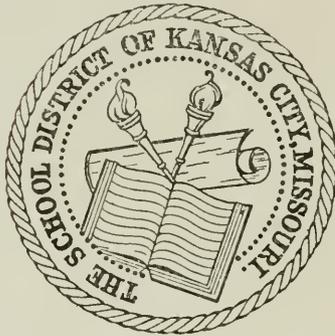


Bound
periodical

1001.959

Kansas City
Public Library



This Volume is for
REFERENCE USE ONLY

7-38-6m-P

PUBLIC LIBRARY
KANSAS CITY
MO

From the collection of the

o P^{z n m}re^ainger
v L^{t p}ibrary

San Francisco, California
2008

YHABILLI OLURUN
YTIQ ZACMAN
OH

PUBLIC LIBRARY
KANSAS CITY

THE BELL SYSTEM TECHNICAL JOURNAL

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

EDITORIAL BOARD

F. B. JEWETT

A. F. DIXON

S. BRACKEN

H. P. CHARLESWORTH

O. E. BUCKLEY

M. J. KELLY

W. WILSON

W. H. HARRISON

O. B. BLACKWELL

G. IRELAND

R. W. KING, *Editor*

J. O. PERRINE, *Associate Editor*

TABLE OF CONTENTS

AND

INDEX

VOLUME XVIII

1939

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

VIA AIR MAIL
YTD 2024
ON

Sound
Periodical

PRINTED IN U. S. A.

1001959

FE 9 '40

THE BELL SYSTEM TECHNICAL JOURNAL

VOLUME XVIII, 1939

Table of Contents

JANUARY, 1939

Electrostatic Electron-Optics— <i>Frank Gray</i>	1
Equivalent Modulator Circuits— <i>E. Peterson and L. W. Hussey</i> ..	32
An Improved Three-Channel Carrier Telephone System— <i>J. T. O'Leary, E. C. Blessing and J. W. Beyer</i>	49
Crossbar Dial Telephone Switching System— <i>F. J. Scudder and J. N. Reynolds</i>	76
A Twelve-Channel Carrier Telephone System for Open-Wire Lines — <i>B. W. Kendall and H. A. Affel</i>	119
Recent Developments in the Measurement of Telegraph Trans- mission— <i>R. B. Shanck, F. A. Cowan and S. I. Cory</i>	143
Contemporary Advances in Physics, XXXII. Particles of the Cosmic Rays— <i>Karl K. Darrow</i>	190
Hurricane and Flood—September 1938— <i>W. H. Harrison</i>	218
A Terrain Clearance Indicator— <i>Lloyd Espenschied and R. C. Newhouse</i>	222
Transcontinental Telephone Lines— <i>J. J. Pilliod</i>	235

APRIL, 1939

Some Ceramic Manufacturing Developments of the Western Electric Company— <i>A. G. Johnson and L. I. Shaw</i>	255
The Production of Ultra-High-Frequency Oscillations by Means of Diodes— <i>F. B. Llewellyn and A. E. Bowen</i>	280
A Representation of the Sunspot Cycle— <i>C. N. Anderson</i>	292
The Number of Impedances of an n Terminal Network— <i>John Riordan</i>	300
Copper Oxide Modulators in Carrier Telephone Systems— <i>R. S. Caruthers</i>	315

Some Applications of the Type "J" Carrier System—
L. C. Starbird and J. D. Mathis 338

Line Problems in the Development of the Twelve-Channel Open-
 Wire Carrier System—
L. M. Ilgenfritz, R. N. Hunter and A. L. Whitman 363

JULY, 1939

Frequency-Modulation: Theory of the Feedback Receiving Circuit
 —*John R. Carson* 395

The Application of Negative Feedback to Frequency-Modulation
 Systems—*J. G. Chaffee* 404

Survey of Magnetic Materials and Applications in the Telephone
 System—*V. E. Legg* 438

Impedance Properties of Electron Streams—*Liss C. Peterson* 465

Plastic Materials in Telephone Use—
J. R. Townsend and W. J. Clarke 482

The Dielectric Properties of Insulating Materials—
E. J. Murphy and S. O. Morgan 502

OCTOBER, 1939

Experience in Applying Carrier Telephone Systems to Toll Cables
 —*W. B. Bedell, G. B. Ransom and W. A. Stevens* 547

The Toronto-Barrie Toll Cable—*M. J. Aykroyd and D. G. Geiger* . 588

The Computation of the Composite Noise Resulting from Random
 Variable Sources—*E. Dietze and W. D. Goodale, Jr.* 605

Load Rating Theory for Multi-Channel Amplifiers—
B. D. Holbrook and J. T. Dixon 624

The Quantum Physics of Solids, I—The Energies of Electrons in
 Crystals—*W. Shockley* 645

Dial Clutch of the Spring Type—*C. F. Wiebusch* 724

Index to Volume XVIII

A

- Affel, H. A., and B. W. Kendall*, A Twelve-Channel Carrier Telephone System for Open-Wire Lines, page 119.
Altimeter: A Terrain Clearance Indicator, *Lloyd Espenschied and R. C. Newhouse*, page 222.
Amplifiers, Multi-Channel, Load Rating Theory for, *B. D. Holbrook and J. T. Dixon*, page 624.
Anderson, C. N., A Representation of the Sunspot Cycle, page 292.
Aykroyd, M. J. and D. G. Geiger, The Toronto-Barrie Toll Cable, page 588.

B

- Bedell, W. B., G. B. Ransom and W. A. Stevens*, Experience in Applying Carrier Telephone Systems to Toll Cables, page 547.
Beyer, J. W., J. T. O'Leary and E. C. Blessing, An Improved Three-Channel Carrier Telephone System, page 49.
Blessing, E. C., J. T. O'Leary and J. W. Beyer, An Improved Three-Channel Carrier Telephone System, page 49.
Bowen, A. E., and F. B. Llewellyn, The Production of Ultra-High-Frequency Oscillations by Means of Diodes, page 280.

C

- Cable, Toll, The Toronto-Barrie, *M. J. Aykroyd and D. G. Geiger*, page 588.
Cables, Toll, Experience in Applying Carrier Telephone Systems to, *W. B. Bedell, G. B. Ransom and W. A. Stevens*, page 547.
Carrier System, Type "J," Some Applications of the, *L. C. Starbird and J. D. Mathis*, page 338.
Carrier System, Twelve-Channel Open-Wire, Line Problems in the Development of the, *L. M. Ilgenfritz, R. N. Hunter and A. L. Whitman*, page 363.
Carrier Telephone System, An Improved Three-Channel, *J. T. O'Leary, E. C. Blessing and J. W. Beyer*, page 49.
Carrier Telephone System, A Twelve-Channel, for Open-Wire Lines, *B. W. Kendall and H. A. Affel*, page 119.
Carrier Telephone Systems, Copper Oxide Modulators in, *R. S. Caruthers*, page 315.
Carrier Telephone Systems, Experience in Applying to Toll Cables, *W. B. Bedell, G. B. Ransom and W. A. Stevens*, page 547.
Carson, John R., Frequency-Modulation: Theory of the Feedback Receiving Circuit, page 395.
Caruthers, R. S., Copper Oxide Modulators in Carrier Telephone Systems, page 315.
Ceramic Manufacturing Developments of the Western Electric Company, Some, *A. G. Johnson and L. I. Shaw*, page 255.
Chaffee, J. G., The Application of Negative Feedback to Frequency-Modulation Systems, page 404.
Clarke, W. J., and J. R. Townsend, Plastic Materials in Telephone Use, page 482.
Clearance Indicator, A Terrain, *Lloyd Espenschied and R. C. Newhouse*, page 222.
Clutch, Dial, of the Spring Type, *C. F. Wiebusch*, page 724.
Contemporary Advances in Physics, XXXII. Particles of the Cosmic Rays, *Karl K. Darrow*, page 190.
Copper Oxide Modulators in Carrier Telephone Systems, *R. S. Caruthers*, page 315.
Cory, S. I., R. B. Shanck and F. A. Cowan, Recent Developments in the Measurement of Telegraph Transmission, page 143.
Cosmic Rays, Particles of the. Contemporary Advances in Physics, XXXII, *Karl K. Darrow*, page 190.
Cowan, F. A., R. B. Shanck and S. I. Cory, Recent Developments in the Measurement of Telegraph Transmission, page 143.

Crossbar Dial Telephone Switching System, *F. J. Scudder and J. N. Reynolds*, page 76.
 Crystals, The Energies of Electrons in—The Quantum Physics of Solids, I, *W. Shockley*, page 645.

D

Darrow, Karl K., Contemporary Advances in Physics, XXXII. Particles of the Cosmic Rays, page 190.
 Dial Clutch of the Spring Type, *C. F. Wiebusch*, page 724.
 Dial Telephone Switching System, Crossbar, *F. J. Scudder and J. N. Reynolds*, page 76.
 Dielectric Properties of Insulating Materials, The, *E. J. Murphy and S. O. Morgan*, page 502.
Dietze, E. and W. D. Goodale, Jr., The Computation of the Composite Noise Resulting from Random Variable Sources, page 605.
 Diodes, The Production of Ultra-High-Frequency Oscillations by Means of, *F. B. Llewellyn and A. E. Bowen*, page 280.
Dixon, J. T. and B. D. Holbrook, Load Rating Theory for Multi-Channel Amplifiers, page 624.

E

Electron-Optics, Electrostatic, *Frank Gray*, page 1.
 Electron Streams, Impedance Properties of, *Liss C. Peterson*, page 465.
 Electrons in Crystals, The Energies of—The Quantum Physics of Solids, I, *W. Shockley*, page 645.
Espenschied, Lloyd and R. C. Newhouse, A Terrain Clearance Indicator, page 222.

F

Feedback, Negative, The Application of to Frequency-Modulation Systems, *J. G. Chaffee*, page 404.
 Feedback Receiving Circuit, Theory of the. Frequency-Modulation: *John R. Carson*, page 395.
 Flood, and Hurricane—September 1938, *W. H. Harrison*, page 218.
 Frequency-Modulation: Theory of the Feedback Receiving Circuit, *John R. Carson*, page 395.
 Frequency-Modulation Systems, The Application of Negative Feedback to, *J. G. Chaffee*, page 404.

G

Geiger, D. G., and M. J. Aykroyd, The Toronto-Barrie Toll Cable, page 588.
Goodale, W. D., Jr., and E. Dietze, The Computation of the Composite Noise Resulting from Random Variable Sources, page 605.
Gray, Frank, Electrostatic Electron-Optics, page 1.

H

Harrison, W. H., Hurricane and Flood—September 1938, page 218.
Holbrook, B. D., and J. T. Dixon, Load Rating Theory for Multi-Channel Amplifiers, page 645.
Hunter, R. N., L. M. Ilgenfritz and A. L. Whitman, Line Problems in the Development of the Twelve-Channel Open-Wire Carrier System, page 363.
 Hurricane and Flood—September 1938, *W. H. Harrison*, page 218.
Hussey, L. W. and E. Peterson, Equivalent Modulator Circuits, page 32.

I

Ilgenfritz, L. M., R. N. Hunter and A. L. Whitman, Line Problems in the Development of the Twelve-Channel Open-Wire Carrier System, page 363.
 Insulating Materials, The Dielectric Properties of, *E. J. Murphy and S. O. Morgan*, page 502.

J

Johnson, A. G., and L. I. Shaw, Some Ceramic Manufacturing Developments of the Western Electric Company, page 255.

K

Kendall, B. W., and H. A. Affel, A Twelve-Channel Carrier Telephone System for Open-Wire Lines, page 119.

L

Legg, V. E., Survey of Magnetic Materials and Applications in the Telephone System, page 438.

Line Problems in the Development of the Twelve-Channel Open-Wire Carrier System, L. M. Ilgenfritz, R. N. Hunter and A. L. Whitman, page 363.

Lines, Open-Wire, A Twelve-Channel Carrier Telephone System for, B. W. Kendall and H. A. Affel, page 119.

Lines, Transcontinental Telephone, J. J. Pilliod, page 235.

Llewellyn, F. B., and A. E. Bowen, The Production of Ultra-High-Frequency Oscillations by Means of Diodes, page 280.

Load Rating Theory for Multi-Channel Amplifiers, B. D. Holbrook and J. T. Dixon, page 624.

M

Magnetic Materials and Applications in the Telephone System, Survey of, V. E. Legg, page 438.

Mathis, J. D., and L. C. Starbird, Some Applications of the Type "J" Carrier System, page 338.

Measurement of Telegraph Transmission, Recent Developments in the, R. B. Shanck, F. A. Cowan and S. I. Cory, page 143.

Modulation, Frequency: Theory of the Feedback Receiving Circuit, John R. Carson, page 395.

Modulation Systems, Frequency—The Application of Negative Feedback to, J. G. Chaffee, page 404.

Modulator Circuits, Equivalent, E. Peterson and L. W. Hussey, page 32.

Modulators, Copper Oxide, in Carrier Telephone Systems, R. S. Caruthers, page 315.

Morgan, S. O., and E. J. Murphy, The Dielectric Properties of Insulating Materials, page 502.

Moulding: Plastic Materials in Telephone Use, J. R. Townsend and W. J. Clark, page 482.

Multi-Channel Amplifiers, Load Rating Theory for, B. D. Holbrook and J. T. Dixon, page 624.

Murphy, E. J., and S. O. Morgan, The Dielectric Properties of Insulating Materials, page 502.

N

Network, n Terminal, The Number of Impedances of an, John Riordan, page 300.

Newhouse, R. C., and Lloyd Espenschied, A Terrain Clearance Indicator, page 222.

Noise, the Composite, Resulting from Random Variable Sources, The Computation of, E. Dietze and W. D. Goodale, Jr., page 605.

O

O'Leary, J. T., E. C. Blessing and J. W. Beyer, An Improved Three-Channel Carrier Telephone System, page 49.

Open-Wire Carrier System, Twelve-Channel, Line Problems in the Development of the, L. M. Ilgenfritz, R. N. Hunter and A. L. Whitman, page 363.

Open-Wire Lines, A Twelve-Channel Carrier Telephone System for, B. W. Kendall and H. A. Affel, page 119.

Optics, Electrostatic Electron—Frank Gray, page 1.

Oscillations, Ultra-High-Frequency, The Production of by Means of Diodes, F. B. Llewellyn and A. E. Bowen, page 280.

P

Peterson, E., and L. W. Hussey, Equivalent Modulator Circuits, page 32.

Peterson, Liss C., Impedance Properties of Electron Streams, page 465.

Physics, XXXII, Contemporary Advances in. Particles of the Cosmic Rays, Karl K. Darrow, page 190.

- Physics, The Quantum, of Solids, I—The Energies of Electrons in Crystals, *W. Shockley*, page 645.
Pilliod, J. J., Transcontinental Telephone Lines, page 235.
 Plastic Materials in Telephone Use, *J. R. Townsend and W. J. Clarke*, page 482.

Q

- Quantum Physics, The, of Solids, I—The Energies of Electrons in Crystals, *W. Shockley*, page 645.

R

- Ransom, G. B., W. B. Bedell and W. A. Stevens*, Experience in Applying Carrier Telephone Systems to Toll Cables, page 547.
Reynolds, J. N., and F. J. Scudder, Crossbar Dial Telephone Switching System, page 76.
Riordan, John, The Number of Impedances of an n Terminal Network, page 300.

S

- Scudder, F. J., and J. N. Reynolds*, Crossbar Dial Telephone Switching System, page 76.
Shanck, R. B., F. A. Cowan and S. I. Cory, Recent Developments in the Measurement of Telegraph Transmission, page 143.
Shaw, L. I., and A. G. Johnson, Some Ceramic Manufacturing Developments of the Western Electric Company, page 255.
Shockley, W., The Quantum Physics of Solids, I—The Energies of Electrons in Crystals, page 645.
 Spring Type, Dial Clutch of the, *C. F. Wiebusch*, page 724.
Starbird, L. C., and J. D. Mathis, Some Applications of the Type "J" Carrier System, page 338.
Stevens, W. A., W. B. Bedell and G. B. Ransom, Experience in Applying Carrier Telephone Systems to Toll Cables, page 547.
 Sunspot Cycle, A Representation of the, *C. N. Anderson*, page 292.
 Switching System, Crossbar Dial Telephone, *F. J. Scudder and J. N. Reynolds*, page 76.

T

- Telegraph Transmission, Recent Developments in the Measurement of, *R. B. Shanck, F. A. Cowan and S. I. Cory*, page 143.
 Telephone Lines, Transcontinental, *J. J. Pilliod*, page 235.
 Telephone System, An Improved Three-Channel Carrier, *J. T. O'Leary, E. C. Blessing and J. W. Beyer*, page 49.
 Telephone System for Open-Wire Lines, A Twelve-Channel Carrier, *B. W. Kendall and H. A. Affel*, page 119.
 Terrain Clearance Indicator, A, *Lloyd Espenschied and R. C. Newhouse*, page 222.
 Toll Cable, The Toronto-Barrie, *M. J. Aykroyd and D. G. Geiger*, page 588.
 Toll Cables, Experience in Applying Carrier Telephone Systems to, *W. B. Bedell, G. B. Ransom and W. A. Stevens*, page 547.
Townsend, J. R., and W. J. Clarke, Plastic Materials in Telephone Use, page 482.
 Transcontinental Telephone Lines, *J. J. Pilliod*, page 235.

U

- Ultra-High-Frequency Oscillations by Means of Diodes, The Production of, *F. B. Llewellyn and A. E. Bowen*, page 280.

W

- Whitman, A. L., L. M. Ilgenfritz and R. N. Hunter*, Line Problems in the Development of the Twelve-Channel Open-Wire Carrier System, page 363.
Wiebusch, C. F., Dial Clutch of the Spring Type, page 724.

THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS
OF ELECTRICAL COMMUNICATION

Electrostatic Electron-Optics— <i>Frank Gray</i>	1
Equivalent Modulator Circuits— <i>E. Peterson and L. W. Hussey</i>	32
An Improved Three-Channel Carrier Telephone System — <i>J. T. O'Leary, E. C. Blessing and J. W. Beyer</i>	49
Crossbar Dial Telephone Switching System — <i>F. J. Scudder and J. N. Reynolds</i>	76
A Twelve-Channel Carrier Telephone System for Open-Wire Lines— <i>B. W. Kendall and H. A. Affel</i>	119
Recent Developments in the Measurement of Telegraph Transmission— <i>R. B. Shanck, F. A. Cowan and S. I. Cory</i>	143
Contemporary Advances in Physics, XXXII. Particles of the Cosmic Rays— <i>Karl K. Darrow</i>	190
Hurricane and Flood—September 1938— <i>W. H. Harrison</i> . .	218
A Terrain Clearance Indicator — <i>Lloyd Espenschied and R. C. Newhouse</i>	222
Transcontinental Telephone Lines— <i>J. J. Pilliod</i>	235
Abstracts of Technical Papers	246
Contributors to this Issue	251

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

THE BELL SYSTEM TECHNICAL JOURNAL

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



EDITORIAL BOARD

F. B. Jewett	H. P. Charlesworth	W. H. Harrison
A. F. Dixon	O. E. Buckley	O. B. Blackwell
D. Levinger	M. J. Kelly	H. S. Osborne
	W. Wilson	
R. W. King, <i>Editor</i>	J. O. Perrine, <i>Associate Editor</i>	



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, 1939
American Telephone and Telegraph Company



The Bell System Technical Journal

Vol. XVIII

January, 1939

No. 1

Electrostatic Electron-Optics

By FRANK GRAY

Certain types of electrostatic fields may be used as lenses to focus electron beams. The theory of these lenses is developed for electric fields that are symmetrical about a central axis. The introduction of two velocity functions exactly reduces the partial differential equations of electron motion to a series of ordinary differential equations. The first equation describes the action of a lens for electron paths near the axis; the remaining equations determine the higher order aberration terms. Sections on the following subjects are included: the general equations of electron-optics, thin lenses, thick lenses, aberration, the reduction of aberration, apertured plates, and concentric tubes. A list of symbols and lens equations is also included at the end of the article.

IN certain types of modern vacuum tubes, a beam of electrons is brought to a focus by an electrostatic field whose action on the beam is analogous to that of an optical lens on a beam of light. An electrostatic field which acts in this manner is called an electron lens. Such lenses are rapidly finding applications in amplifier tubes, television and oscillograph tubes, electron microscopes, and various types of experimental apparatus. As the extent of their application widens, the theory of these lenses naturally assumes a corresponding importance.

The first articles on the new science of electron-optics were published by Bush¹ in 1926-1927, and the next important step in its development was taken by Davisson and Calbick² and by Brüche and Johannson³ working independently in 1931-1932. The following years marked an increased interest in the subject, with comprehensive articles by various authors, and its literature expanded rapidly. An

¹ H. Bush, *Ann. d. Physik*, 81, 974, 1926 and *Arch. f. Elektrotech.*, 18, 583, 1927.

² C. J. Davisson and C. J. Calbick, *Phys. Rev.*, 38, 585, 1931 and *Phys. Rev.*, 42, 580, 1932.

