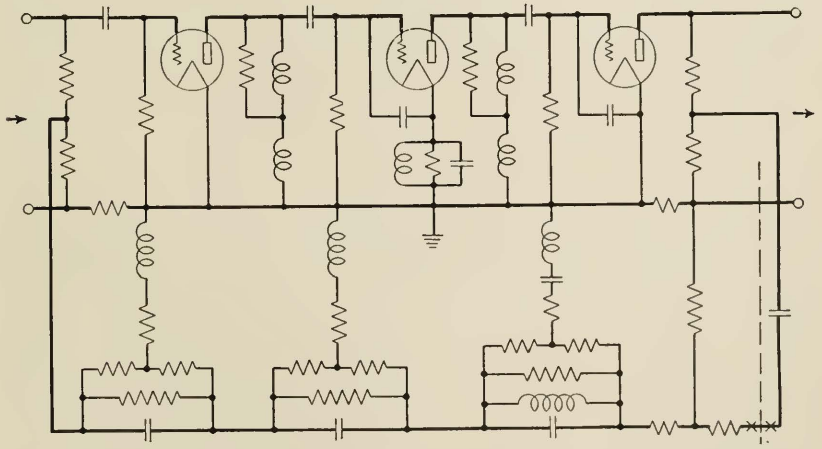


Fig. 9—Circuit diagram of the feed-back amplifier used in testing the stability criterion. The dashed line indicates the point at which the loop was broken for measurement of the transfer factor. At the left of this line is shown the resistance attenuator provided to vary the gain around the feed-back loop.

It has three stages, the first two tubes being space charge grid pentodes, and the last one a triode. The interstage coupling circuits were made up of simple inductances and resistances as shown. The amplifier was designed by E. L. Norton and E. E. Aldrich to provide a transfer factor characteristic having the desired shape, i.e., a loop crossing the zero phase axis in the neighborhood of 10 kilocycles. It will be observed that the feed-back circuit is connected between bridge networks in both input and output circuits, which were provided to eliminate reaction of the input and output circuits upon the feed-back network.[11]

Experimental Results

The transfer factor was measured for a zero setting of the feed-back attenuator over a frequency range of 0.5 to 1200 kilocycles. The results are shown in Fig. 10. The method of plotting this figure requires some discussion. In order to keep the curve within a reasonable size and still show the necessary details the scale has been made logarithmic by plotting the gain around the loop in decibels instead of the corresponding numerical ratios. It is of course impossible to carry this out completely on a polar diagram since the transfer factor goes to zero at high frequencies. To take care of this the scale is made logarithmic only above zero gain, corresponding to unit transfer ratio, and is linear below. It should be noted that if the logarithmic portion of the scale is translated outward so that the zero decibel point lies successively in the regions marked A, B, C, and D, the indicated amplifier conditions correspond to those designated above as (1), (2), (1) and (3) respectively. Experimentally an increase of the feed-back attenuator corresponds to such a translation of the logarithmic scale by an amount equal to the increase in attenuation. Therefore, the transition from one condition to another should occur when the attenuator setting is equal to the gain at a zero phase point in the curve as measured with a zero attenuator setting.

The test of the stability criterion consists of a determination of the attenuator settings at which oscillations begin, and a comparison of these settings with those at which a transition from a stable to an unstable condition is predicted by the theory. Experimentally oscillations were found to occur in regions A and C and not in regions B and D which is in qualitative agreement with Nyquist's predictions. Quantitatively the measured and predicted transition points agreed within one decibel which is estimated to be within the experimental error.

It should be noted that the plotted curve has been drawn up for $A(j\omega)$, no points of $A(-j\omega)$ being shown, although both are required

[11] H. S. Black, *Bell Sys. Tech. Jour.*, January, 1934.

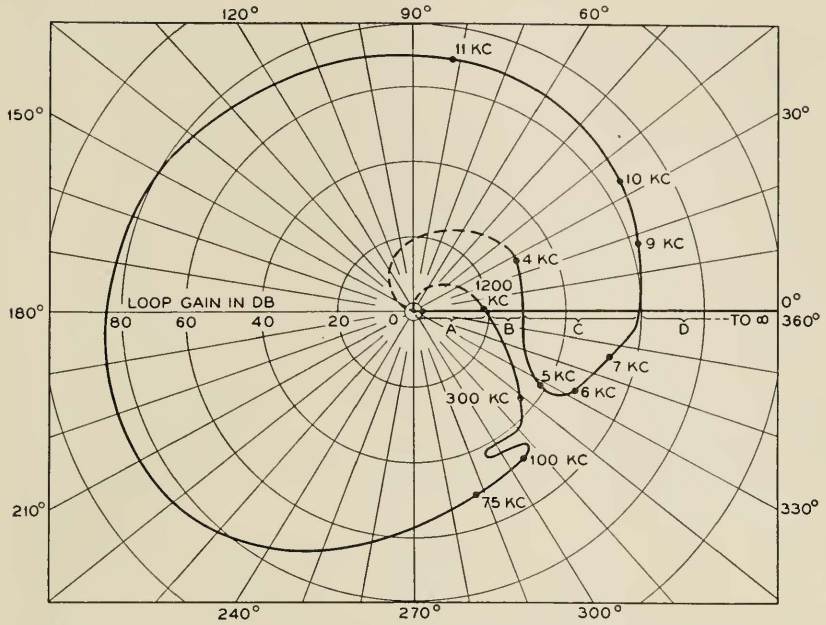


Fig. 10—Transfer factor diagram for the amplifier of Fig. 9 with the feed-back attenuator set at zero decibel.

by the theoretical derivation. Where the transfer factor is zero at zero frequency, only $A(j\omega)$ is required since the loop then closes for positive values of ω. In amplifiers transmitting d.c., however, both positive and negative values of ω are needed [8] to form a closed loop. In any case $A(-j\omega)$ is the mirror image of $A(j\omega)$ about the x-axis.

Extensions of the Criterion

Nonlinear Amplifier

The stability criterion which was verified by the experiments reported in the preceding section is framed for linear systems, those in which the steady state response is linearly proportional to the applied force. In vacuum tube circuits, linearity is best approximated at small force amplitudes, and is departed from to an extent dependent upon the impressed potentials, as well as upon tube and circuit characteristics. The divergence from linearity becomes well marked when the load capacities of the tubes are approached, or when grid current is made to flow through large grid impedances. The question then arises as to the form which the stability criterion takes when a tube circuit is

[8] Loc. cit.

operated in a nonlinear region—let us say by impressing upon the circuit a sufficiently large alternating potential provided by an external independent generator.

To answer this question we may consider the response of the amplifier, loaded by the independent generator, to a small alternating potential introduced for test purposes. Since the response of the system is known to be linear from the theory of perturbations, we might attempt to apply the linear criterion to the small superposed force. To do this it is necessary to measure the transfer factor for the small superposed force over the frequency range at a particular load of interest.

Application of the experimental technique to this extended criterion introduces difficulties since the opening of the feed-back loop for measuring purposes disturbs to a certain extent the distribution of these loads, particularly the harmonics, and modulation products in general. This makes it difficult to get the same loading effect when the loop is opened for measuring purposes as obtained when the loop is closed. Another consideration is that the response to the small component may be expected to vary in general at different points on the loading wave, so that the measuring procedure averages the response over a cycle of the loading wave. A method of measurement analogous to that of the flutter bridge would be required to evaluate the transfer factor at points of the loading cycle. Further, the measuring apparatus is affected by the presence of the loading currents when these are sufficiently large. In the present case in which the loading frequency (60 kilocycles) was far removed in the frequency scale from the test frequencies, it was found possible to approximate the necessary measurements by the insertion of selective circuits.

The curves of Fig. 11 represent portions of the transfer factor polar diagram for an amplifier similar to the one previously described, measured by the visual method with different loading amplitudes. The effect of the load on this particular amplifier is to change both phase shift and amplitude so that the curves shrink both radially and tangentially, pulling the loop back across the zero phase axis until, at the heaviest load, the two low-frequency crossings are completely eliminated. If the extended criterion is valid, we should expect the amplifier to be stable at any setting of the feed-back attenuator. As the load is decreased from this value, the crossings occur at successively higher gains so that the start of oscillations would occur at progressively higher settings of the feed-back attenuator.