³ E. Brüche and N. Johannson, *Ann. d. Physik*, 15, 145, 1932.

excellent review of this literature and the history of electron-optics are given in a symposium⁴ of papers published in 1936, and the various practical applications of electron lenses are well described in the books on that subject.⁵

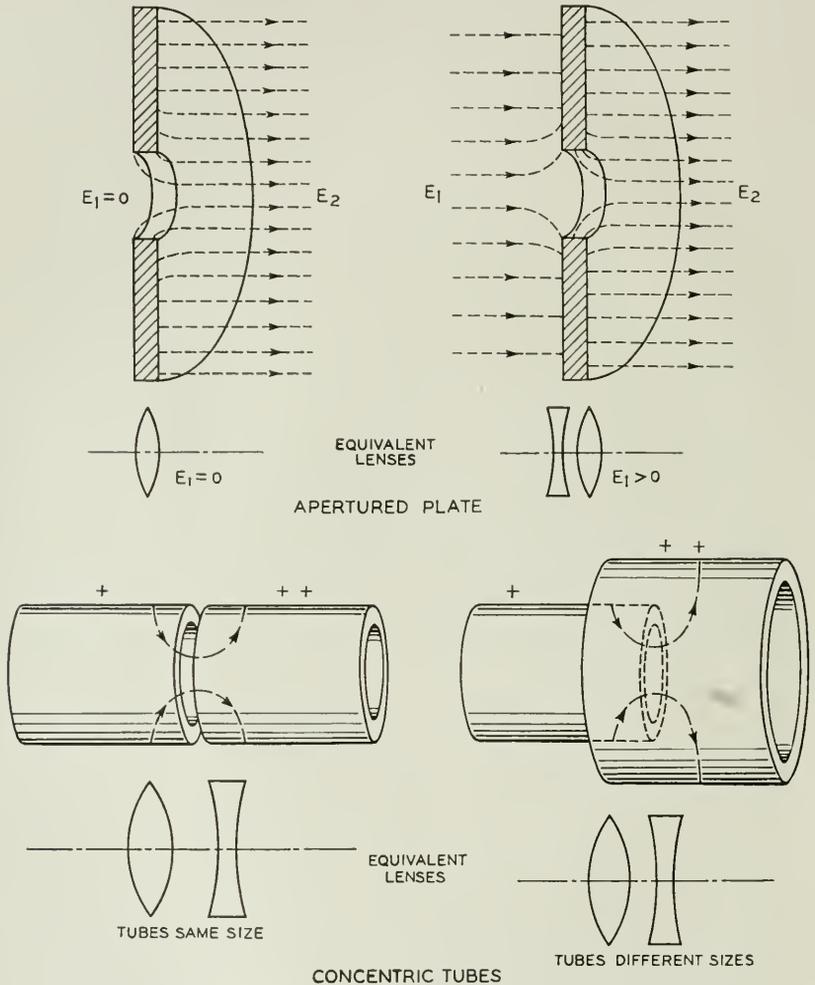


Fig. 1—Lines of force in typical electron lenses.

⁴ *Zeit. f. techn. Physik*, 17, 584-645, 1936.

⁵ E. Brüche and O. Scherzer, *Geometrische Electronenoptik* (Springer, 1934). J. T. MacGregor-Morris and J. A. Henley, "Cathode Ray Oscillography" (Monographs on Electrical Engineering), 1936. Maloff and Epstein, "Electron Optics in Television" (McGraw-Hill, 1938).

The theory of electron-optics is thus well established and any further attempts at the subject must lead to substantially the same results. There is, however, a need for a precise development of the theory in a simpler manner. With this need in mind, the present article approaches the subject in a manner that appeals to the reader who is more familiar with electrical theory than he is with the concepts of geometrical optics, and this approach leads clearly to the various approximations that are needed in the development of the theory. With the aid of two velocity functions, the partial differential equations of electron motion are briefly and exactly reduced to a series of ordinary differential equations; the theory is then developed in terms of their approximate solutions.

Attention is confined to systems in which the electric fields are symmetrical about a central axis. In such systems any field having a radial component of electric intensity changes the radial velocity of an electron passing through it, and thus behaves—to some extent at least—as an electron lens. A uniform field parallel to the axis and field-free space are the only regions in which there is no lens action. Typical electron lenses are shown in the figures on the second page. As illustrated by these examples, a practical electron lens is characterized by a short region in which there is an abrupt change in the electric intensity parallel to the axis. Lines of force are continuous, and the field parallel to the axis can change only by lines of force coming into it, or going out from it, in a radial direction. In the region of the abrupt change, there are consequently strong radial fields which can deflect an electron in a radial direction. The region changes the focus of an electron beam passing through it, and its action is analogous to that of an optical lens.

SECTION I—THE GENERAL EQUATIONS

In the present paper it is assumed that the initial electron source has perfect symmetry of form about the central axis, and that the electrons have no appreciable velocities of emission from the source. An electron thus has no angular velocity about the axis, and its motion may be described in terms of a coordinate z taken along the axis and a radial coordinate r measured from the axis.

If an electron's velocity vector is projected at any point along its path, it intersects the axis at some point p , as illustrated in Fig. 2, and the electron may be regarded as instantaneously moving either away from, or else toward, this point of intersection. The distance d along the axis from the electron to the point of intersection is called

the instantaneous focal distance.⁶ Defined in this manner, the focal distance conforms with the optics of light; it is positive when an electron is moving toward a focal point; and is negative when the

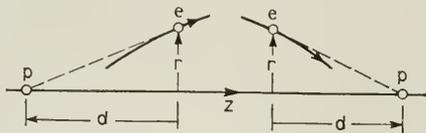


Fig. 2—Focal distance.

electron is moving away from such a point. From the geometry of the figure, it is seen that

$$d = -\frac{r\dot{z}}{\dot{r}}, \quad (1)$$

where \dot{r} and \dot{z} are the instantaneous components of electron velocity.

The focal distance of an electron varies continuously as the electron moves along. The simplest variation occurs in field-free space, where the electron travels in a straight line and the focal point remains stationary; but even then the focal distance varies as the electron moves; for the focal distance is measured from the moving electron to the stationary focal point, as illustrated in Fig. 3. In an electron

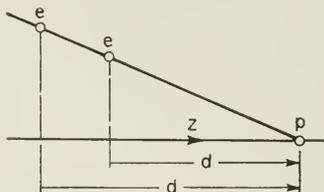


Fig. 3—Focal distances, field-free space.

lens, the focal point of an electron also shifts continuously as the electron moves through the lens and the focal distance varies in a complicated manner.

The values of d at the two sides⁷ of an electron lens, for any electron path, are called conjugate focal distances of the lens, and are usually designated as d_1 and d_2 . The theory of electron-optics is largely concerned with the derivation of an equation relating these conjugate focal distances.

⁶ The term is here used in a broad sense to include the distance to any intersection point on the z -axis, even though the latter is not the point of convergence of an electron beam.

⁷ The value of d as an electron enters the non-uniform field of the lens, and the value of d as it leaves the non-uniform field.

Before passing on to such a derivation, it is well to introduce another quantity, which is analogous to focal distance and very useful in making approximations. Suppose that, from any point along its path, an electron were to continue on with its instantaneous velocity in a straight line. Its velocity along the axis would continue to have the instantaneous value \dot{z} , and the electron would travel over the distance d and arrive at the focal point in a period of time T given by

$$T = d/\dot{z} \quad (2)$$

or from equation 1

$$T = -r/\dot{r}. \quad (3)$$

This period of time is analogous to focal distance, and we therefore call it focal time. The values of T at the two sides of an electron lens, for any electron path, are in a corresponding manner called conjugate focal times of the lens.

To obtain an equation relating the conjugate focal distances of a lens, we must consider the path of an electron through the lens. The path is determined by the initial velocity and coordinates of the electron as it enters the lens and by its acceleration in the electric field of the lens. By defining electrical units in the proper manner the ratio e/m is eliminated from the equations of acceleration and they assume the simple form

$$\ddot{r} = \frac{\partial \Phi}{\partial r}, \quad (4)$$

$$\ddot{z} = \frac{\partial \Phi}{\partial z}, \quad (5)$$

where Φ is the potential at points in space.⁸

The first solution of these equations gives the well known energy relation

$$\dot{r}^2 + \dot{z}^2 = 2\Phi, \quad (6)$$

where the electron source is taken as zero potential.

With the exception of special cases, the equations are not further soluble in the usual sense, and one resorts to solution in series.

As they stand, the two equations for acceleration are inconvenient; they involve partial derivatives of potential with respect to space and ordinary derivatives of velocity with respect to time, and the latter cannot be transformed to partial derivatives with respect to space, for the simple reason that the velocity of an electron does not exist

⁸ The final equations of electron-optics involve the potentials only in the form of ratios which are independent of the electrical units.

at points off its path. The equations may, however, be reduced to a more convenient form by the introduction of two velocity functions⁹ defined as follows.

Let u and w be any two functions of r and z that satisfy the equations

$$\frac{\partial u}{\partial z} - \frac{\partial w}{\partial r} = 0, \quad (7)$$

$$u^2 + w^2 = 2\Phi. \quad (8)$$

Consider now an imaginary point moving with velocity components

$$\dot{r} = u, \quad \dot{z} = w. \quad (9)$$

The derivative of \dot{r} with respect to time is

$$\ddot{r} = \frac{\partial u}{\partial r} \dot{r} + \frac{\partial u}{\partial z} \dot{z} = \frac{\partial u}{\partial r} u + \frac{\partial u}{\partial z} w \quad (10)$$

and from equations 7 and 8

$$\ddot{r} = \frac{\partial u}{\partial r} u + \frac{\partial w}{\partial r} w = \frac{\partial \Phi}{\partial r} \quad (11)$$

the component \dot{r} thus satisfies differential equation (4) for electron motion. In a similar manner it may be shown that the velocity component \dot{z} satisfies equation (5). The motion of the imaginary point is thus the same as the motion of an electron, and the velocity functions u and w are therefore the velocity components of electron motion.

The velocity functions are solutions of equations 7 and 8, one of which is a simple algebraic equation and the other a partial differential equation with respect to space alone. The inconvenient time derivatives have been eliminated in these new equations for electron velocity.

The existence of a velocity function is not confined to a single electron path; it exists over the electric field in general. Any pair of particular solutions for u and w thus corresponds to an infinite number of possible electron paths. In the converse manner, there are an infinite number of particular solutions for any electric field, and there is a pair of particular solutions corresponding to any given electron path through the field.¹⁰

⁹ These functions are the components of the generalized vector function described in Appendix 4.

¹⁰ The existence of such solutions is proved by the existence of the series solutions, which are derived in the following pages.

Solutions for the velocity functions are obtained by expressing them as power series in r .

$$u = Ar + Br^3 + Cr^5 + \dots, \quad (12)$$

$$w = a + br^2 + cr^4 + \dots, \quad (13)$$

where the coefficients are functions of z alone. The above powers of r are the ones required in a system symmetrical about the z -axis. In such a system \dot{r} reverses in sign with r and the u -series is odd; \dot{z} does not reverse sign with r and the w -series is even. Aside from such reasoning, the choice of the two series is justified provided they lead to solutions of the differential equation in a form suitable for the purposes of electron-optics.

The potential Φ obeys the equation

$$\Delta\Phi = \frac{\partial^2\Phi}{\partial z^2} + \frac{\partial^2\Phi}{\partial r^2} + \frac{1}{r}\frac{\partial\Phi}{\partial r} = 0 \quad (14)$$

and it may likewise be expressed as a power series in r . This well known series is

$$\Phi = v - \frac{v''}{2^2}r^2 + \frac{v''''}{(2 \cdot 4)^2}r^4 \dots, \quad (15)$$

where v is the potential on the axis of the system, and the primes indicate differentiation with respect to z .

On substituting the three series in equations 7 and 8 and equating the coefficients of the various powers of r in each equation we obtain a series of ordinary differential equations for the coefficients of the u -series.

$$\sqrt{2v}A' + A^2 = -\frac{v''}{2}, \quad (16)$$

$$\sqrt{2v}B' + 4AB = \frac{v''''}{16} - \frac{(A')^2}{2}, \quad (17)$$

$$\sqrt{2v}C' + 6AC = -\frac{v'''''}{384} - 3B^2 - 3/4A'B', \quad (18)$$

.

and the coefficients of the w -series are

$$\begin{aligned} a &= \sqrt{2v}, \\ b &= A'/2, \\ c &= B'/4. \end{aligned} \quad (19)$$

.

The solution of the partial differential equations for electron velocity is thus reduced to the solution of a series of ordinary differential equations, which in themselves contain no approximations.

From equation 1, the inverse focal distance is now obtained by dividing u by r and w , which gives

$$\frac{1}{d} = - \frac{A + Br^2 + Cr^4 + \dots}{\sqrt{2v} + \frac{A'}{2}r^2 + \frac{B'}{4}r^4 + \dots} \quad (20)$$

This is the general equation for focal distance as it is affected by aberration. In using this equation, we are at liberty to set the higher coefficients equal to zero at the incident side of the lens. This determines the initial value of A in terms of the first conjugate focal distance. The second conjugate focal distance is then determined by solving for the coefficients at the exit side of the lens. Due to the presence of the terms in r , this second focal distance varies slightly with the radial distance at which an electron passes through the lens, and the focus is therefore diffused along the axis. This diffusion of the focus is called aberration.

The coefficient A is of particular importance in the theory of electron-optics. For paraxial rays, that is, rays near the axis, the higher terms in the two series are negligibly small compared to their first terms, and for such rays

$$\frac{1}{d} = - \frac{A}{\sqrt{2v}} \quad (21)$$

With the exception of aberration, the single coefficient A thus determines the complete performance of a lens, and the principal constants of a lens are determined by its differential equation alone. In lenses where the rays are confined to a region near the axis with proper diaphragms, the aberration terms are small and the coefficient A describes the performance of a lens sufficiently well.

The next section is devoted to the derivation of the principal lens equations from this coefficient. The aberration terms are considered only in the last section of the paper.

SECTION II—RAYS NEAR THE AXIS

For rays near the axis the optical characteristics of an electric field are determined by the differential equation for A alone,

$$\sqrt{2v}A' + A^2 = - \frac{v''}{2} \quad (16)$$

For such rays the higher terms in r may be neglected in the general equations and we obtain the following useful relations

$$A = \dot{r}/r = -\frac{\sqrt{2v}}{d}, \quad (22)$$

$$\dot{z} = \sqrt{2v}, \quad (23)$$

$$T = d/\sqrt{2v} = -\frac{1}{A}, \quad (24)$$

$$dt = dz/\dot{z} = dz/\sqrt{2v}. \quad (25)$$

A Uniform Electric Field

A uniform electric field parallel to the axis is not usually regarded as an electron lens,¹¹ but it does shift the focal point of a beam of electrons passing through it. In a uniform field, v'' is zero and the differential equation for A may be written in the form

$$\frac{dA}{A^2} = -\frac{dz}{\sqrt{2v}}. \quad (26)$$

An integration of this equation from any point z_1 to any other point z_2 , in the uniform field, gives

$$\frac{1}{A_2} - \frac{1}{A_1} = \frac{2(z_2 - z_1)}{\sqrt{2v_2} + \sqrt{2v_1}}, \quad (27)$$

where A_1 and A_2 are the values of A at z_1 and z_2 . On substituting $-\sqrt{2v}/d$ for A in this equation, it may be transformed to

$$\left(1 + \sqrt{\frac{v_2}{v_1}}\right) d_1 - \left(1 + \sqrt{\frac{v_1}{v_2}}\right) d_2 = 2(z_2 - z_1), \quad (28)$$

which is the equation relating the conjugate focal distances at any two planes—located at z_1 and z_2 —in the uniform field.

The shift in the focal point of an electron beam as it passes through a uniform field is illustrated in Fig. 4.

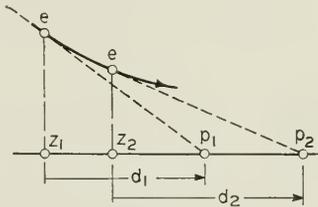


Fig. 4—Focal distances in a uniform field.

¹¹ Electron rays parallel to the axis are not bent by the field, and it does not magnify an electron image.

Thin Lenses

Approximate solutions of the differential equation 16 for A are obtained more clearly by first changing the space variables to time variables. This is done by using relations 24 and 25, which transform the equation to

$$\frac{1}{T^2} \left(\frac{dT}{dt} + 1 \right) = -\frac{v''}{2} \quad (29)$$

or

$$\frac{1}{T^2} \frac{d}{dt} (T + t) = -\frac{v''}{2}. \quad (30)$$

The new equation tells how the focal time T varies with time as an electron moves along.¹²

A thin lens is defined as a region of non-uniform field extending over such a short distance along the axis that an electron traverses it in a period of time small compared to the focal times involved: the thickness of the lens is small compared to the conjugate focal distances. By taking the origin of time t at the middle of an electron's period of transit through a lens, t in the lens is not greater than half the period of transit, and t may therefore be neglected in comparison to T in a thin lens. With this approximation in equation 31, it reduces to

$$\frac{d}{dt} \left(\frac{1}{T} \right) = \frac{v''}{2}. \quad (31)$$

In integrating this equation through a lens we choose two points z_1 and z_2 at the approximate boundaries of the non-uniform field, that is, the points where v'' substantially drops to zero as illustrated in Fig. 5. Then, remembering that dt is $dz/\sqrt{2v}$, an integration from z_1

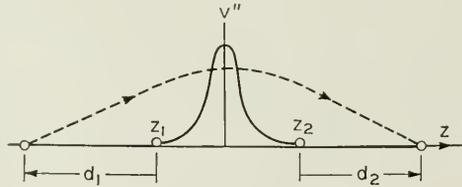


Fig. 5—Conjugate focal distances, thin lens.

to z_2 gives

$$\frac{1}{T_2} - \frac{1}{T_1} = \frac{1}{F}, \quad (32)$$

¹² The period of time that a train requires to reach its terminal point also varies with time as the train moves along.

where the inverse focal term is

$$\frac{1}{F} = \int_{z_1}^{z_2} \frac{v''}{2\sqrt{2v}} dz \quad (33)$$

or on integration by parts

$$\frac{1}{F} = \frac{1}{2} \left[\left(\frac{v'}{\sqrt{2v}} \right)_2 - \left(\frac{v'}{\sqrt{2v}} \right)_1 \right] + \frac{1}{2} \int_z^{z_2} (v')^2 (2v)^{-3/2} dz. \quad (34)$$

The substitution in equation 32 of the values for T_1 and T_2 as given by equation 24 now gives the lens equation

$$\frac{\sqrt{2v_2}}{d_2} - \frac{\sqrt{2v_1}}{d_1} = \frac{1}{F}. \quad (35)$$

This equation is analogous to the equation for a thin optical lens

$$\frac{\mu_2}{d_2} - \frac{\mu_1}{d_1} = \frac{1}{F}, \quad (36)$$

bounded on its two sides by media with different refractive indices μ_1 and μ_2 , the $\sqrt{2v}$ corresponding to refractive index.

Electron rays parallel to the axis do not come to a focus at a distance F from an electron lens; in other words, F is not a principal focal distance. There are, in general, two principal focal points on opposite sides of an electron lens. Their principal focal distances f_1 and f_2 are found by setting first d_1 , and then d_2 , equal to infinity in equation 35. This gives

$$f_1 = -\sqrt{2v_1}F, \quad f_2 = \sqrt{2v_2}F \quad (37)$$

as the two principal focal distances. It may be shown from equation 33 that these principal focal distances really involve the voltages only in the form of the ratio v_2/v_1 . By substituting them in the lens equation 35, it may be written in the convenient form

$$\frac{f_2}{d_2} + \frac{f_1}{d_1} = 1, \quad (38)$$

which likewise involves the voltages only in the form of a ratio.

There are two types of electron lenses that deserve special consideration. The first is a small aperture in a thin plate separating two uniform fields of different intensities—as a special case one of the fields may be zero. An example of such a lens is illustrated in Fig. 6.

In this type of lens, the non-uniform field at the aperture covers a distance along the axis about equal to its diameter.¹³ If the diameter is small compared to v/v' , there is little change in potential throughout

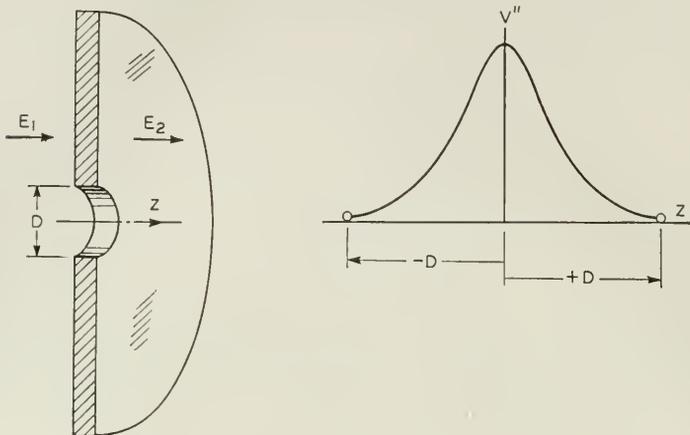


Fig. 6—An apertured plate.

the lens and the $\sqrt{2v}$ may be considered as a constant in the integration 33 for the inverse focal term. With this approximation,

$$\frac{1}{F} = \frac{1}{2} \left[\frac{v_2' - v_1'}{\sqrt{2v}} \right], \quad (39)$$

when v is the potential of the plate and v_1' and v_2' are the electric intensities of the two uniform fields. The lens equation 35 is then

$$\frac{\sqrt{2v}}{d_2} - \frac{\sqrt{2v}}{d_1} = \frac{1}{2} \left[\frac{v_2' - v_1'}{\sqrt{2v}} \right], \quad (40)$$

which may be written in the simpler form

$$\frac{1}{d_2} - \frac{1}{d_1} = \frac{v_2' - v_1'}{4v}. \quad (41)$$

In this type of lens, the electrical refractive index $\sqrt{2v}$ is the same on both sides of the lens and the two principal focal distances are equal, just as they are for a thin optical lens when it is bounded by air on both sides.¹⁴

¹³ "Two Problems in Potential Theory," T. C. Fry, *Bell Telephone System Monograph B-671*.

¹⁴ A complete electron-optical system usually involves a combination of lenses. The calculations for a combination are illustrated by the example in Appendix 1.

The apertured plate between two uniform fields is the only lens that permits such a simple calculation of focal distances. In all other lens structures the potential varies appreciably throughout the lens and the integration for the focal term is complicated. The actual numerical calculations have been carried out for only a few of these cases.

The second type of lens deserving special consideration is a lens bounded on both sides by field-free space. For such boundaries the first term in the last member of equation 34 vanishes, and $1/F$ is determined by the integral term alone. This integral is inherently positive, and a lens bounded on both sides by field-free space is thus always a convergent lens. The two concentric tubes of Fig. 7 give

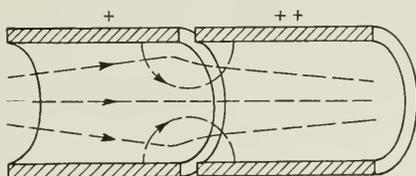


Fig. 7—Concentric tubes—lines of force and electron paths.

a lens of this type, the electric field in each tube dropping to zero at a short distance from its end. It is true that there is a divergent field of the same intensity as the convergent field; but an electron is at a higher potential in the divergent field and traveling faster, so it receives a smaller radial deflection in that field and the lens is convergent. It is interesting to note that the lens is still positive even when the potentials on the electrodes are reversed; in other words, a lens of this type is positive irrespective of the direction of the electric field.¹⁵

An Approximation for Certain Thick Lenses

In certain electron lenses there is a short region of strong lens action accompanied by more extended regions of weaker action; the large values of the derivative v'' are confined to a short distance along the axis, but the derivative does have appreciable values over a more extended region. A lens of this type can be treated in the following approximate manner, provided that there is but one maximum of $|v''|$ in the lens.

For this purpose, the differential equation 31 is rewritten in the form

$$d\left(\frac{1}{T+t}\right) = \frac{v''}{2} \left(1 + \frac{t}{T}\right)^{-2} dt, \quad dt = dz/\sqrt{2v} \quad (42)$$

¹⁵ The principal focal distances of concentric tubes are calculated in Appendix 2.

and the lens equation is derived by integrating it from a point z_1 to a point z_2 , where the two points are taken at the substantial boundaries of the non-uniform field. In carrying out this integration, the origin of time is taken at the instant that the electron is at the maximum of $|v''|$, as illustrated in Fig. 8, and for convenience the origin of z is

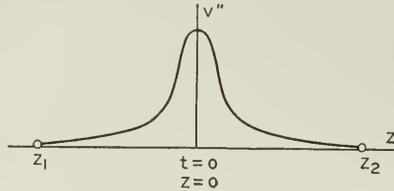


Fig. 8—Coordinates for a thick lens.

also taken at that point. With this choice of the origin, the term t/T in the second member of the equation is small compared to unity in the region where v'' is large and not very important in the regions where v'' is small. This term may therefore be neglected in lenses when the time of transit is not too great a fraction of the focal times involved. The integration of the equation then gives

$$\frac{1}{T_2 + t_2} - \frac{1}{T_1 + t_1} = \frac{1}{F} \quad (43)$$

when the inverse focal term is again

$$\frac{1}{F} = \int_{z_1}^{z_2} \frac{v''}{2\sqrt{2v}} dz \quad (44)$$

and

$$t_2 = \int_0^{z_2} \frac{dz}{\sqrt{2v}}, \quad t_1 = \int_0^{z_1} \frac{dz}{\sqrt{2v}}. \quad (45)$$

A transformation to space variables by means of equations 24 and 25 gives the lens equation in a form analogous to that for a thick optical lens,

$$\frac{\sqrt{2v_2}}{d_2 - \alpha_2} - \frac{\sqrt{2v_1}}{d_1 - \alpha_1} = \frac{1}{F}, \quad (46)$$

where

$$\alpha_2 = -\sqrt{2v_2}t_2 = -\int_0^{z_2} \sqrt{\frac{v_2}{v}} dz, \quad (47)$$

$$\alpha_1 = -\sqrt{2v_1}t_1 = -\int_0^{z_1} \sqrt{\frac{v_1}{v}} dz.$$

A plane located at a distance α_1 from the point z_1 is the approximate first principal plane of the lens; and a plane located at a distance α_2 from the point z_2 is the approximate second principal plane of the lens. In the lens equation, $d_1 - \alpha_1$ and $d_2 - \alpha_2$ are the conjugate focal distances measured from the principal planes. If the focal distances measured in this manner are designated as D_1 and D_2 respectively, the lens equation assumes the simpler form

$$\frac{\sqrt{2v_2}}{D_2} - \frac{\sqrt{2v_1}}{D_1} = \frac{1}{F}. \quad (48)$$

An electron lens frequently has both a positive and a negative maximum of v'' , and the preceding approximation cannot be applied to the lens as a whole. There is, however, necessarily a point between the two maxima where v'' is zero and by taking this as a division point, the lens can be separated into two components. The approximation can then be separately applied to each component, and the whole lens treated as a combination of two lenses.

The General Theory of Thick Lenses.

The equation for the coefficient A ,

$$\frac{dA}{dz} + \frac{A^2}{\sqrt{2v}} = -\frac{v''}{2\sqrt{2v}},$$

is a Racciti equation and, with the exception of special cases, it has no exact solution in the usual sense. Particular solutions can be obtained only by integration in series. It is, however, possible to express the general solution of a Racciti equation in terms of any two particular solutions, and this property enables us to develop the general theory of a thick lens in terms of its principal focal distances.

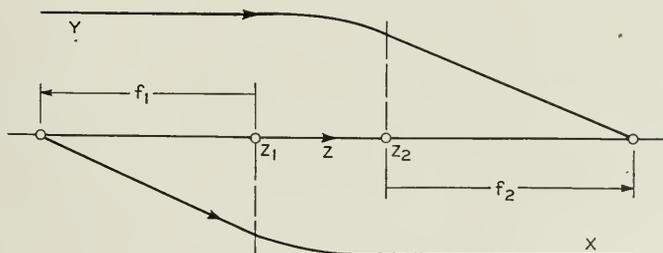


Fig. 9—Paths corresponding to X and Y .

In considering a thick lens, two points z_1 and z_2 are again taken at the substantial boundaries of the non-uniform field constituting the

lens. The differential equation for A necessarily has a particular solution equal to zero at z_1 . This solution is designated as Y , and it corresponds to an electron ray entering the lens parallel to the axis. At z_2 this solution is equal to $-\sqrt{2v_2}/f_2$, where f_2 is the second principal focal distance measured from z_2 . The path of such a ray is illustrated in Fig. 9. This particular solution obeys the same differential equation as A . By subtracting the differential equation of A from that of Y and making a slight transformation, we obtain

$$\frac{d}{dz} \log(A - Y) = -\frac{A + Y}{\sqrt{2v}}, \quad (49)$$

and it should be noted that

$$A/\sqrt{2v} = \frac{\dot{r}}{r\sqrt{2v}} = \frac{d}{dz} \log r. \quad (50)$$

An integration from z_1 to z_2 and a transformation to focal distances then gives the relation

$$\frac{(f_2 - d_2)d_1}{f_2d_2} = k_2 \sqrt{\frac{v_1 r_1}{v_2 r_2}}, \quad (51)$$

where k_2 is a constant of the lens, given by

$$1/k_2 = \exp. \int_{z_1}^{z_2} \frac{Y dz}{\sqrt{2v}}. \quad (52)$$

By proceeding in the same manner with a particular solution X for a ray leaving the lens parallel to the axis, we obtain a second relation

$$\frac{(f_1 - d_1)d_2}{f_1d_1} = k_1 \sqrt{\frac{v_2 r_2}{v_1 r_1}}, \quad (53)$$

where

$$k_1 = \exp. \int_{z_1}^{z_2} \frac{X dz}{\sqrt{2v}}. \quad (54)$$

The differential equations of X and Y may also be subtracted and integrated, and this gives a third relation

$$f_1/f_2 = -\frac{k_2}{k_1} \sqrt{\frac{v_1}{v_2}}. \quad (55)$$

A multiplication of the first two relations 51 and 53 gives

$$(f_2 - d_2)(f_1 - d_1) = k_1 k_2 f_1 f_2, \quad (56)$$

which is one form of the equation relating the conjugate focal distances of a lens. This equation may be converted into a more useful form by the following considerations.

A combination of the three preceding relations gives

$$\frac{r_2}{d_2}(d_2 - \alpha_2) = \frac{r_1}{d_1}(d_1 - \alpha_1), \quad (57)$$

where

$$\alpha_1 = f_1(1 - k_1), \quad \alpha_2 = f_2(1 - k_2). \quad (58)$$

To interpret this equation, we erect two imaginary planes as shown in Fig. 10. The first plane is located at a distance α_1 from z_1 . If the

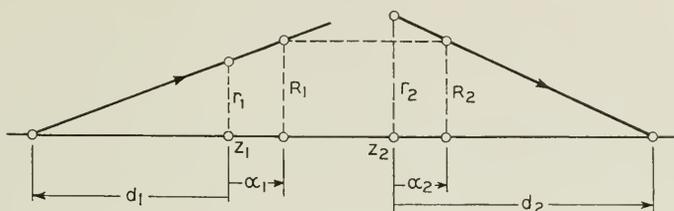


Fig. 10—The principal planes.

path of the incident ray is projected it intersects this plane at some radial distance R_1 . The second plane is erected at a distance α_2 from z_2 . The path of the exit ray intersects it at a radial distance R_2 . The equation says—from simple geometry—that the two radial distances R_1 and R_2 are equal. The path of an electron through the lens is therefore the same as if the electron proceeded in a straight line to the first plane, passed parallel to the axis to the second plane, and then proceeded again in a straight line to the second conjugate focal point. These two planes are called the first and second principal planes of the lens. The action of a thick lens is the same as if the space between the principal planes were non-existent, leaving them in coincidence, and a thin lens were located at the plane of coincidence.

The principal planes of a lens may lie either inside or outside of the lens. In most convergent lenses, α_1 is positive and α_2 negative, and the two planes both lie inside the lens.

The first conjugate focal distance measured from the first principal plane is designated as D_1 , and the second conjugate focal distance measured from the second principal plane is designated as D_2 . When they are measured in this manner, the two conjugate distances are

$$D_1 = d_1 - \alpha_1, \quad D_2 = d_2 - \alpha_2. \quad (59)$$

The two principal focal distances measured from the principal planes are, in a similar manner, designated as F_1 and F_2 ; then, from equation 58,

$$\begin{aligned} F_1 &= f_1 - \alpha_1 = k_1 f_1, \\ F_2 &= f_2 - \alpha_2 = k_2 f_2, \end{aligned} \quad (60)$$

and, from equation 55,

$$F_1/F_2 = -\sqrt{\frac{2v_1}{2v_2}}. \quad (61)$$

Substitution of the new focal distances in the lens equation 56 now gives

$$(F_2 - D_2)(F_1 - D_1) = F_1 F_2 \quad (62)$$

or

$$\frac{F_2}{D_2} + \frac{F_1}{D_1} = 1. \quad (63)$$

This is the general equation relating the conjugate focal distances in any lens. With the aid of equation 61, it may be written in the more familiar form

$$\frac{\sqrt{2v_2}}{D_2} - \frac{\sqrt{2v_1}}{D_1} = \frac{1}{F}, \quad (64)$$

where

$$\frac{1}{F} = \frac{\sqrt{2v_2}}{F_2} = -\frac{\sqrt{2v_1}}{F_1}. \quad (65)$$

The Principal Points of a Lens

The points locating the two principal planes on the axis of a lens and its two principal focal points are called the cardinal points of the lens. The preceding theory of a thick lens shows that its performance is completely determined by the locations of these four points.¹⁶ The theory does not furnish a general method for calculating their locations, but it does show that they can be determined from a knowledge of two so-called principal rays. The first is a ray leaving the lens parallel to the axis. If its entrance and exit paths are projected, they intersect as shown in Fig. 11, and the intersection locates the first principal plane. The projected incident ray also intersects the axis, and this intersection locates the first principal focal point. The second principal plane and the second principal focal point may be located in a similar manner from the entrance and exit paths of a ray entering the lens parallel to the axis.

¹⁶ We are here speaking only of rays near the axis.

The required paths of the two rays must in general be determined either by a series or step-by-step integration of the differential equation for A , or else by actual measurements on the physical lens.¹⁷

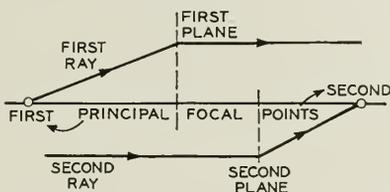


Fig. 11—The cardinal points of a lens.

Magnification

Electron object and electron image are defined as their optical analogies. The electron object may be an actual source of electrons, or the real image of such a source, or it may be a virtual image from which the electrons are apparently coming as they enter a lens.

The magnification by an electron lens may be treated in the following manner. Let S_1 be the size of an electron object located at a distance D_1 from the first principal plane of a lens. Two electron rays from the edge of the object are considered—as shown in Fig. 12. The ray

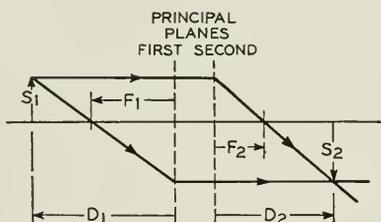


Fig. 12—Magnification.

entering the lens parallel to the axis may be regarded as passing on to the second principal plane and then bending sharply to pass through the second principal focal point; the ray through the first principal focal point may be considered as passing on to the first principal plane and then proceeding parallel to the axis. The intersection of the two rays locates the electron image and determines its size S_2 . The magnification M is defined as S_2/S_1 , and it follows from simple geometry that

$$M = \frac{D_2 - F_2}{F_2} = \frac{F_1}{D_1 - F_1}. \quad (65)$$

¹⁷ Other step-by-step methods can be used when a map of the equipotential surface is available.

A more convenient expression for magnification is now obtained by combining the two preceding expressions to give

$$M = \frac{F_1 D_2}{F_2 D_1} \quad (66)$$

and from equation 61

$$M = -\sqrt{\frac{2v_1 D_2}{2v_2 D_1}}. \quad (67)$$

The magnification is not in general equal to the ratio of the image distance to the object distance, as it is for an optical lens in air. It is only equal to that ratio when the voltage is the same on both sides of the lens.

SECTION III—ABERRATION IN A LENS

Returning to the first section, we see that the general expression for focal distance is

$$\frac{1}{d} = -\frac{A + Br^2 + Cr^4 \dots}{\sqrt{2v} + \frac{A'}{2}r^2 + \frac{B'}{4}r^4 \dots}. \quad (20)$$

The exact focal distance of an electron thus depends on its radial coordinate r , and a ray passing through a lens at a distance from the axis does not come to the same focus as a ray near the axis. A precise, general theory for rays at a distance from the axis could—in theory at least—be derived by solving the differential equations for as many of the higher coefficients as desired and substituting them in the above equation. Such a general solution would, however, be very difficult indeed, and one is content—as he usually is in optics—to treat the performance of a lens in a much more restricted manner.

The equation for focal distance can be simplified to some extent by noting that its denominator is the velocity component \dot{z} . With the aid of the energy equation 6, this component can be written in the form

$$\dot{z} = \sqrt{2v} \left[1 + \left(\frac{r}{d} \right)^2 \right]^{-1/2}. \quad (68)$$

In most lenses r is small compared to d , and the last factor in the above equation may be approximately set equal to unity. This approximation is accurate to one per cent even for a lens with an angular aperture corresponding to F3.5—an F2 lens is a very fast camera lens. With this approximation the inverse focal distance is

$$\frac{1}{d} = -\frac{A + Br^2 + Cr^4 + \dots}{\sqrt{2v}}. \quad (69)$$

The presence of the terms in r causes a diffusion of the focus in a lens, and a clearer picture of this diffusion is obtained by expressing it as lateral aberration. So we now proceed to derive an expression for this aberration, and the meaning of the term becomes apparent from the derivation. For this purpose we consider electrons entering a lens as if they all came from a point source at a distance d_1 from the first side of the lens. We are at liberty to set the higher coefficients equal to zero at that side of the lens, and this gives

$$A_1 = -\frac{\sqrt{2v_1}}{d_1}, \quad (70)$$

$$B_1 = C_1 = \dots = 0.$$

At the exit side of the lens, the focal distance is

$$\frac{\sqrt{2v_2}}{d_2} = -(A_2 + B_2r^2 + C_2r^4 + \dots), \quad (71)$$

where the coefficients are solutions of their differential equations subject to the initial conditions 70. The focal distance d_0 for rays near the axis is given by

$$\frac{\sqrt{2v_2}}{d_0} = -A_2. \quad (72)$$

The difference between the focal distance d_2 of a ray leaving the lens at a distance r from the axis and the focal distance d_0 of a ray near the axis is

$$d_2 - d_0 = \frac{d_0^2}{\sqrt{2v_2}} (B_2r^2 + C_2r^4 + \dots). \quad (73)$$

This difference is called the longitudinal aberration of the lens. It is the distance that the focal point is diffused along the axis, when the lens is limited by an exit diaphragm or radius r .

If a screen is placed at a distance d_0 from the lens, rays near the axis will come to a point focus on the screen; but rays leaving the lens at a distance r from the axis will strike the screen along a circular line. The radius s of this circle is called the lateral aberration of the lens. It follows rather simply, from the value of the longitudinal aberration, that the lateral aberration is

$$s = \frac{d_0}{\sqrt{2v_2}} (B_2r^3 + C_2r^5 + \dots). \quad (74)$$

This is the radius of the diffuse image of a point source, when the lens is limited by a diaphragm of radius r .

The differential equations 17, 18 . . . for the aberration coefficients are linear and subject to solution in the usual manner when A and v are known functions of z . The solutions for the higher coefficients would of course be quite involved. The higher terms are, however, small compared to the second term, which causes most of the aberration, and the approximate distortion is given by the second term alone. This term is called the second order aberration term.

The Reduction of Aberration

The coefficient of any aberration term vanishes when conditions are arranged so that the last member of its differential equation is zero, for the coefficient may be arbitrarily set equal to zero at the first side of the lens, and the solution of its linear equation is then zero throughout the lens.

The important second order aberration term can thus be made to vanish by arranging conditions so that

$$\frac{V''''}{16} - \frac{(A')^2}{2} = 0. \quad (75)$$

In a lens that is not too thick compared to the focal distances involved, we have seen that the term A^2 may be neglected in the differential equation for A , and

$$A' = -\frac{v''}{2\sqrt{2v}}. \quad (76)$$

The substitution of this value for A' in the above equation gives

$$v'''' - \frac{(v'')^2}{v} = 0 \quad (77)$$

as the differential equation for electric fields that are approximately free from second order aberration, when the focal distances are reasonably large compared to the length of the field along the axis.

The general solution of this equation is a series solution, but several particular solutions have been obtained in terms of known functions. The potentials corresponding to these particular solutions are given by the following equations:

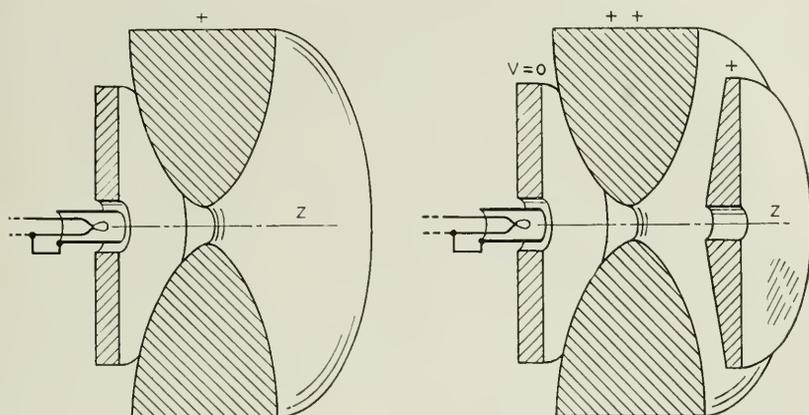
$$\Phi = ae^{\pm\omega z} J_0(\omega r), \quad (78)$$

$$\Phi = (a \sin \omega z + b \cos \omega z) J_0(i\omega r), \quad (79)$$

$$\Phi = (a \sinh \omega z + b \cosh \omega z) J_0(\omega r), \quad (80)$$

$$\Phi = 3az^{3/2} \left[\frac{1}{3} - \frac{1}{4} \left(\frac{r}{2z} \right)^2 + \frac{3}{64} \left(\frac{r}{2z} \right)^4 \cdots \right]. \quad (81)$$

Any one of these electric fields can be produced by shaping and positioning electrodes to correspond with two of its equipotential surfaces. These fields are, however, in general not well adapted to production with practical electrode structures. The one exception is the field defined by equation 79, and electrodes for producing it in a practical form are shown in Fig. 13. They are suitable for giving an



Figs. 13, 14—Lenses with reduced aberration.

electron stream its initial acceleration. The electric field constitutes a divergent lens, as do practically all initial accelerating fields.