The curves of Fig. 12 show the attenuator settings predicted by the extended criterion and those determined by direct observation of the attenuator setting required for oscillations when the feed-back circuit

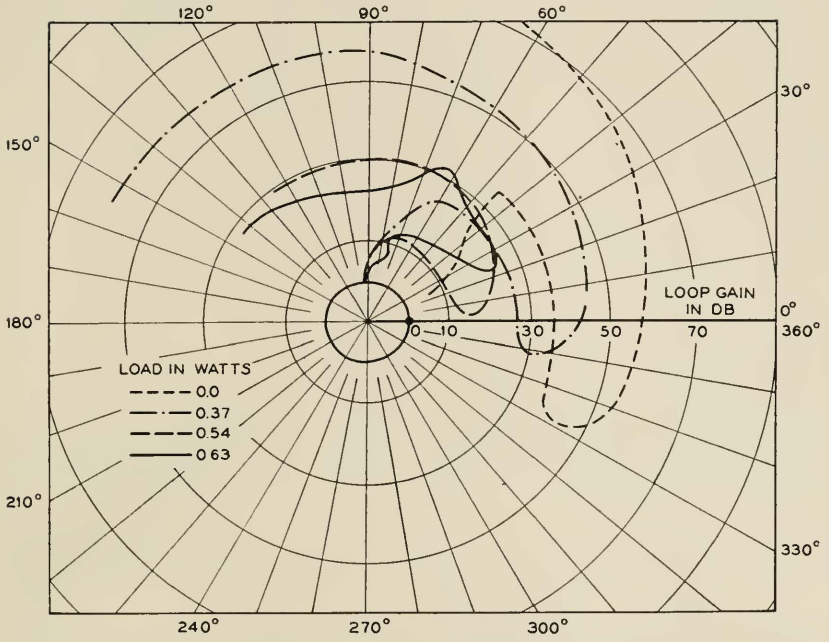


Fig. 11—The transfer factor diagram for the amplifier of Fig. 9 with the feed-back attenuator set at zero decibel. The four curves shown correspond to different amounts of the 60-kilocycle load.

was closed. Two sets of curves are shown, one for each of the low-frequency crossings. These are plotted against the loading amplitude. The agreement between the experimental and predicted values is close for the higher gain crossing at small loading amplitudes, but a divergence is apparent at high loads. For the lower crossing there is a divergence of 1.5 decibels at low loads, which changes sign and becomes greater at the higher loads. These divergences may be ascribed to a variety of causes among which probably the most important are the effects of harmonics upon the amplifier loading, overloading of the measuring apparatus by harmonics of the loading electromotive force, and phase shifts introduced by the selective circuits. The last two causes may be eliminated by improved technique, but the first cause in general introduces a fundamental difficulty, particularly important when large nonlinearities are involved.

Negative Impedances

One of the early forms of stability criterion mentioned in the first section was that relating to the measured impedance of the circuit.

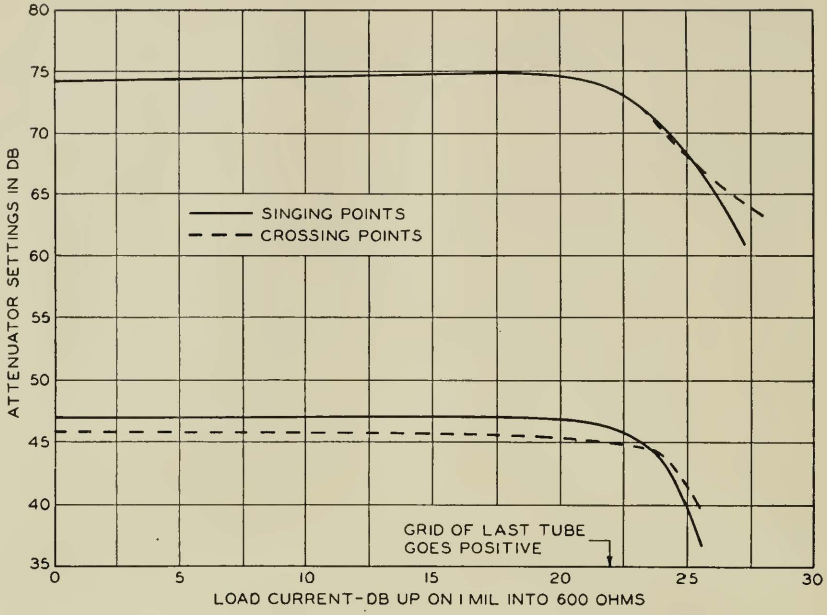


Fig. 12—Comparison of feed-back attenuator settings required for the starting of oscillations, and those deduced from the transfer diagram, plotted as functions of the 60-kilocycle load. The two dashed curves correspond to the two points (roughly 4.5 and 8 kilocycles) at which the transfer factor diagram (Fig. 11) crosses the zero phase axis. The gains of Figs. 11 and 12 cannot be compared directly because of a change made in the amplifier circuit of Fig. 12 which increased the loop gain.