As expressed by equation 79, this field is followed by a symmetrically reversed field, and for some purposes it may be desirable to include the reversed field. This is done by locating a low potential electrode along its corresponding equipotential surface as shown in Fig. 14. A small aperture may be cut in this electrode for the passage of electrons. The aperture then acts as a lens to bring the beam to a focus, but this lens has its own aberration, and the whole system is then only partially free from first order aberration.

APPENDIX I—CALCULATIONS FOR A COMPLETE SYSTEM

The electrode arrangement of Fig. 15 is chosen for giving a simple example of the calculations for a complete optical system. The final focal distance is found by calculating the focal distances at the points m , n , o , p in succession. Electrons leave the cathode and travel parallel to the axis in the uniform field between the first and second plates, so their focal distance is $-\infty$ when they arrive at the point m . The electrons then pass through the aperture in the second plate,

and their focal distance at n is calculated from the lens equation 41, which gives

$$\frac{1}{d_n} + \frac{1}{\infty} = \frac{1}{4v_1} \left[\frac{v_1 - v_2}{l} - \frac{v_1}{l} \right] \quad (1)$$

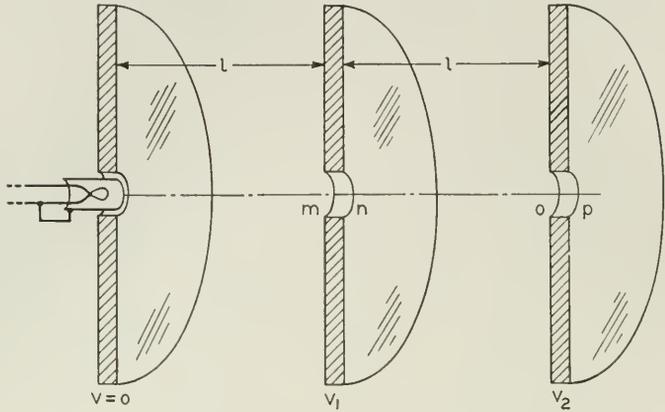


Fig. 15—Example of a complete system.

and the focal distance at n is

$$d_n = \frac{4l}{\beta^2 - 2}, \quad (2)$$

where β is $\sqrt{v_2/v_1}$. The beam then passes through the uniform field between the second and third plates, and the focal distance at 0 is calculated from equation 28 for a uniform field, which becomes

$$(1 + \beta)d_n - (1 + 1/\beta)d_0 = 2l \quad (3)$$

and gives

$$d_0 = 2\beta l \frac{4 + 2\beta - \beta^2}{(1 + \beta)(\beta^2 - 2)}. \quad (4)$$

The beam then passes through the aperture in the third plate into field-free space, and the lens equation for this aperture is

$$\frac{1}{d_p} - \frac{1}{d_0} = \frac{1}{4v_2} \left[0 - \frac{v_2 - v_1}{l} \right]. \quad (5)$$

Substitution for d_0 now gives

$$\frac{1}{d_p} = \frac{1 + \beta}{2\beta l} \left[\frac{\beta^2 - 2}{4 + 2\beta - \beta^2} + \frac{1 - \beta}{2\beta} \right], \quad (6)$$

which is the reciprocal of the final focal distance measured from the last plate.

In complete lens systems, where the symbolic calculations are complicated, it is frequently simpler to introduce specific numerical values and carry the successive steps of the calculation through in a numerical manner. By doing this for a few suitably chosen numerical values one can obtain the particular information that is desired.

APPENDIX II—CONCENTRIC TUBES

Two concentric tubes at different potentials form an electron lens that is well adapted to practical tube construction. When the two tubes are of the same diameter, the approximate constants of the lens may be determined as follows.¹⁸

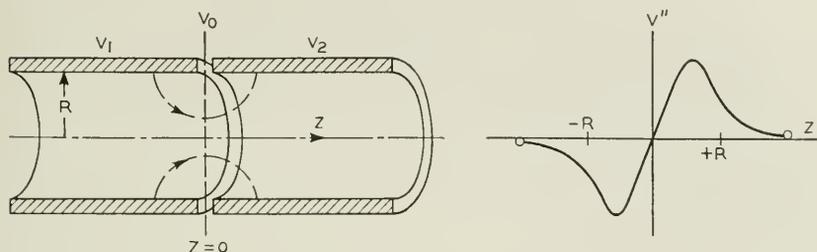


Fig. 16—Concentric tubes.

In this type of lens, the electric intensity is symmetrical with respect to an imaginary plane drawn between the two tubes—as illustrated in Fig. 16—and the plane is therefore an equipotential surface. Its potential v_0 is the mean potential of the two tubes. This plane is regarded as a division plane separating the lens into two component electric fields.

We first consider the component to the right of the plane. The solution for the potential inside of the tube may be obtained in the form of a Bessel Function series, and it follows from this series that the potential on the axis is

$$v = v_2 - (v_2 - v_0) \sum_{\mu} \frac{2}{\mu J_1(\mu)} \exp. \left(-\frac{\mu z}{R} \right), \quad (1)$$

where R is the radius of the tubes, and μ takes on discrete values equal to the successive roots of

$$J_0(\mu) = 0. \quad (2)$$

We find that an approximation to the exponential series is given by

$$\sum_{\mu} \frac{2}{\mu J_1(\mu)} \exp. \left(-\frac{\mu z}{R} \right) = 1 - \tanh \omega z, \quad (3)$$

¹⁸ We assume that the separation between their ends is negligibly small compared to their diameter.

where ω is equal to $1.32/R$. The closeness of this approximation is shown in Fig. 17. Its introduction gives

$$v = v_0 + (v_2 - v_0) \tanh \omega z. \quad (4)$$

A similar approximation is found for the potential on the axis to the left of the division plane,

$$v = v_0 - (v_1 - v_0) \tanh \omega z, \quad (5)$$

and it turns out that the potential on the axis of both tubes can be expressed by the single equation

$$v = \frac{1}{2}[(v_2 + v_1) + (v_2 - v_1) \tanh \omega z]. \quad (6)$$

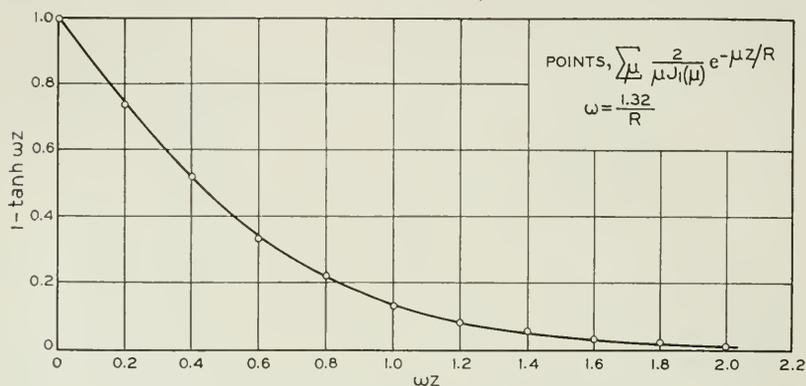


Fig. 17—An approximation for the exponential series.

With the potential on the axis expressed in terms of a known function of z , various series methods may be used for locating the principal planes and calculating the principal focal distances. They are, however, complicated and it may be preferable to use the approximate lens equation obtained by treating the structure as a thin lens.

When treated in this manner, the expression 33 for the inverse focal term can be exactly integrated, and the lens equation is

$$\frac{\sqrt{2v_2}}{d_2} - \frac{\sqrt{2v_1}}{d_1} = \frac{\omega\sqrt{2}}{3(\sqrt{v_2} + \sqrt{v_1})} (\sqrt{v_2} - \sqrt{v_1})^2, \quad \omega = 1.32/R. \quad (7)$$

Division by either $\sqrt{2v_2}$ or $\sqrt{2v_1}$ —as desired—reduces this equation to one that involves the voltages only in the form of a ratio. The error in a focal distance d calculated from this equation is of the order of R , when the focal distance is measured from the division plane of the tubes.

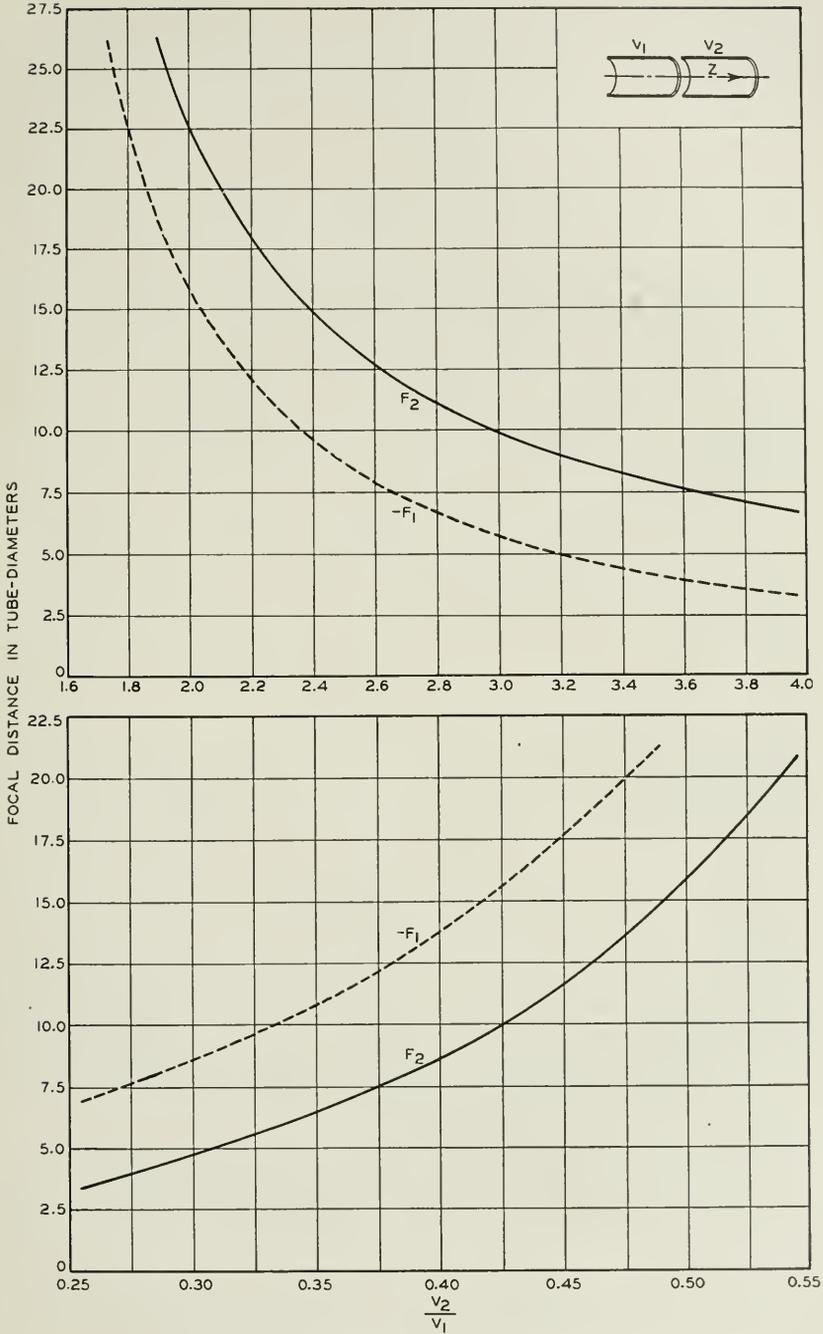


Fig. 18—Principal focal distances—concentric tubes.

The principal focal distances for various voltage ratios are given in terms of the tube diameter by the curves of Fig. 18. For rough calculations, these plotted values may be used in the lens equation

$$\frac{f_2}{d_2} + \frac{f_1}{d_1} = 1, \quad (8)$$

where focal distances are again measured from the division plane of the tubes.

The electric field of the concentric tubes has two maxima of $|v''|$ located symmetrically with respect to the division plane, as illustrated in Fig. 16. Each maximum is located at a distance $.5R$ from the plane. The electron lens may therefore be treated in a somewhat more exact manner by considering it as two thin lenses located at these points. The inverse focal term of the equivalent lens to the left of the plane is

$$\frac{1}{F} = \frac{\omega\sqrt{2}}{v_0 - v_1} \left[v_0(\sqrt{v_0} - \sqrt{v_1}) - \frac{v_0^{3/2} - v_1^{3/2}}{3} \right], \quad (9)$$

and the inverse focal term of the equivalent lens to the right of the plane is

$$\frac{1}{F} = \frac{\omega\sqrt{2}}{v_2 - v_0} \left[v_0(\sqrt{v_2} - \sqrt{v_0}) - \frac{v_2^{3/2} - v_0^{3/2}}{3} \right]. \quad (10)$$

The final focal distance in any particular case is found by carrying out the calculations for the two lenses in succession, with their separation taken equal to R .

APPENDIX III

A Plane Electrode at the End of a Tube.—In addition to their above application, the last two equations may be used for other purposes. In electron devices, one frequently puts a plane electrode at the end of another, tubular electrode.¹⁹ The approximate lens action of the electric field between the plate and tube is then described by one or the other of these equations. Equation 9 applies when the plane follows the tube in the direction of electron motion; and equation 10 applies when the plane precedes the tube.

In structures of this type, the plate is usually pierced with an aperture for the passage of electrons. When the aperture is small compared to the tube diameter, the lens system can be treated in the following manner.

¹⁹ We assume the separation between the plate and the end of a tube to be negligible compared to the tube diameter.

We first consider the case of the plane preceding the tube. The electric intensity at the plate is found by differentiating equation 4 with respect to z and then setting z equal to zero. A substitution of this intensity in equation 41 of the text gives the lens equation of the aperture. In addition to this lens there is an equivalent thin lens located inside the tube at a distance $.5R$ from the plate, and having the inverse focal distance of equation 10. The system is considered as a combination of the two lenses and the calculations are carried through in the usual manner. When the plane follows the tube, the constants of the two lenses are determined from equations 5 and 9, and the combination is treated in a similar manner.

APPENDIX IV—THE VELOCITY FUNCTION

The auxiliary functions u and w are a special case of the components of a generalized vector function that is useful in developing series solutions for electron motion. The equations of this function are equivalent to the Hamilton-Jacobi equation; they are briefly outlined in the present system of units as follows.

In a field that may comprise both an electric intensity E and a magnetic intensity H , let v be any vector function of x , y , z that satisfies the equations

$$\text{curl } v = H/c, \quad (1)$$

$$1/2|v|^2 = \phi + W, \quad (2)$$

where W is a constant equal to the energy of electron emission from the source. Then v is a possible vector velocity for electron motion in the field.

If the magnetic intensity is zero, the vector function v has a potential ψ , which may be any solution of the equation

$$1/2|\text{grad } \psi|^2 = \phi + W \quad (3)$$

and $\text{grad } \psi$ is then a possible vector velocity for electron motion in the field.

The validity of these equations is established by transforming them to the usual equations for electron acceleration.

A LIST OF THE MORE IMPORTANT SYMBOLS AND EQUATIONS

In the present theory of electron-optics, *all distances* along the axis are measured in the direction of motion, as they are in the optics of light.

r, z —cylindrical coordinates
 t —time

- Φ —potential at point in space, the electron source taken as zero potential
- v —potential on the axis
- v' —derivative of v with respect to z
- v —is also used for the voltage of electrodes
- d —focal distance in general
- T —focal time in general
- A —the important coefficient for rays near the axis, a function of z alone
- u, w —velocity functions corresponding to \dot{r} and \dot{z}
- d_1, d_2 —conjugate focal distances measured from the two sides of a lens
- f_1, f_2 —principal focal distances measured from the two sides of a lens
- As an approximation in thin lenses, the focal distances are measured either from the mid-point of the lens, or from the point where $|v''|$ is a maximum, provided that there is but one maximum in the lens.
- α_1, α_2 —location of the principal planes with respect to the sides of a lens
- D_1, D_2 —conjugate focal distances measured from the principal planes
- F_1, F_2 —principal focal distances measured from the principal planes
- F —the focal term of a lens, not a focal distance

Equations for Rays Near the Axis

$$\dot{z} = \sqrt{2v},$$

$$A = -\frac{\sqrt{2v}}{d} = -\frac{1}{T} = \dot{r}/r,$$

$$\sqrt{2v}A' + A^2 = -\frac{v''}{2},$$

$$\frac{1}{T^2} \frac{d}{dt} (T + t) = -\frac{v''}{2}.$$

The important equations for a thin lens are:

$$\frac{\sqrt{2v_1}}{d_2} - \frac{\sqrt{2v_1}}{d_1} = \frac{1}{F},$$

$$\frac{f_2}{f_1} = -\sqrt{\frac{2v_2}{2v_1}},$$

$$\frac{f_2}{d_2} + \frac{f_1}{d_1} = 1,$$

$$\frac{1}{F} = \int_{z_1}^{z_2} \frac{v''}{2\sqrt{2v}} dz.$$

The following equations hold for any lens :

$$\frac{\sqrt{2v_2}}{d_2 - \alpha_2} - \frac{\sqrt{2v_1}}{d_1 - \alpha_1} = \frac{1}{F},$$

$$\frac{\sqrt{2v_2}}{D_2} - \frac{\sqrt{2v_1}}{D_1} = \frac{1}{F},$$

$$\frac{F_2}{F_1} = -\sqrt{\frac{2v_2}{2v_1}},$$

$$\frac{F_2}{D_2} + \frac{F_1}{D_1} = 1,$$

$$M = \frac{F_1 D_2}{F_2 D_1} = -\sqrt{\frac{2v_1}{2v_2}} \frac{D_2}{D_1},$$

where M is magnification.

Equivalent Modulator Circuits

By E. PETERSON and L. W. HUSSEY

Equivalent modulator circuits are developed in the form of linear resistance networks. They are equivalent in the sense that the current magnitude in any mesh of the network is equal to the current amplitude of a corresponding frequency component in the modulator. The elements of the network are determined by the properties of the modulator, while the terminating resistances are those physically existent in the connected circuit.

With this correspondence demonstrated, the operating features of the modulator may be deduced from the known properties of linear networks. Among the properties considered are the transfer efficiency from signal to sideband, and the input resistance to signal as affected by the sideband load resistance.

Equivalent networks are worked out for a number of interesting cases, involving different impedances to unwanted modulation products, together with different non-linear characteristics. The equivalents come out comparatively simple in form under the restrictions noted and followed in the text, which make the carrier large compared to the signal, and the circuit elements purely resistive.

CONSIDERED from the circuit standpoint, a number of modulator performance features are important in any application. Among these features might be mentioned the efficiency of power transfer from signal input to sideband output, and associated with it the question of how the signal input energy is distributed among the different frequency components and dissipated in the modulator itself. Then, too, we need to know how the impedance of the modulator to any component depends upon the modulator structure and upon the connected impedances to other products.

In attempting to get answers to these questions by mathematical analysis, we encounter lengthy and cumbersome expressions in general which do not lend themselves to ready physical interpretation. The physical interpretation of these equations may be facilitated by introducing equivalent circuits of familiar form. One form commonly used in the past replaces the non-linear system by a circuit including a series of generators and linear impedances.

This may be illustrated by reference to a simple non-linear circuit, in which carrier and signal generators are connected through an

external resistance to a two-terminal non-linear element such as a diode, or a copper oxide rectifier. The effects of non-linearity show up in the change of modulator resistance with changes in applied potentials and in the appearance of new frequency components. These effects may be reproduced quantitatively if we replace the non-linear element by its equivalent consisting of a linear internal resistance together with a series of internal generators—indicated at the right of the dashed line of Fig. 1-A. It is easy to see from this

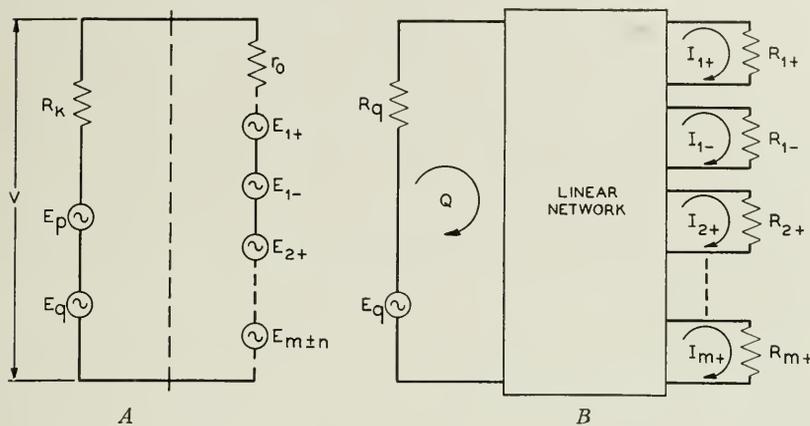


Fig. 1—(A) Equivalent modulator circuit in which the modulator is replaced by a fixed internal resistance together with a series of generators. (B) Equivalent modulator circuit replacing the modulator by a network of linear elements which serves to couple the signal circuit into the paths followed by the modulation products. In this circuit the mesh currents represent the amplitudes of the various frequency components.

circuit¹ what the amplitude of any current component should be; for the general component

$$I_{m\pm n} = E_{m\pm n} / (r_0 + R_{m\pm n}).$$

Despite the apparent simplicity of this relation, a difficulty arises as soon as we attempt to state the internal generator e.m.f.'s explicitly, since they are found to be tied up with the impressed potentials, the

¹ Here we denote the carrier frequency by $p/2\pi$ and the signal frequency by $q/2\pi$; the corresponding generator potentials are E_p and E_q respectively, and the external resistance is R_k where k indicates the frequency at which the resistance is effective.