Nyquist's criterion involving the transfer factor may be transformed so as to formulate a more complete criterion involving such an impedance.

To do this we have to express the factor $(1 - A)$, on which the stability criterion was based, in terms of the circuit impedances. For illustrative purposes we may quote the results obtained with the two fundamental forms of feed-back circuits, the series and shunt types.[12] These results, while obtained for the input circuit of the amplifier, are valid for any other point of the feed-back loop. Further, combinations of the shunt and series type feed-back circuits may be used.

Series Feed-Back

The series circuit is shown in Fig. 13, so called because the feed-back is applied in series with the amplied electromotive force and the amplifier input. The passive impedances marked are those existing when the feed-back loop is broken and terminated as indicated by the

[12] Crisson, *Bell Sys. Tech. Jour.*, vol. X, p. 485.

dotted lines. By direct circuit analysis, the current and voltage amplitudes in the feed-back condition are related by

$$E = (Z + Z_0 + Z_i)(1 - A)I,$$

where A and the Z's are functions of frequency. The total effective circuit impedance is obtained as the multiplifier of I in the right mem-

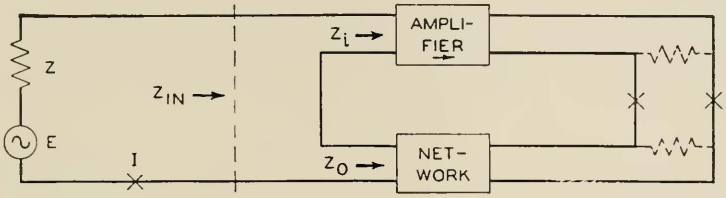


Fig. 13—Series type feed-back circuit. The dotted resistances indicate the terminations applied when the feed-back circuit was broken, to which the passive impedances (Z_iZ_0) apply. Z_{in} represents the effective input impedance with the feed-back circuit connected through.

ber. Subtracting the generator impedance Z from the total, the input impedance becomes

$$Z_{IN} = (Z_0 + Z_i)(1 - A) - AZ,$$

from which,

$$1 - A = \frac{Z}{Z + Z_0 + Z_i}\left(1 + \frac{Z_{IN}}{Z}\right).$$

Of the two factors of the right member, the first one, involving passive impedances alone, can have no roots with positive real part. Any such roots must, therefore, be contained in the second bracketed factor and then only when Z_{in} is negative. Hence paraphrasing the transfer factor criterion, if we plot $-Z_{in}/Z$ over the frequency range, the circuit is stable when the point (1, 0) is not enclosed by the resultant curve.

Shunt Feed-Back

Proceeding as in the series case with the circuit of Fig. 14 we get

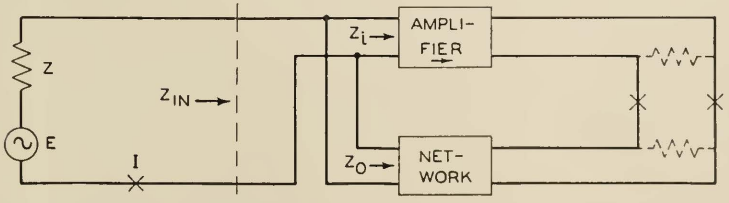


Fig. 14—Shunt type feed-back circuit. The notation corresponds to that of Fig. 13.

$$1 - A = \frac{Z_a}{Z + Z_a}\left(1 + \frac{Z}{Z_{in}}\right),$$

where Z_a represents the impedance of Z_0 and Z_i in parallel. Again only the bracketed term can yield undamped transients so that the criterion involves plotting $-Z/Z_{in}$ over the frequency range; if the resultant curve does not enclose (1, 0) the circuit is stable.

It may be remarked that these results are applicable to circuits including two-terminal negative impedances such as the oscillating arc and the dynatron, which are of the series and the shunt type respectively.

ACKNOWLEDGMENTS

The authors are indebted to Mr. L. W. Hussey for discussions of theoretical points, and to Mr. P. A. Reiling for his cooperation in the experiments.

Abstracts of Technical Articles from Bell System Sources

Shared Channel Broadcasting.[1] C. B. AIKEN. This paper deals with the experimental studies made on the character and causes of interference noticeable in shared channel broadcasting, such as heterodyning, flutter, sideband interference and wobbling. Valuable data are included on the characteristics of square-law and linear detectors anent to interference.

The Determination of Dielectric Properties at Very High Frequencies.[2] J. G. CHAFFEE. A simple method of determining the dielectric constant and power factor of solid dielectrics at frequencies as high as 20 megacycles, with an accuracy which is sufficient for most purposes, is described. The major sources of error are discussed in detail, and several precautions which should be observed are pointed out.

Measurements of the dielectric properties at 18 megacycles of a number of commonly used materials have shown that in general the power factor and dielectric constant are not widely different from those which obtain at frequencies of the order of one megacycle.

In addition, the results of an investigation of the input impedance of vacuum tube voltmeters at high frequencies are described as an illustration of the further application of this method of measurement.

Optical Factors in Caesium-Silver-Oxide Photoelectric Cells.[3] H. E. IVES AND A. R. OLPIN. This paper describes an investigation of the part played by the angle of incidence and state of polarization of the exciting light in producing the enhanced or selective emission of photoelectrons in the red region of the spectrum which is characteristic of photoelectric cells made by treating a silver surface with oxygen and caesium vapor (Fig. 1). This question is one which has been raised in connection with all types of photoelectric cells having composite surfaces and which exhibit spectrally selective emission. It has thus been an open question whether the selective peaks in the spectral response curves exhibited by the alkali hydride cells are to be ascribed to an enhanced effect of the perpendicular vector of obliquely incident

[1] *Radio Engineering*, June, 1934.
[2] *Proc. I. R. E.*, August, 1934.
[3] *Jour. Op. Soc. Am.*, August, 1934.

radiation, or whether the spectral selectivity is in the nature of a locally intrinsic emissive power, such as would be caused by an optical absorption band or an electronic transmission band. In order to answer this question, it is necessary to have emitting surfaces of a specular character. Such surfaces have not been prepared with the alkali hydrides, but it has been found possible to make the caesium-silver-oxide cells on specular plates of silver so that they retain their specular character in the final sensitized surface. Cells of this sort were used in this study, and have made possible a clear separation of the emissive singularities due to optical conditions and the singularities which may be described as intrinsic to the material.

Both from their method of preparation and from their optical behavior, we have felt justified in considering the caesium-silver-oxide photoelectric cells prepared with specular silver surfaces as consisting of silver surfaces overlaid with a thick layer of transparent refracting material, on the top of which is a thin photosensitive layer. The silver plates, after oxidation, exhibit interference colors, the exact color depending upon the amount of oxidation. Viewed at an angle through a nicol prism, these oxidized plates exhibit the well-known properties of thin refractive layers on a metal base. Thus when the plane of polarization is changed from the plane of incidence to the plane perpendicular thereto, no change of hue takes place for small angles of incidence; but at large angles, the color changes to a complementary hue. After the silver oxide surface has been exposed to caesium vapor and given a heat treatment, these optical properties are still usually observable, but degraded. The softening of the interference colors may be due either to a change in thickness of the refracting medium as caesium oxide is formed or to the introduction of a general body color. In a few less common cases the colors faded out completely, the plate at the end of the heat treatment being metallic in appearance yet still exhibiting a pronounced selective response to red and infrared light.

The behavior of a thin photoelectric sheet separated from a specular metal surface by a layer of refracting medium has been treated in an earlier paper where a layer of caesium was deposited on the top of a quartz-coated platinum plate. The data obtained in this earlier paper are immediately applicable to the present problem, granting the similarity of conditions which we have assumed. It has been convenient to pursue this present study on the assumption of such a similarity and to arrive at conclusions from the agreement with, or deviation from, the results obtained from the simpler materials and conditions previously studied.

Phase Angle of Vacuum Tube Transconductance at Very High Frequencies.[4] F. B. LLEWELLYN. Theoretical considerations indicate that the transconductance of a vacuum tube exhibits a phase angle when the transit time of electrons from cathode to anode becomes an appreciable fraction of the high-frequency period. Measurements show that such a phase angle actually occurs and that its behavior is in general agreement with the theoretical predictions.

Application of Sound Measuring Instruments to the Study of Phonetic Problems.[5] JOHN C. STEINBERG. This paper gives the results of a period by period analysis of the vowel sound waves occurring when the sentence "Joe took father's shoe bench out" was spoken. Such an analysis gives an approximate picture of the time variations in r.m.s. amplitude of the wave, frequency of voice fundamental, and frequency regions of overtone reenforcement. Although the study is confined to a few sounds and one speaker's voice, it illustrates a method of approach to studies of speech production and measurement.