The new frequencies are usually made up of sums and differences of integral multiples of carrier and signal frequencies. In general, they may be represented by $(mp \pm nq)/2\pi$, where m and n are integers or zero. It is advantageous to adopt an abbreviated notation for the voltage component, say, of any frequency by which the general component is indicated as $E_{m\pm n}$. Further when n is unity it is omitted from the subscript, so that the generator e.m.f. of frequency $(mp \pm q)/2\pi$ is indicated as $E_{m\pm}$. One of the restrictions mentioned further on results in limiting n to unity.

modulator characteristics, and the external circuit impedances. For this reason the equivalent circuit of Fig. 1-*A* reveals only part of the story, and in general the relation between the amplitudes of impressed and generated components remains somewhat obscured.

In a number of cases of practical interest it is possible to represent the connection between the amplitudes of various frequency components by means of a different type of equivalent circuit. Figure 1-*B* is an illustration of this type, in which the paths of the current components are shown individually. The connection between the various circuits is effected by means of a linear network which contains no internal generators. In this equivalent network the magnitude of any mesh current is equal to the amplitude of a corresponding frequency component in the modulator circuit. The purpose of this paper is to demonstrate the validity of this representation, and to show in detail what the linear network looks like when applied to various types of modulating elements, in a variety of interesting cases.

In order to develop such equivalent networks in simple and useful form, the following restrictions are imposed. The system includes only one non-linear element.² The terminating impedances are purely resistive, although they may be functions of frequency. The signal amplitude must be much smaller than that of the carrier. Finally, the slope of the modulator current-voltage characteristic never becomes negative. Under these conditions a number of modulating systems can be treated, including variable resistance modulators with a variety of current-voltage characteristics, and the variable resistance microphone.

The section following deals with the modulator as a resistance (or conductance) varying at carrier frequency. Succeeding sections consider the behavior of such variable elements under different circuit conditions.

I. CARRIER CONTROLLED RESISTANCE

In setting up equations from which the equivalent networks are obtained, the restriction on signal amplitude permits us to assume the modulator to be a resistance or conductance varying at carrier frequency. This commonly used assumption may be arrived at with the aid of Fig. 2, which shows a typical non-linear current-voltage

² Other cases are to be found in a paper by R. S. Caruthers on "Copper Oxide Modulators in Carrier Telephone Systems," presented at the A.I.E.E. Winter Convention, January 1939.

Modulators including a plurality of elements can frequently be replaced by an equivalent structure with a single modulating element. This is true of the rectifier type of modulator. In the double-balanced or ring type, however, under certain conditions the equivalent circuit involves merely an ideal transformer connecting signal and sideband circuits.

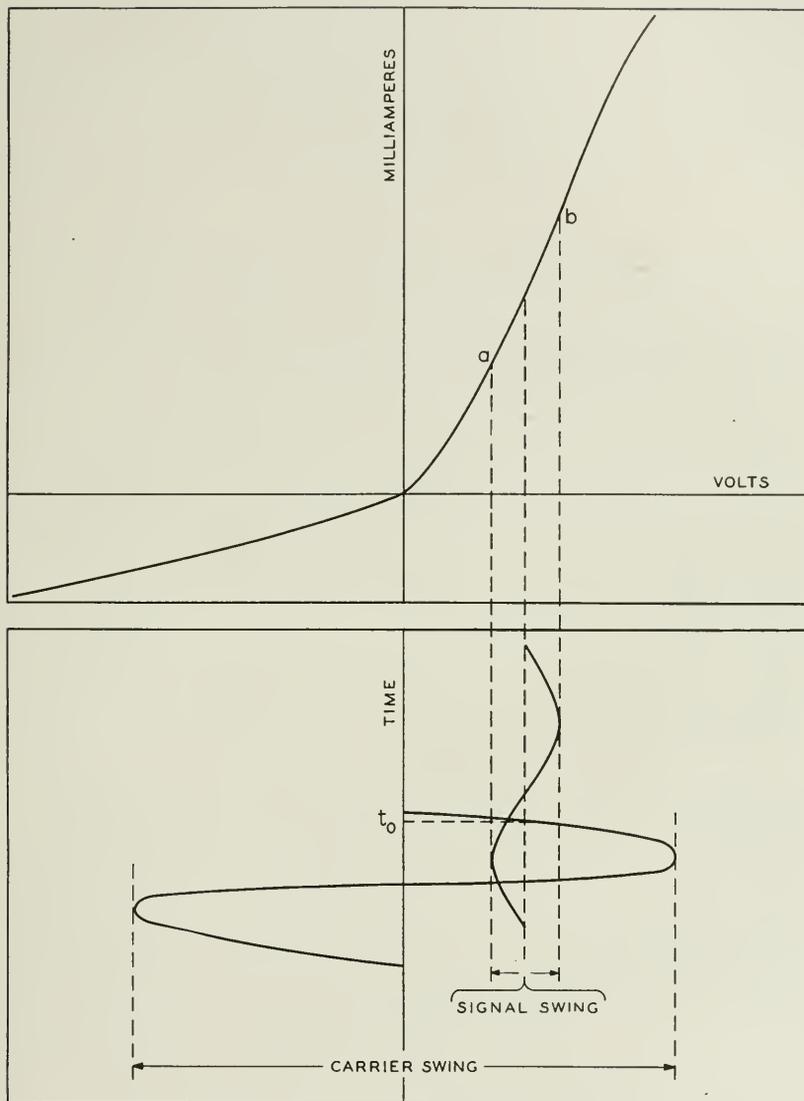


Fig. 2—Non-linear current-voltage relation representative of certain types of modulators. The impressed carrier voltage is large in comparison to the signal, the variation of which is represented in the neighborhood of the potential reached by the carrier wave at time t_0 .

curve. The variation with time of a large carrier voltage and a small signal voltage are also indicated. Now if we consider the carrier to provide a varying bias, then at any typical point t_0 we can consider the signal voltage to sweep over a small segment (a, b) of the modulator characteristic. The resistance is practically constant over this segment and of magnitude

$$R = \frac{dv}{di}, \quad (1)$$

the derivative being evaluated at the carrier voltage under consideration. In general this resistance varies from point to point of the carrier cycle. Thus the carrier enters the signal-sideband relation only through the variation of a resistance facing the signal and modulation products.

If the current-voltage characteristic is a smooth curve the resistance varies smoothly over the carrier cycle. If the characteristic is made up of two straight lines the resistance switches between two constant values. This latter is approximated by most rectifiers, such as diode and copper-oxide rectifiers with suitably large carrier amplitudes; it is called by analogy a commutator modulator.

As a simple example of a variable resistance consider the characteristic

$$i = av + bv^2, \quad (2)$$

from which

$$1/R = \frac{di}{dv} = a + 2bv. \quad (3)$$

If the impressed carrier potential, v , is $P \cos pt$, the conductance is

$$G = \frac{1}{R} = a + 2bP \cos pt. \quad (4)$$

More generally, the resistance or conductance for a given characteristic can be expressed as a series

$$R = r_0 + \sum_1^{\infty} 2r_n \cos npt, \quad (5a)$$

or

$$G = g_0 + \sum_1^{\infty} 2g_n \cos npt. \quad (5b)$$

Here the coefficients depend only on the modulator characteristic and the carrier amplitude. In special cases some of the coefficients vanish.

Thus an expansion for the linear rectifier includes only those coefficients for which n is odd, whereas one for a modulator exhibiting odd symmetry in its current-voltage relation, such as thyrite, includes only coefficients for which n is even.

The choice between (5a) and (5b) in any given case is usually a matter of convenience.³ This will be made clear by the forthcoming examples. In every case use will be made of Ohm's law in one of the two forms

$$v = Ri$$

or

$$i = Gv.$$

For simplicity, we select the relation which leads to the smallest number of terms in the expansion. Thus if, from the form of the terminating impedance, we know that i involves only a small number of significant frequency components, whereas the voltage involves a large number, (5a) will be used. If the potential, v , across the modulating element is known to be the simpler, (5b) will be used.

In the practical application of modulators to carrier systems the impedance characteristics of the connected selective circuits for taking out the desired sideband energy provide, to a good approximation, just such simplification. Thus a filter is substantially resistive in its pass band and the suppression regions may be designed to have either a very high or a very low impedance. If very high, no currents flow in these frequency regions and (5a) applies; if very low no potentials appear across it in these frequency regions so that (5b) applies.

II. SINGLE SIDEBAND—HIGH IMPEDANCE OUTSIDE BAND

We will first consider a single sideband modulator involving any variable resistance which can be expressed in the form (5a). The terminating resistance is R_q to signal and R_{1+} to the upper second order sideband. Because of the high terminating impedance which we assume to all other products, all current components other than signal (Q) and sideband (I_{1+}) are negligibly small.

The total current flowing in the circuit is then

$$i = Q \cos qt + I_{1+} \cos (p + q)t. \quad (6)$$

The potential across the non-linear element ($v = Ri$) is obtained from (5a) and (6) as

$$v = \left[r_0 + \sum_1^{\infty} 2r_n \cos npt \right] [Q \cos qt + I_{1+} \cos (p + q)t]. \quad (7)$$

³ Except for those cases in which the occurrence of an infinity in any one of these two quantities prohibits its use.

After multiplying, and separating out the different frequency components, each frequency component of v is equated to the corresponding terminating generator e.m.f. minus the potential drop across the external impedance. Carrying out this process for signal and sideband, respectively,

$$\begin{aligned} E_q &= (R_q + r_0)Q + r_1 I_{1+}, \\ 0 &= r_1 Q + (R_{1+} + r_0)I_{1+}. \end{aligned} \quad (8)$$

If Q and I_{1+} are considered as mesh current amplitudes in a simple linear circuit, it is evident that r_1 represents a mutual resistance, and that Fig. 3 represents an equivalent network. In this system the

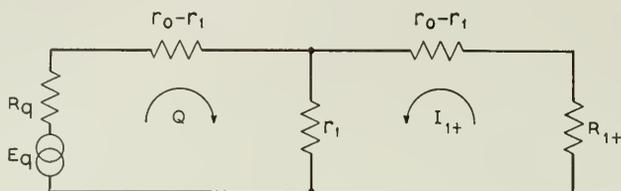


Fig. 3—Equivalent modulator network connecting signal and sideband when other modulation current components are suppressed by high circuit impedances.

signal source is connected to the sideband load by a simple T network.⁴ It will be found in subsequent cases similarly, that the connection between signal and sideband circuits may be effected by a network comparatively simple in form. Hence we can make our deductions concerning the performance of modulator circuits by reference to the well known properties of such equivalent networks. In the present case, for example, we can draw the following conclusions.

1. The modulator loss becomes negligibly small if the series arm resistance, $r_0 - r_1$, is very small and the shunt arm resistance, r_1 , is relatively large.
2. Considering the modulator network as fixed, maximum power is transferred when the signal and sideband resistances match the characteristic resistance of the network, so that

$$R_q = R_{1+} = \sqrt{r_0^2 - r_1^2}. \quad (9)$$

3. Under matched impedance conditions the power efficiency is

$$\eta = \left[\frac{r_1}{r_0 + \sqrt{r_0^2 - r_1^2}} \right]^2. \quad (10)$$

⁴ While the results come out most simply in terms of a T network, the various possible transformations (for example to a π or to a lattice network) are of course equally valid.

The term *power efficiency*—as used here—means the ratio of the power delivered to the load resistance (R_{1+}) to that introduced at the input side of the network by the signal source. The corresponding current ratio of sideband to signal is the square root of η . If the sideband resistance is shorted the current ratio rises to its maximum. The ratio of voltage at the network output to that at the network input when the load resistance is made infinite coincides with this value (r_1/r_0).

If we consider the various possible kinds of resistance variation with time under the restrictions noted, it appears that the greatest attainable value of the power ratio is unity. It is evident from the equivalent circuit that this limit corresponds to no loss from signal to sideband. The closest approach to this no-loss condition is obtained in a modulating element presenting a resistance which, over a carrier cycle, varies between widely different resistances, taking on one extreme value for a small fraction of the cycle and remaining near the other extreme for the remainder of the cycle. Under these conditions the series arm of the equivalent net tends to zero, the shunt arm to infinity. There are practical limitations to the extent to which these conditions can be approached in practical modulators. For example the best attainable values of the two resistance extremes usually depend upon the modulator characteristic, and upon the carrier amplitude employed which may be limited by heat dissipation or by voltage breakdown. Further limitation is imposed by parasitic capacitances, which effectively limit the maximum attainable modulator resistance.

The commutator modulator may be used to illustrate the results of analysis above. Figure 4 shows the variation of the resistance of

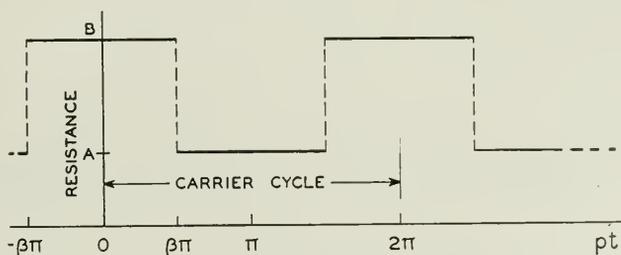


Fig. 4—Variation of resistance with time in a commutator modulator, in which the resistance is switched from A to B , remaining at the higher value B for the fraction β of the carrier cycle.

such a modulator over a carrier cycle. B and A are the two values of the resistance ($B > A$) and β is the fraction of the carrier cycle

over which the resistance is B . The coefficients of the resistance expansion are readily shown to be

$$r_0 = \beta B + (1 - \beta)A, \quad (11)$$

$$r_k = (B - A) \frac{\sin k\beta\pi}{k\pi}. \quad (12)$$

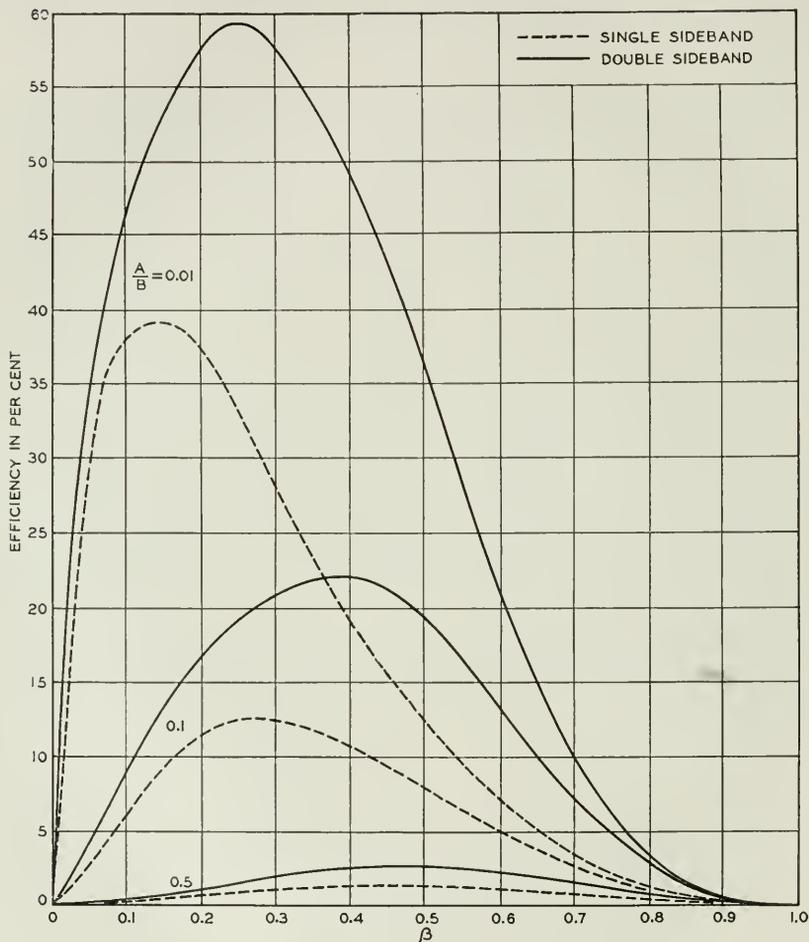


Fig. 5—Efficiency of a commutator modulator shown as a function of the pulse fraction β with the ratio of the two commutator resistances as parameter. Here the resistance terminations to signal and to sideband are optimum and the circuit impedance is made high to unwanted components. Full line applies to double sideband output, dashed line to single sideband output.

The efficiency (with optimum termination) depends only on β and the ratio A/B . The dashed curves on Fig. 5 show the variation of effi-

ciency with these parameters. It is evident from the equivalent circuit that best efficiency is obtained with $r_0 - r_1$ small and r_1 large. In this case r_0 increases linearly with β , and r_1 varies sinusoidally with β —having its maximum at $\beta = 0.5$. The immediate conclusion is that the optimum conditions are obtained when β is less than 0.5. In fact the efficiency approaches the limiting value of 100 per cent when β is very small and B is much greater than A , as may be seen from (11) and (12).

III. DOUBLE SIDEBAND—HIGH IMPEDANCE OUTSIDE BAND

In double sideband operation both upper and lower sideband currents flow, but all other modulation products are suppressed as in the previous case. Here the equations for signal and upper and lower sideband respectively are

$$\begin{aligned} (r_0 + R_q)Q + r_1 I_{1+} + r_1 I_{1-} &= E_q, \\ r_1 Q + (r_0 + R_{1+})I_{1+} + r_2 I_{1-} &= 0, \\ r_1 Q + r_2 I_{1+} + (r_0 + R_{1-})I_{1-} &= 0. \end{aligned} \tag{13}$$

Comparing (13) with the equations for a three-mesh circuit, we obtain the equivalent network of Fig. 6. It is obvious from the symmetry of this network that the two sidebands are equal when $R_{1+} = R_{1-}$. Conditions for optimum efficiency may be put in form permitting convenient comparison with the single sideband case when we assume equal resistances to both sidebands.

Efficiency curves of a commutator modulator are shown on Fig. 5 for both single and double sideband cases. They differ primarily in that the utilization of two sidebands gives greater efficiency, except in limiting cases. The outstanding difference is that the unsymmetric network has optimum signal and sideband resistances which are not equal except at three values of β equal to 0, 1/2 and 1. Modulators are often operated with β approximately 1/2, so that in this case the results here check with the common experience that the two terminating resistances should be equal. It may be remarked that only in highly efficient modulators would unequal terminations make an appreciable difference in the efficiency of power transfer.

A comparison of Figs. 3 and 6 gives some light on the difference in efficiency of the single and double sideband cases. The comparison is made when $R_{1+} = R_{1-}$. From the symmetry of the circuit of Fig. 6, I_{1+} then equals I_{1-} and the mutual resistance ($-r_2$) may be eliminated, leaving a simple T network connecting the input and the load. This is, with two exceptions, the T network of Fig. 3 with all

elements doubled in magnitude. The input series arm is decreased ($r_0 - 2r_1$ instead of $2r_0 - 2r_1$) and the output series arm is increased by an element $2r_2$. Since $2r_2$ is generally much smaller than r_0 there is a net gain in efficiency.

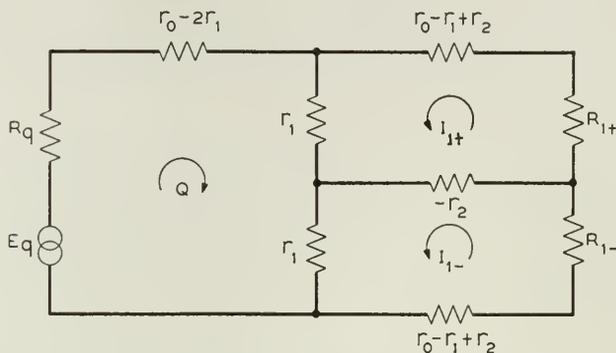


Fig. 6—Equivalent modulator circuit showing the connection between signal and two sideband circuits. All other modulation current components are suppressed by means of high circuit impedances.

IV. LOW IMPEDANCE OUTSIDE BAND

The foregoing systems involved high impedances, suppressing the flow of current at all but two or three frequencies. Circuits are as readily obtained in the case of low impedance to all but two or three of the modulation products. The physical systems this approximates are the same as those previously discussed except that the terminating filters must present a low impedance to frequencies outside the band.

All the external potential drops across the modulating element are taken as negligible except components at the signal and one or two sideband frequencies. The analysis, corresponding to that of the previous sections, uses Ohm's law in the form $i = Gv$. Thus this analysis, and the equivalent circuits, involve the expansion of a conductance instead of a resistance. The equations corresponding to (8) and (13) are equations in V_q , V_{1+} and V_{1-} . In the single sideband case a resulting equivalent circuit is that of Fig. 7. This is a simple symmetric π network. From its well known characteristics the optimum terminating conductance and maximum efficiency are immediately available:

$$G_{1+} = \frac{1}{R_{1+}} = \sqrt{g_0^2 - g_1^2}, \quad (14)$$

$$\eta = \left[\frac{g_1}{g_0 + \sqrt{g_0^2 - g_1^2}} \right]^2. \quad (15)$$

These expressions are identical in form with corresponding ones obtained for the high-impedance single-sideband case, in which conductance components replace resistance components. This con-

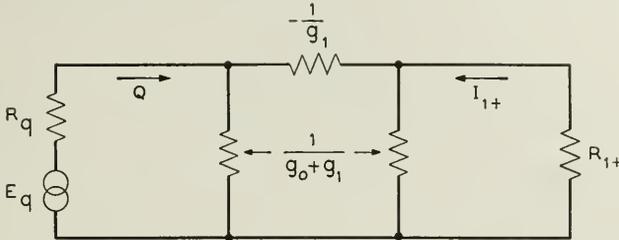


Fig. 7—Equivalent modulator circuit showing connection between signal and a single sideband. All other modulation voltage components are suppressed by means of low circuit impedances. The use of a Π network and the specification of element values as conductances are both matters of convenience.

firms what has been observed in several special cases: that there is no theoretical advantage of either impedance extreme over the other in the general case.⁵ Which one to use in any particular case depends upon the special characteristics of the modulator or upon the practicality of obtaining required impedance conditions.