[4] *Proc. I. R. E.*, August, 1934.
[5] *Jour. Acous. Soc. Am.*, July, 1934.

Contributors to this Issue

KARL K. DARROW, B.S., University of Chicago, 1911; University of Paris, 1911–12; University of Berlin, 1912; Ph.D., University of Chicago, 1917. Western Electric Company, 1917–25; Bell Telephone Laboratories, 1925–. Dr. Darrow has been engaged largely in writing on various fields of physics and the allied sciences.

LLOYD ESPENSCHIED. Mr. Espenschied is High Frequency Transmission Development Director in the Bell Telephone Laboratories. He joined the Bell System in 1910, having graduated from Pratt Institute the previous year. He has taken an important part in practically all of the Bell System radio developments, beginning with the first long-distance radio-telephone tests of 1915, at which time he received the voice in Hawaii from Arlington, Virginia. He has participated in a number of international conferences on electric communications.

J. G. KREER, B.S. in Electrical Engineering, University of Illinois, 1925; M.A., Columbia University, 1928. Bell Telephone Laboratories, 1925–. Mr. Kreer has been engaged in research work on carrier frequency systems.

S. A. LEVIN, E.E., Chalmers Technical Institute, Gothenburg, 1919; Technische Hochschule, Berlin, 1920–21; Technische Hochschule, Dresden, 1921–23. Radio Department, General Electric Company, Schenectady, N. Y., 1923–26; Engineering Department, National Electric Light Association, New York, N. Y., 1926–30. Bell Telephone Laboratories, 1930–. Mr. Levin's work has to do with the development of high-frequency measuring equipment for carrier systems.

G. L. PEARSON, A.B., Willamette University, 1926; M.A. Stanford University, 1929. Bell Telephone Laboratories, 1929–. Mr. Pearson has been engaged in a study of the noise inherent in electric circuits.

D. B. PENICK, B.S. in Electrical Engineering, University of Texas, 1923; B.A., 1924; M.A. in Physics, Columbia University, 1927. Western Electric Company, Engineering Department, 1924–25; Bell Telephone Laboratories, 1925–. Mr. Penick has been engaged in special problems related to the development of vacuum tubes.

E. Peterson, Cornell University, 1911–14; Brooklyn Polytechnic, E.E., 1917; Columbia, A.M., 1923; Ph.D., 1926; Electrical Testing Laboratories, 1915–17; Signal Corps, U. S. Army, 1917–19. Bell Telephone Laboratories, 1919–. Dr. Peterson's work has been largely in theoretical studies of carrier current apparatus.

Liss C. Peterson, E.E., Chalmers Technical Institute, Gothenburg, 1920; Technische Hochschule, Charlottenburg, 1920–21; Technische Hochschule, Dresden, 1921–22; Signal Corps, Swedish Army, 1922–23. American Telephone and Telegraph Company, 1925–30; Bell Telephone Laboratories, 1930–. Mr. Peterson is engaged in the study of modulation and other problems connected with high frequency carrier systems.

S. A. Schelkunoff, B.A., M.A., in Mathematics, The State College of Washington, 1923; Ph.D. in Mathematics, Columbia University, 1928. Engineering Department, Western Electric Company, 1923–25. Bell Telephone Laboratories, 1925–26. Department of Mathematics, State College of Washington, 1926–29. Bell Telephone Laboratories, 1929–. Dr. Schelkunoff has been engaged in mathematical research, especially in the field of electromagnetic theory.

M. E. Strieby, A.B., Colorado College, 1914; B.S., Harvard, 1916; B.S. in E.E., M.I.T., 1916; New York Telephone Company, Engineering Department, 1916–17; Captain, Signal Corps, U. S. Army, A. E. F., 1917–19. American Telephone and Telegraph Company, Department of Development and Research, 1919–29; Bell Telephone Laboratories, 1929–. Mr. Strieby has been associated with various phases of transmission work, more particularly with the development of long toll circuits. At the present time, in his capacity as Carrier Transmission Research Engineer, he directs studies of new and improved methods of carrier frequency transmission over existing or new facilities.

L. A. Ware, B.E., Engineering College, University of Iowa, 1926; M.S., University of Iowa, 1927; PhD., Physics Department, University of Iowa, 1930. Instructor in Physics, University of Iowa, 1926–29. Bell Telephone Laboratories, 1929–. Dr. Ware's work has been chiefly in connection with regenerative amplifier development.