The equivalent network for the corresponding double sideband system is shown in Fig. 8. Similarly to the previous double sideband case, the symmetry shows that when $R_{1+} = R_{1-}$ the sideband current amplitudes are equal and there is no potential across the coupling resistance $-(1/g_2)$. Thus it may be shorted, reducing the circuit to a simple unsymmetric π . The matching resistances and maximum efficiency may be obtained as before.

The results again are identical with the high-impedance case, if resistances are replaced by conductances. The comment made on the single-sideband case still holds—that there is no general theoretical advantage of either a high- or low-impedance system over the other as far as maximum possible efficiency is concerned.

The curves of Fig. 5 are evidently immediately applicable to the low-impedance circuits provided all resistances are replaced by conductances.

There are, of course, practical advantages of the high- or the low-impedance circuit in particular cases. For example, it is commonly easier to make the terminating impedance to unwanted frequency

⁵ The form of the equations in corresponding high- and low-impedance cases suggests that the impedance and efficiency relations for one case could be deduced from those of the other through the principle of duality. See Guillemin, "Communication Networks," Vol. 2.

components very small rather than very large. The impedance matching may be simpler in one case than in the other since, for the same efficiency, the matching resistances are quite different in the two cases.

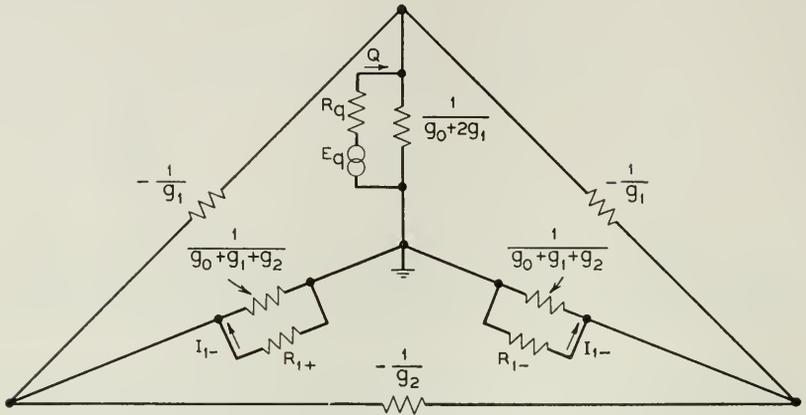


Fig. 8—Equivalent modulator circuit showing connection between signal and two sidebands. All other modulation voltage components are suppressed by low circuit impedances. When one of the sideband paths is shorted, the network reduces to that of Fig. 7.

V. FINITE RESISTANCE TO DISTORTION PRODUCTS

One more extension may be made, without excessively complicating the equivalent circuit. This is an extension to the case of a constant resistance R to all unwanted products, which yields information unobtainable from the limiting cases previously treated of $R = \infty$ and $R = 0$.

This problem can be handled in a simple way by the artifice of incorporating R within the modulator proper. In that case the external impedances to signal and sideband must be reduced by R to keep the total circuit resistance at its correct value, while the external resistance to any other modulation product then becomes zero. This brings the situation down to that considered and solved in the section immediately preceding.

By this manipulation the equivalent circuit is obtained as that of Fig. 9 in the single sideband case. The primes indicate coefficients in the expansion of the modified characteristic. These coefficients are immediately available in the case of the commutator, since the sole change there is an increase in both values of the variable resistance by the amount R . Comparing the efficiency of transformation with that obtained with extreme values of R as in the cases preceding, it appears

desired, they may be obtained from equations (16), using the known values of I_0 and I_1 .

The two-mesh circuit for I_0 and Q can immediately be put in the form of an unsymmetric T terminated at one end by the battery and its internal impedance and at the other end by R_q . The optimum terminating resistances and corresponding efficiencies are obtainable as in previous cases but it is evident from the network, without further computation, that losses are minimized (with suitable termination) if $R_0 - R$ and R_i are small compared with R . R_i is decreased by decreasing R_{nq} ($n > 1$). These conditions mean that the best electrical efficiency is obtained when the resistance variation is large and the unwanted signal harmonics are short circuited.

VII. EXTENSIONS AND SUMMARY

Equivalent networks can be obtained in some cases when the restrictions on the relative amplitudes of signal and carrier are removed. It is evident from Fig. 2 that the value of the variable resistance at any instant then depends not only on the carrier amplitude but also on the signal amplitude. Thus the equivalent networks are no longer made up of constant resistances, but depend upon the magnitudes of both signal and sideband components. Further, new components appear involving multiples of the signal frequency. The equivalent for this case lacks the simplicity of those discussed here, a simplicity which appears when one of the two input components is much greater than the other.

The reason for the restriction to pure resistances becomes evident when one attempts to generalize the results. The current components will then have phase angles differing from zero in general. Consideration of lower sidebands then shows that the phase angles must have their signs reversed in certain circumstances, which leads to obvious complexities. Again in purely resistive circuits it is possible to determine the instantaneous current-voltage relation and hence to specify the resistance variation as a function of time. In a reactive circuit, however, additional difficulty arises in that the relations are much more complex and in general impossible to specify in simple terms.

To summarize, the presentation has been limited to the simplest circuits used for modulation by means of a variable resistance. In each example, the inter-relations between modulation product amplitudes, terminating resistances, and types of modulator characteristics are shown in terms of familiar linear resistance networks. From these, qualitative information concerning the properties of the system is

more readily obtained than from the equations and, in some cases, the solutions for effective impedances and current and voltage amplitudes are obtainable without further recourse to the equations.

ACKNOWLEDGMENT

The writers are indebted to several of their associates in the Bell Telephone Laboratories for the use of unpublished material in this paper. In particular, acknowledgment is due to Mr. R. S. Caruthers, Mr. J. M. Manley, Dr. G. R. Stibitz, and Mr. R. O. Wise, who originally obtained some of the impedance and efficiency relations.

An Improved Three-Channel Carrier Telephone System

By J. T. O'LEARY, E. C. BLESSING and J. W. BEYER

This paper describes an improved three-channel carrier telephone system for use on open-wire lines. It employs recent advances in the telephone art to bring about many economies and circuit simplifications as compared with previous models of the three-channel system. A new type of automatic regulating equipment is included.

INTRODUCTION

THERE are now in service in the Bell System approximately 750,000 miles of telephone circuit which are furnished by carrier systems. Of this total, almost 90 per cent is provided by some 600 Type C systems, ranging from about 75 miles to over 2000 miles in length. Basically designed to add three carrier channels to the normal voice channel on open-wire lines, the Type C system has also been used in special cases to provide additional circuits over deep sea cables of moderate lengths.

The system was first described in this Journal in the July 1928 issue.¹ Improved designs and the application of new circuit elements have recently permitted a very extensive revision of the terminal and repeater equipment which results not only in striking reductions in size and cost as compared with the older equipment, but also gives a considerable improvement in transmission performance. A new type of automatic regulating equipment has been provided for both the terminal and repeater.

The improved system employs heater type pentode tubes, copper-oxide modulators and demodulators and makes use of the negative feedback type of amplifier at both terminal and repeater points. The terminal band filters are newly designed to give improved transmission frequency characteristics on all channels. Each channel is arranged to terminate on a four-wire basis in the same manner as the Type K system for cables.²

An outstanding feature of the modified design is the large saving in space in comparison with the previous equipment. As shown on Fig. 1, the complete terminal with its regulating equipment occupies a single bay, whereas the older system without regulating equipment required two and one-half bays. The repeater space savings, while not so large, are nevertheless substantial. The number of vacuum

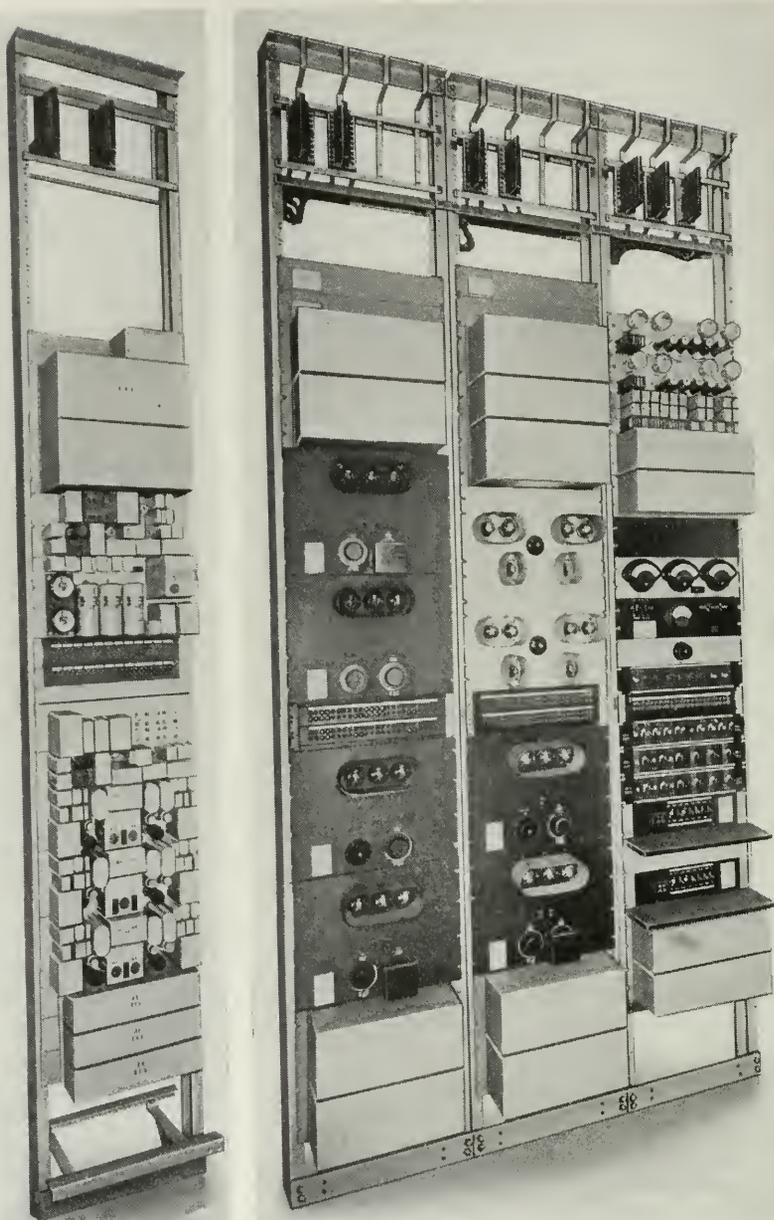


Fig. 1—New and old type C carrier telephone terminals.

tubes required in the system has been reduced, which results in a material saving in power.

Certain features of the improved equipment, notably the automatic regulation, can also be used on the older types of systems and the design objectives were set up with this in view.

FREQUENCY ALLOCATION

The frequency range employed by the system and the allocation of channel bands within that range are shown in Fig. 2. The allocations used in the older systems are also shown for comparison.

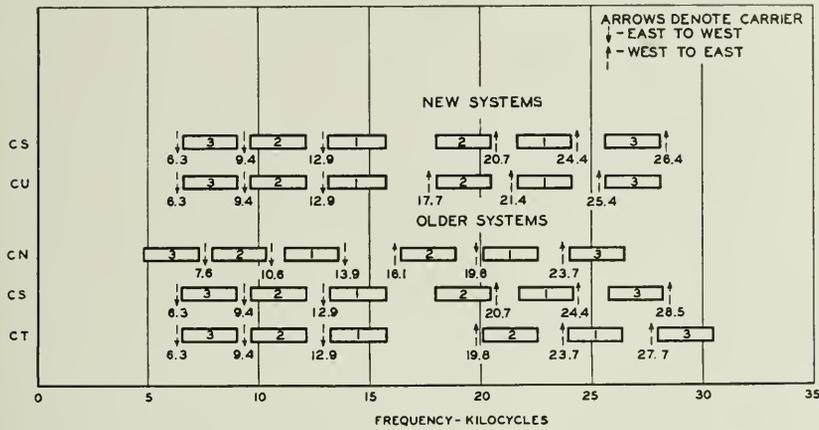


Fig. 2—Frequency allocation of the new systems relative to that of the older systems.

The original selection of the frequency range for the Type C system was the result of many different factors. Foremost among these was the desire to keep the frequencies low in order to minimize line crosstalk and attenuation and changes in the latter due to weather and temperature. On the other hand was the greater filter cost that results from crowding the channels close together. Different frequency bands were used for transmitting in opposite directions in order to avoid the problem of near-end crosstalk and to give the advantages of four-wire transmission. Although consideration was given to the general desirability of increasing the band of frequencies effectively transmitted by the individual channels, the requirement for coordinated operation with older systems already in use precluded any material increase in the frequency space allocated to each channel of the new system. Nevertheless, as will be seen from Fig. 4, the channels show very little distortion within the transmitted band and represent a material improvement over the older systems.

Because the line crosstalk tends to be greater at the higher frequencies, past experience has indicated the advantage of having available two systems between which the crosstalk in the higher frequency group will be unintelligible. Two allocations are provided for this purpose, designated CS and CU. The channel bands are identical in the lower frequency group (East to West) and in the upper frequency group (West to East) differ only in that the carrier frequencies are at opposite ends of the bands. In this group crosstalk between similar bands will have the speech frequencies inverted and will therefore be unintelligible.

This arrangement does not give as high a crosstalk advantage as the arrangement used previously where the bands were not only inverted but also displaced with respect to each other. However, better line crosstalk conditions now prevail due to the application of improved transposition designs and line configurations to the more recently constructed lines and to the use of new methods of mitigating crosstalk on the older lines. This permits the simplification of the frequency allocation, as a result of which one system may be readily converted into the other with fairly simple equipment changes. It will also be possible to use the voice frequency circuit on all pairs as a program circuit transmitting up to 5000 cycles. The advantages of the greater plant flexibility resulting from these two factors are obvious.

The new system may be used on suitably transposed lines with the Type D³ and Type H⁴ single channel systems, whose frequency bands are such that no serious near-end crosstalk problem will arise.

OVERALL SYSTEM

A block diagram of a complete system, consisting of the two terminals and a single intermediate repeater, is shown on Fig. 3. In practice there might be as many as ten or more such repeaters. The two terminals differ from each other only in the frequencies for which their respective transmitting and receiving circuits are designed. The west terminal transmits the high-frequency group of Fig. 2 and receives the lower frequency group while the east terminal does the reverse. The repeater is provided with means for separating the frequencies in the two directions of transmission, amplifying the current to the desired level, and passing them on to the next line section.

A typical overall frequency characteristic for one of the circuits derived from the new system is shown on Fig. 4. This characteristic illustrates the relative freedom from distortion in the transmitted frequency range.

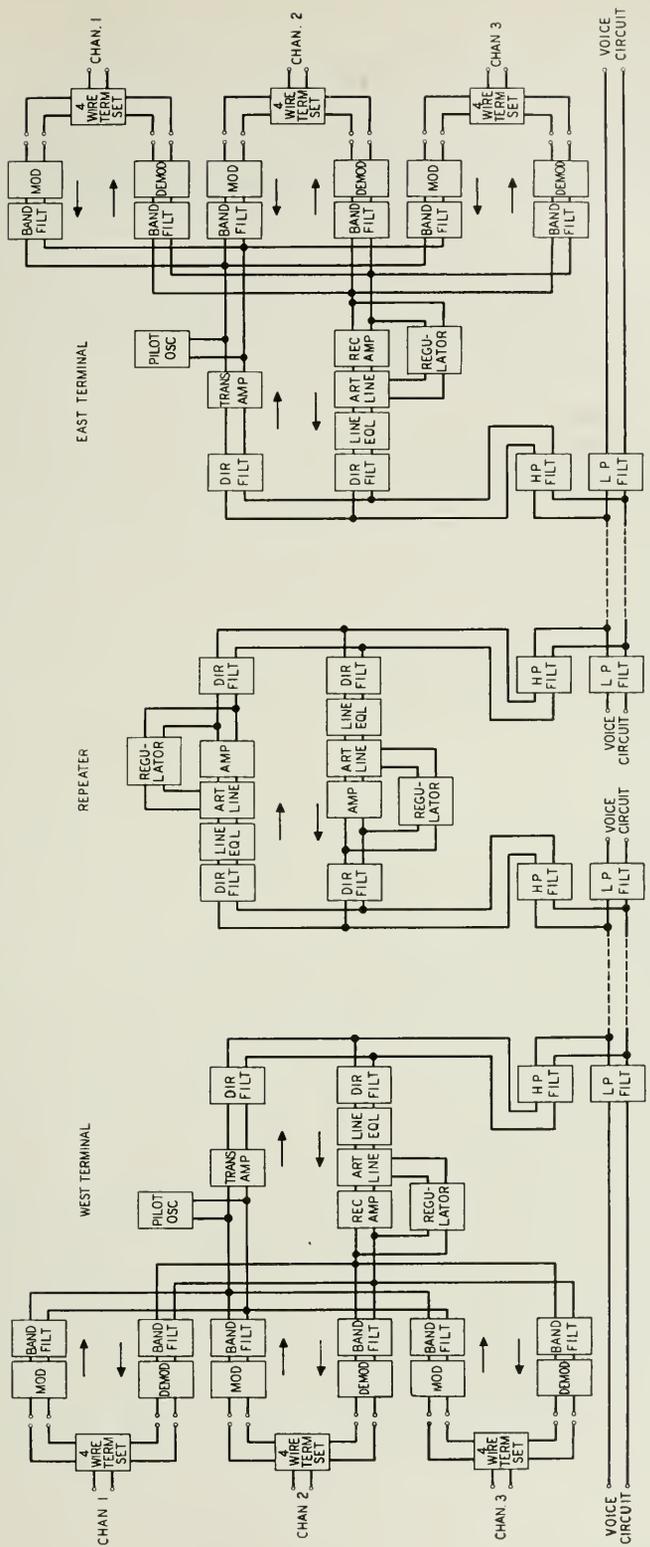


Fig. 3—Schematic of an overall system with one repeater.

The carrier channels are separated from the voice frequency circuit on the same pair of wires by means of a high-pass and low-pass filter combination as shown on Fig. 3. Several different filter sets are available for this purpose differing from each other in the frequency

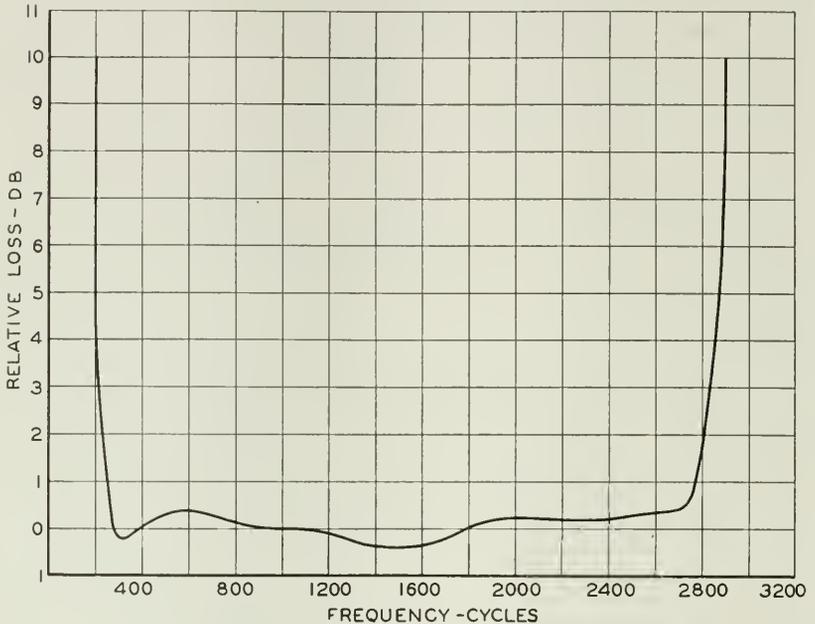


Fig. 4—Typical overall transmission-frequency characteristic.

band which is desired in the voice circuit. Where the voice circuit is an ordinary message circuit the filter will have a cutoff around 3 kc. Where it may be used for program transmission a filter set having a cutoff above 5 kc is provided. For still wider program bands there is a filter set cutting off above 8 kc. The use of this latter filter would, of course, require the sacrifice of the lowest carrier channel since it would be overlapped by the program band.

An important feature of the system is the method of stabilizing the overall transmission. Ahead of the terminal transmitting amplifier in each direction of transmission there is connected to the circuit an oscillator which generates a pilot current. This pilot current has a frequency adjacent to the band of the middle channel. The oscillator is designed to have a relatively high degree of stability with respect to output and frequency. At the output of each repeater and at the receiving terminal the pilot frequency is selected by a high-impedance bridging filter, which has little effect on the through transmission,

and is then used to actuate a regulating mechanism. Changes in the line transmission at this frequency are indicative of the changes at all frequencies and the regulator functions to maintain a nearly constant output level in all three-channel bands.

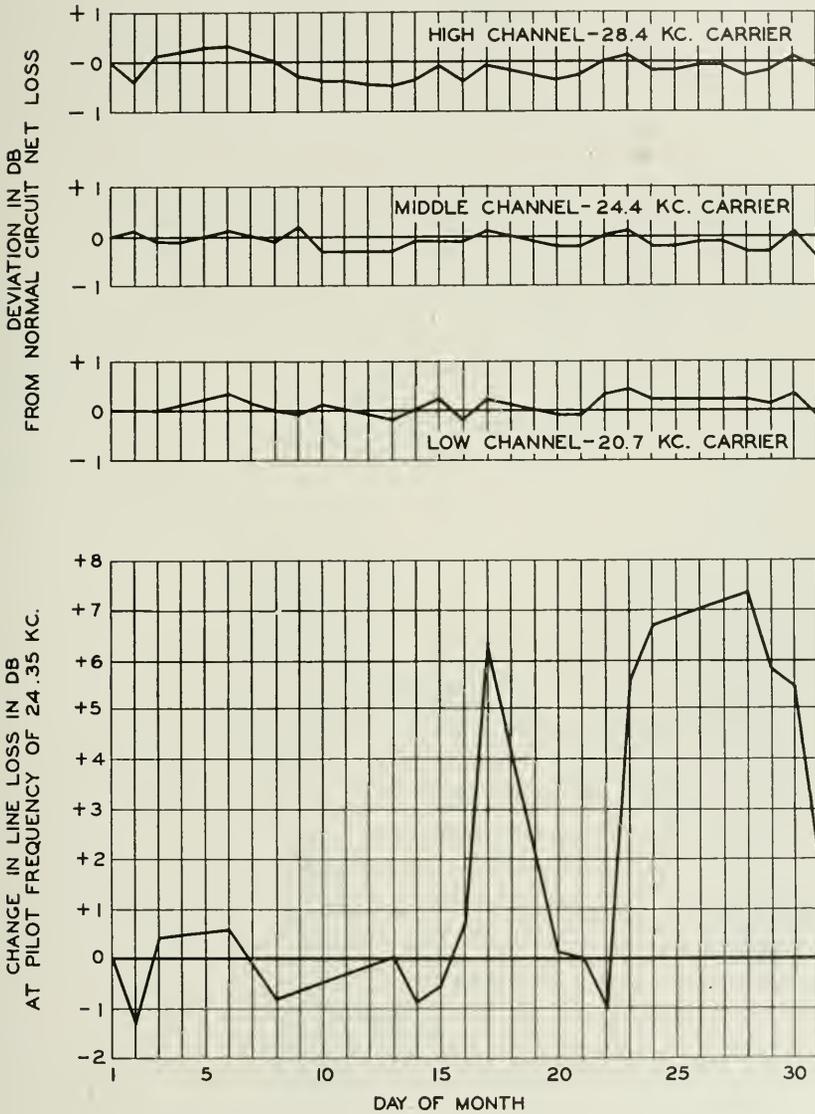


Fig. 5—Chart showing regulation for a period of one month on system with one repeater.

The pilot current is also used to indicate through an audible or visual alarm any trouble which results in large sudden changes in transmission such as would be occasioned by an open or short circuit on the line itself.

The ability of the regulating mechanism to stabilize the transmission over the system is shown on Fig. 5 which shows the deviations recorded in daily measurements on all three channels of a 250-mile system over a period of one month. The actual changes in line loss at the pilot frequency are also shown for comparison. During this period various conditions of temperature, rain and fog were experienced.

With the transmitting level that has been provided and for ordinary line conditions it is found practicable to employ repeater spacings of from 125 miles to over 250 miles. The exact distance in any particular case depends upon many factors, such as: wire size, length of toll entrance or intermediate cables, location of existing offices and the susceptibility of the line to sleet or frost. Where this latter condition is prevalent conservative spacings are desirable.

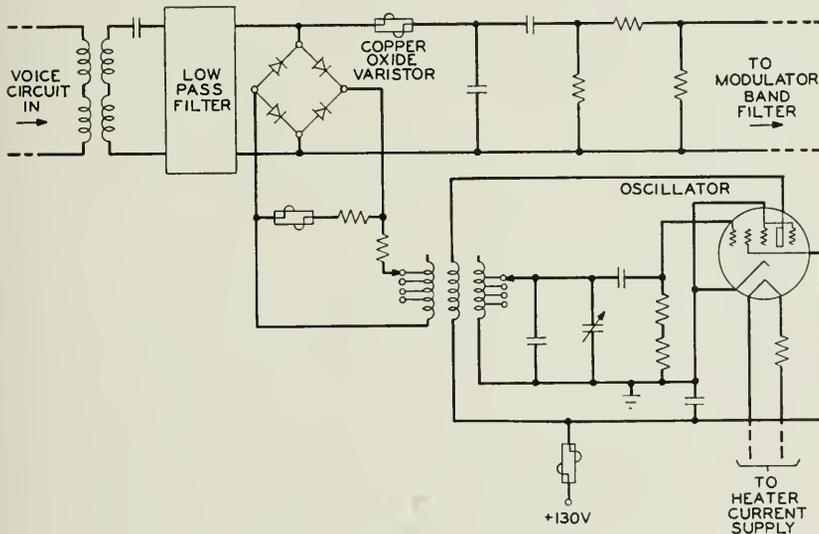


Fig. 7—Schematic of modulator.

TERMINALS

The general theory of operation of the terminal may be understood from the block diagram shown in Fig. 3. On the voice-frequency side each channel terminates as a four-wire circuit. The input to the carrier system from the voice circuit is designed to operate at a level

13 db below the transmitting toll switchboard which is the common reference point. The output from the system is at a level 4 db above that point. Equipment for coupling the system to both two-wire and four-wire circuits has been designed. The circuits employed in each case may be seen in Fig. 6.

The modulator circuit, shown in Fig. 7, uses copper-oxide varistors⁵ for converting the voice frequencies to the higher line frequencies. The high degree of balance obtained in the copper-oxide varistors has the important advantage of making carrier leak a practically negligible factor. This is of particular importance in the case of that channel to which the pilot current is adjacent in the frequency spectrum. The modulator circuit is also designed to limit the peaks of very loud talkers which would otherwise overload the common amplifiers. The effect of this limitation on the quality of the speech transmitted is not noticeable.

The oscillator which supplies the carrier to the modulator is designed to be stable in both output and frequency. When it is once adjusted with the oscillator at the distant end, departures from synchronism will be relatively small. Part of this stability is due to a new circuit design employing coil and condenser elements having opposite temperature coefficients so that changes in one will be compensated for by changes in the other.

The band filters use coils wound on magnetic core material, having improved modulation characteristics, instead of the solenoidal air core coils previously used. This results in a considerable reduction in the space which they occupy.

The transmitting and receiving filters associated with each channel are identical as to band width. They are further characterized by a more abrupt increase in discrimination immediately below and above the pass-band frequencies than was realized in the channel filters for the previous Type C systems and also by less distortion across the pass band. Most of this distortion is in the form of higher loss in the vicinity of the band limiting frequencies. It was deliberately included in the design of the filters for the purpose of masking delay distortion effects on overall transmission quality which might otherwise become noticeable when four or five type C carrier telephone systems are connected in tandem.

The uniformity and symmetry of the various filters are shown by Fig. 8 which gives the characteristics of those in the upper frequency group. This symmetry is required in this group in order to make the CS allocation convertible into the CU by moving the carrier from one end of the band to the other.

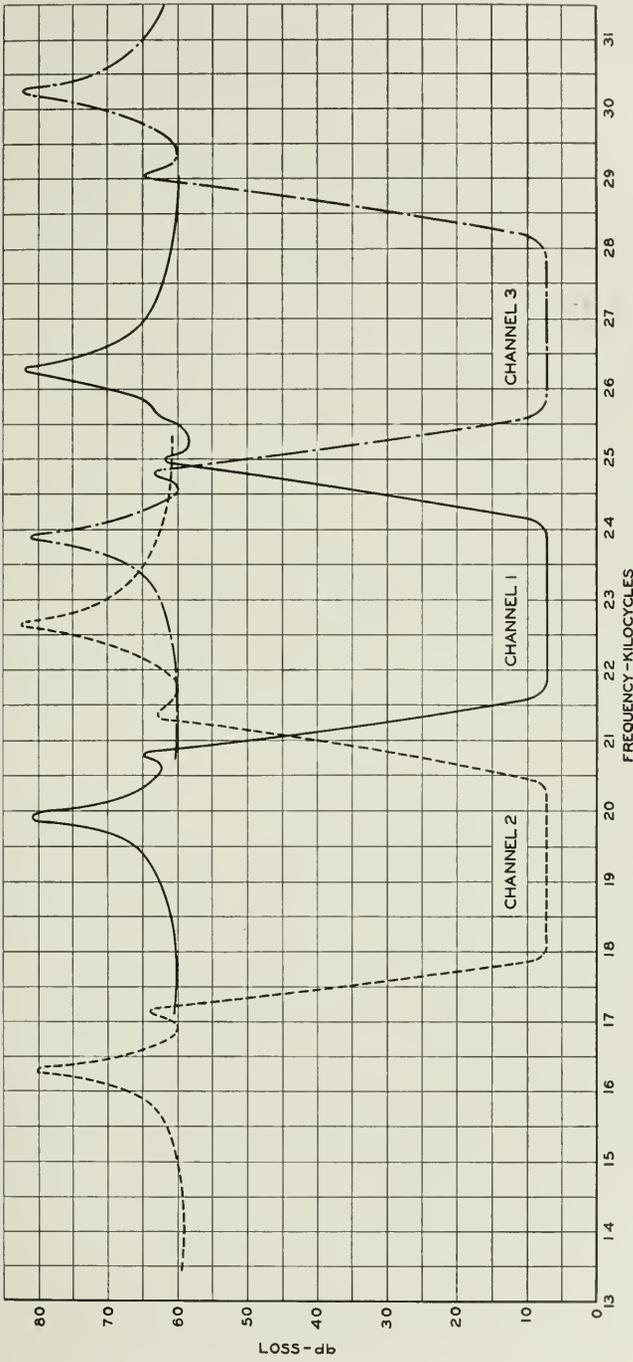


Fig. 8—Typical channel band filter characteristics.

The transmitting amplifier is the same as that used on the receiving side and in both directions of transmission in the repeater. This amplifier is capable of operating at a level 18 db above the transmitting toll switchboard.

On the receiving side, following the directional filters, is the equipment which makes up the system of equalization and regulation. This is identical with that used at the repeaters and is described more fully later. The regulator functions so as to maintain a nearly constant level at the output of the receiving amplifier. The band filters differ from those on the transmitting side only in the frequency bands which are transmitted and are the same as those used at the distant transmitting terminal. The demodulator circuit is of the same general type as the modulator circuit and the oscillator which supplies it with a carrier frequency is practically identical to that used by the modulator. However, because of the low levels at which these copper-oxide units are operated, an amplifier tube is necessary to restore the level to the required value at the output. The gain of this amplifier is continuously adjustable over a range of about 10 db so that precise adjustments of the overall circuit net loss can be made on each channel individually.

On very short non-repeated systems the transmission variations may not be great enough to require the automatic regulating equipment. In such cases a manually operated potentiometer will be used for controlling the gain.

REPEATER

A block diagram of the repeater is shown in Fig. 9. Directional filters on each side separate the two directions of transmission. As in the case of the receiving terminal, the equalizing and regulating equipment maintains all channels at the proper level at the amplifier output. The high cut-off filter shown on the circuit in the west-to-east direction limits the transmission to frequencies below 30 kc. This may be desirable when a system employing still higher frequencies is used on the same pair as the Type C system or on other pairs on the same line.

The repeater provides a maximum gain of about 49 db at the highest frequency in the upper group of the new system and about 43 db at the similar point in the lower group. The exact gains at different points in the frequency range are adjusted by the regulator so as to compensate for the attenuation of the line section preceding the repeater.

A complete repeater with its regulating equipment is mounted on a single equipment bay. A photograph of this bay is shown in Fig. 10.

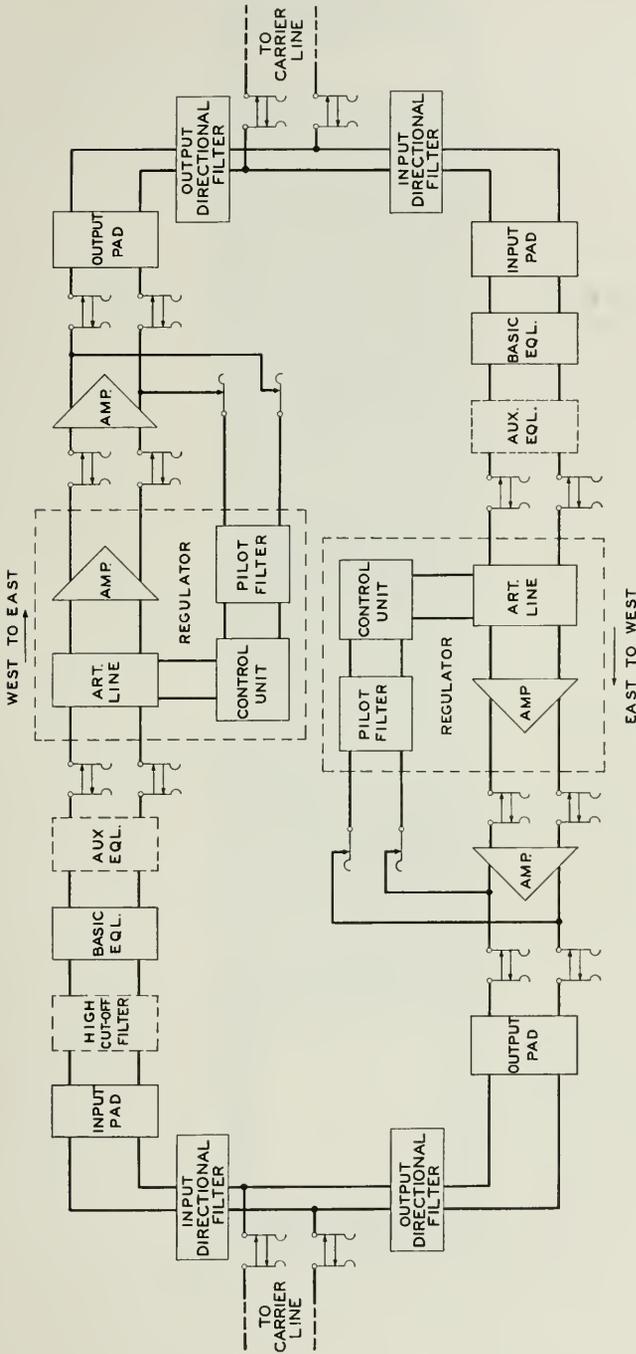


Fig. 9—Schematic of carrier repeater.

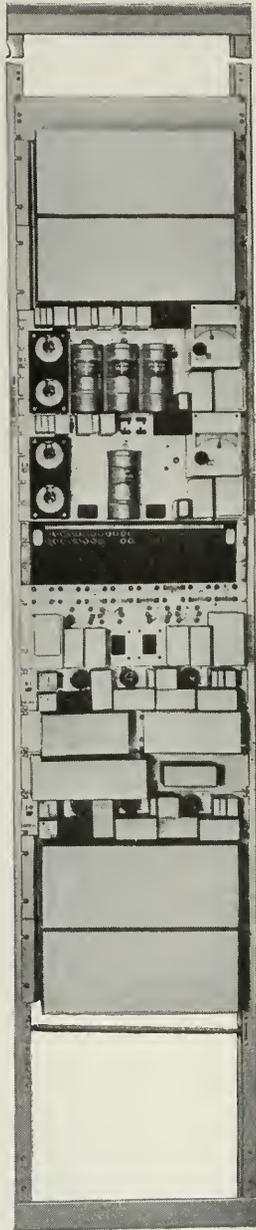


Fig. 10—New repeater for Type C systems.

REGULATION

At the transmitting terminal each channel is adjusted to the same output level and in the operation of the system it is desirable to restore this equality of levels at each repeater point and at the receiving terminal. In each direction of transmission the line losses at the upper end of the frequency range will be higher than at the lower

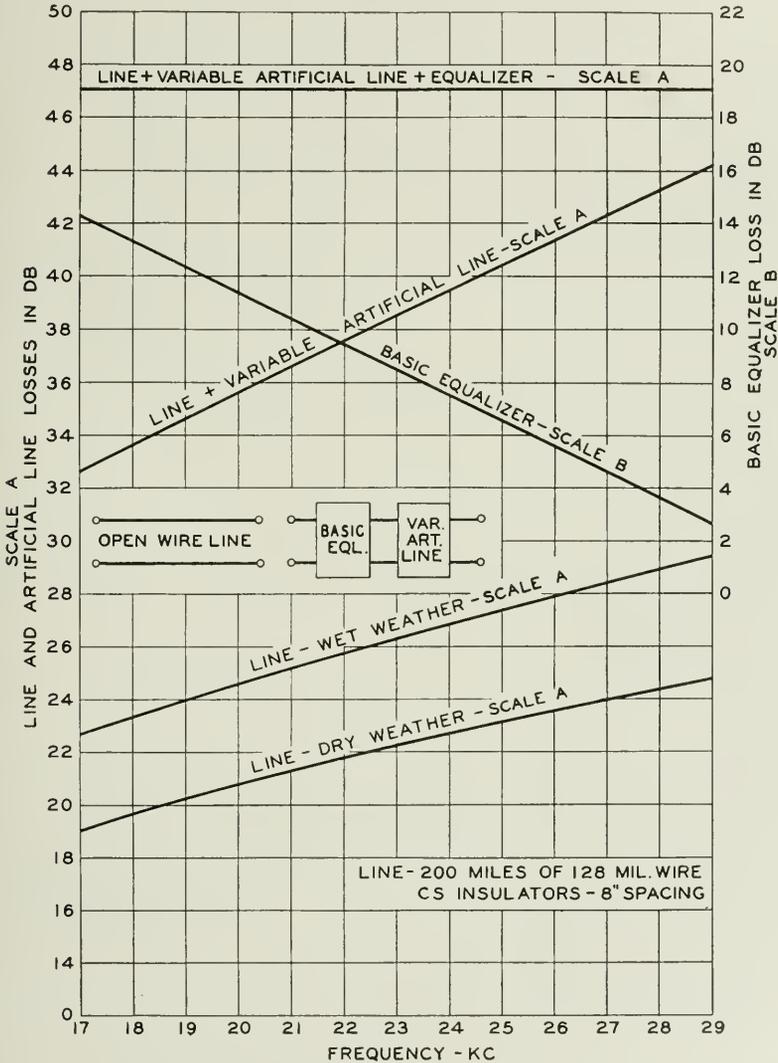


Fig. 11—Theory of regulator.

end and, furthermore, these losses change by varying amounts with temperature and weather conditions. It is the function of the regulating equipment to provide the needed adjustment and equalization of levels over the wide range of line conditions.

The theory of the operation of the regulating system is shown in a simplified manner in Fig. 11. The line circuit is connected at the end of each repeater section to a line equalizer, which is in turn connected to an artificial line unit designed to be continuously variable. The slope of the line equalizer loss characteristic is the reverse of the line slope and is as great as may be found in practice on any ordinary length of line section except under conditions of severe ice or frost. The artificial line slopes in the same direction as the line circuit itself. In lining up a system the artificial line is adjusted so that when added to the real line the slope of the combination neutralizes that of the equalizer leaving the overall transmission very nearly uniform for all frequencies in the range. Then as the loss and slope of the real line change, the artificial line is made to change in the reverse direction leaving the overall transmission still uniform. The action of this artificial line is under the control of the pilot current, referred to before, and the design of the equipment is such that in maintaining the pilot at the proper level the other channels are also properly adjusted.

In practice, the regulating unit must take care of a wide variety of wire sizes, wire spacings and insulator types. It has been found, however, that the change in slope for a given change in attenuation at some reference frequency is very nearly the same for all combinations of the above during ordinary weather conditions. As a result a single unit can be made to serve all cases.

Where conservative repeater spacings are employed there is a large amount of regulating range available to take care of sleet or frost formations on the wires. For the particular use illustrated in Fig. 11 the total range is about twice that required for ordinary wet weather. For shorter sections the available range would be still greater.

Some details of the regulating equipment are shown on Fig. 12. The pilot frequency is selected from the other frequencies on the line by a narrow band filter bridged across the output of the amplifier. This filter has a high impedance so that the bridging loss is small and does not interfere with regulation at succeeding stations.

A copper-oxide varistor is used to convert the selected pilot frequency into direct current which in turn actuates two Weston Sensitrol relays connected in series. The relays are equipped with meter scales on which a needle attached to the armature serves as a pointer. One of these relays controls the action of the regulator, the other functions

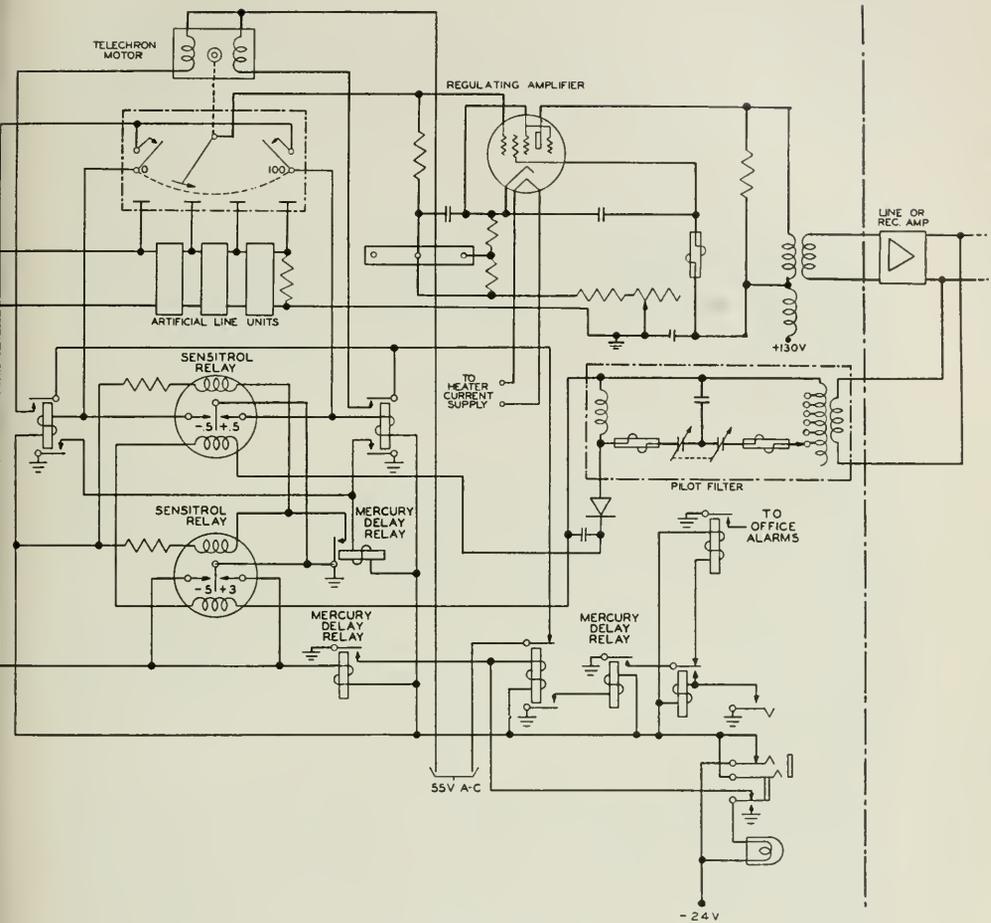


Fig. 12—Regulator circuit.

in an alarm circuit described later. With normal current flowing, corresponding to the proper pilot level at the output of the amplifier, the pointers of these relays are on midscale with a reading of 0 db. When the pilot level changes by more than .5 db in either direction, contacts close on the relay controlling the regulator and cause power to be supplied to a Telechron motor which controls the continuously variable artificial line mentioned above. This will be driven intermittently at a rate of about 1 db per minute until the output level has been restored to within $\pm .5$ db.

As will be seen from the sketch the variable line consists of three sections connected in tandem. The ends of these sections are connected to the four stators of a variable air condenser. The rotor of

the condenser meshes with any stator or with parts of two adjacent stators. The condenser, therefore, serves as a potential divider across the regulating network sections, the loss introduced depending upon the position of the rotor with respect to the stators. Basically these artificial line sections consist of units having the same loss characteristic. The first section, however, may be supplemented by additional units which will be required on the shorter line sections in order to build out the line slope as shown on Fig. 11. Since this section will be the last one to be cut out by the regulator, the less accurate part of the regulating range is thereby reserved for the periods of very high line loss, such as during ice or frost formation which occurs only infrequently.

The second sensitrol relay has contacts which close only on much larger changes in the pilot level such as would result from a failure of the line itself. When it operates it disables the regulating circuit and through a slow operating mercury relay brings in an audible or visual alarm indicating to the attendant that the system is in trouble. The slow operating relay introduces a delay in the operation so that short interruptions will not operate the alarm system.

The principal function of the regulating amplifier shown on Fig. 12 is to provide a high-impedance termination for the regulating network and condenser combination. There is, however, a small amount of gain available which may be useful in some cases.

NEW LINE ANPLIFIER

The amplifier which is used in the repeaters and in the transmitting and receiving branches of each terminal is one of the outstanding developments of the system. It was designed to have satisfactory transmission characteristics over both upper and lower frequency groups. It employs the principle of negative feedback⁶ to achieve a high degree of stability, freedom from modulation and stabilized input and output impedances.

The advances made in the design of this amplifier can be seen by the comparison in the following table with the push-pull amplifier which was used in the older systems. In some cases the latter was supplemented by an auxiliary amplifier where higher gains were needed.

	24-Volt Power— Watts	130-Volt Power— Watts	Panel Space— Inches	Gain db	No. of Tubes
Push-Pull Amplifier.....	52.8	15.1	12 $\frac{1}{4}$	32	6
Push-Pull Amplifier Plus Auxiliary Amplifier.....	76.1	17.2	17 $\frac{1}{2}$	48	8
New Feedback Amplifier.....	16.4	6.3	3 $\frac{1}{2}$	50	2

There should be added to the comparison the fact that the new amplifier does not require selection of tubes to obtain satisfactory modulation results. It is also more stable with variations of power voltages and changes of vacuum tubes. The large space saving will be evident from Fig. 13, which pictures the new amplifier with the old push-pull amplifier and its auxiliary.

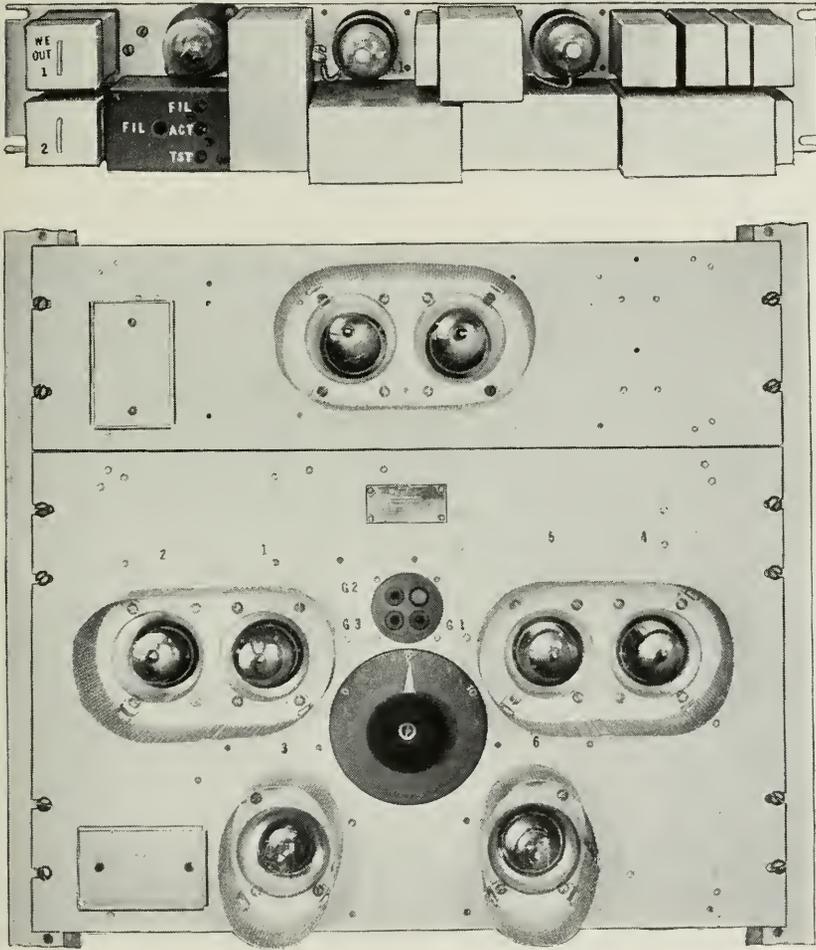


Fig. 13—New amplifier compared with the older type which it replaces.

The circuit of the amplifier is shown in Fig. 14. It is a two-stage circuit using pentode tubes. As will be seen from the circuit the feedback connection is obtained through the use of input and output transformers which are essentially hybrid coils. These coils are

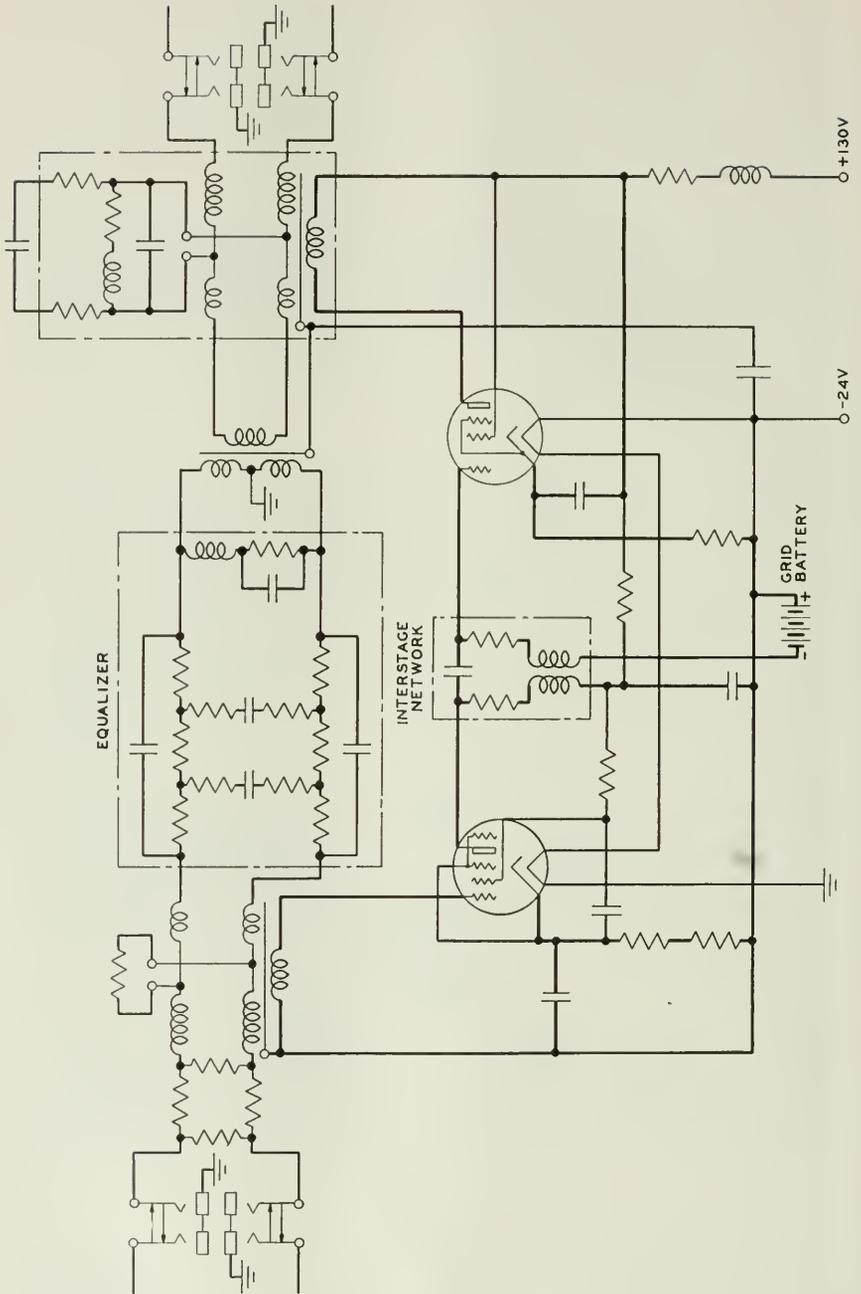


Fig. 14—Schematic of amplifier used in terminal and repeater.

designed to have unequal ratios, minimizing the loss to through transmission at the expense of greater loss in the feedback circuit. Including the transformers in the feedback path makes them also beneficiaries of the feedback with a resultant reduction in impedance irregularities, transmission distortion and modulation products.

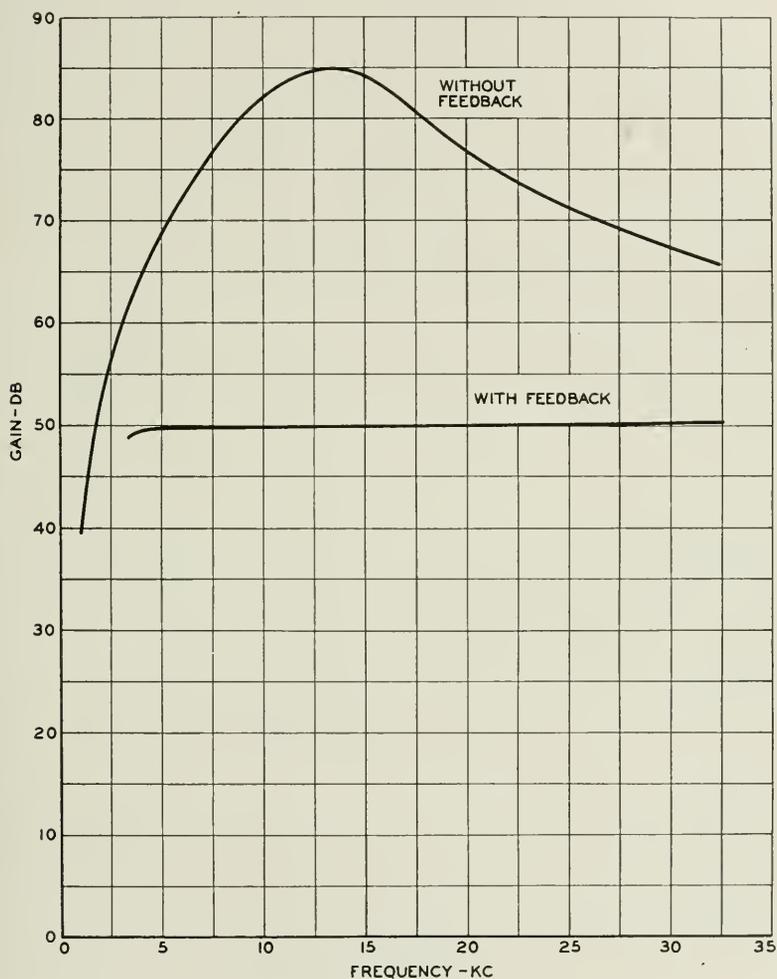


Fig. 15—Gain frequency characteristic of amplifier.

The amplifier has a gain of 50 db with provision for increasing it to 52 db when used as a terminal transmitting amplifier. Gain-frequency characteristics are shown in Fig. 15 for the 50-db gain condition with and without the feedback connection. The effect of feedback on transmission distortion is evident in this figure.

The second and third harmonic products in the amplifier are illustrated in Fig. 16 which shows their relation to the fundamental for different values of fundamental output. This harmonic production is a good measure of the modulation performance. In this respect

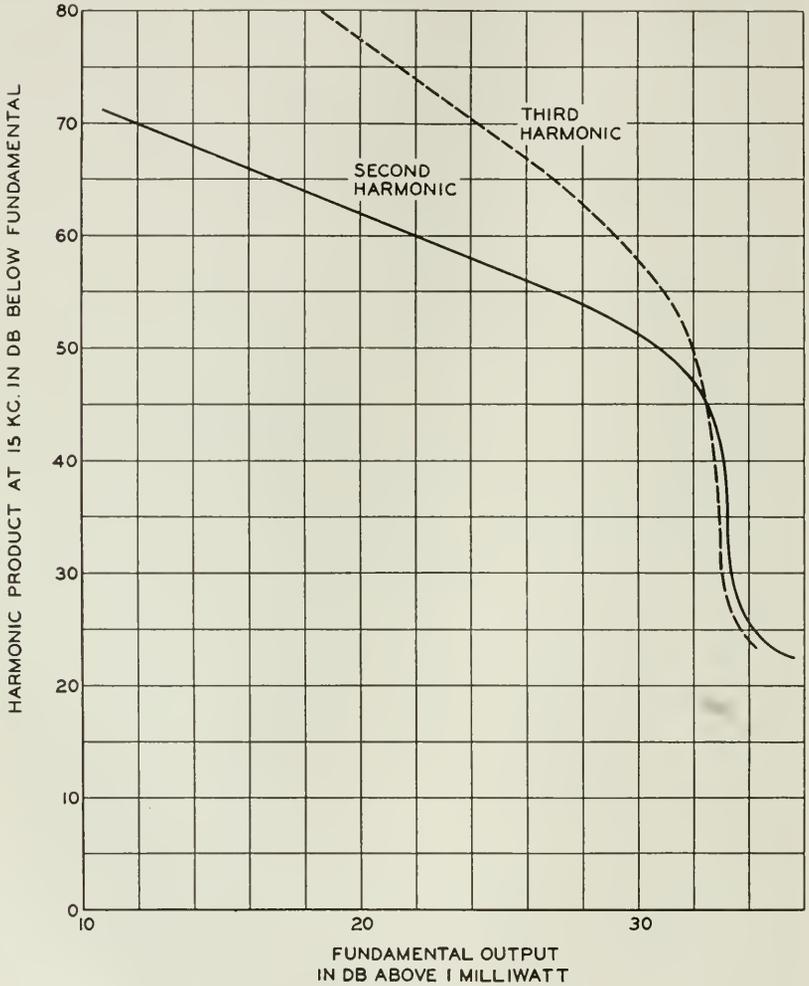


Fig. 16—Modulation in line amplifier.

the new amplifier is as good as or better than the push-pull amplifier which was used in the older system when a periodic selection of tubes was necessary to insure a satisfactory reduction of second harmonics.

A gain-load characteristic of the amplifier, with respect to a single-

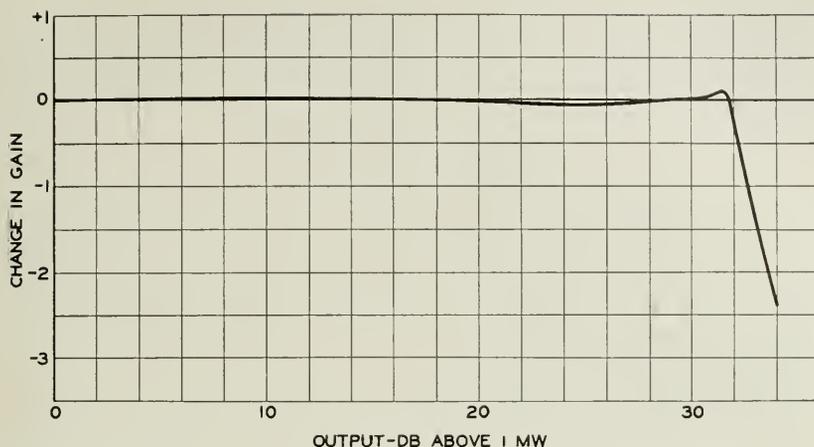


Fig. 17—Gain-load characteristic of amplifier at frequency of 15 kc.

frequency output, is shown in Fig. 17. Translated into terms of three-channel operation this means a permissible level of 18 db above the transmitting toll switchboard without noticeable distortion or interference between channels due to modulation.

POWER SUPPLY

The new system is designed to operate at both terminals and repeaters on standard telephone office battery supply of 24 and 130 volts. A dry-cell battery is required to supply grid bias to the output tube of each amplifier. A small amount of 110-volt a-c power is also required to drive the Telechron motor in the regulator circuit. The total steady power consumption in a terminal is 88 watts and in a repeater is 56 watts. These figures compare with 220 watts and 164 watts, respectively, in the older terminal and repeater.

Where regulated 24-volt battery is not available, tubes having a slightly greater current consumption are used in combination with ballast lamps. In this case the power used will be somewhat greater.

Provision has been made for a separate a-c power conversion unit to be used with the system in offices where the usual d-c voltage is not available. This should prove of great value where the provision of a battery reserve is not warranted.

SIGNALING

Standard voice-frequency signaling equipment can be used with the new terminals. There is also available a new type of ringer circuit in which a single tube functions both as the source of power for an outgoing signal and as a detector for an incoming signal. This

circuit employs heater type tubes and will operate from the a-c power conversion unit mentioned above. Since it also obviates the need for a 1000-cycle generator as a source of signaling current it is particularly well adapted to the small office type of installation.

EQUIPMENT FEATURES

As mentioned before the new terminal is much more compact than those previously used. Formerly $2\frac{1}{2}$ bays of standard size were required for the terminal proper and an additional bay for the automatic regulating equipment. Now a complete terminal including the regulating equipment can be mounted in one such bay with some space left for miscellaneous equipment.

The same degree of compactness has been applied to the new repeater. A bay of standard size was formerly required for the repeater proper with another bay for the automatic regulating equipment when provided. Both are now provided in one bay and, as in the terminal, some space is available for mounting other equipment.

The assembly of the equipment panels of the carrier terminal and repeater generally follows conventional practices, the repeating coils, condensers, vacuum tubes, etc., being mounted on the front of steel panels with the electrical terminals projecting through and the wiring placed on the rear. The filters are in sealed cans with soldering terminals brought out on the rear for wiring connections.

In view of the wide field of use anticipated for the new system, somewhat more than the usual flexibility of assembly and arrangement of parts has been provided. In small terminal offices, that is, offices having one or two systems, the four-wire terminating sets and associated patching jacks and the carrier line equipment and associated jacks may all be in one bay, using for this purpose the miscellaneous bay space referred to above. Similarly, the line filter equipment may be mounted on the repeater bay.

In the larger terminal offices, in order to facilitate operation and maintenance, the four-wire voice-frequency patching jacks can be located in a central patching bay with similar jacks from other carrier channels. Testing and monitoring equipment provided at such a point will, therefore, be common to many circuits. In the same manner the carrier frequency patching jacks of a large number of terminals or repeaters in an office can be grouped at a central point.

LINE CONSIDERATIONS

Since the new system occupies practically the same frequency range as its predecessor, it can be applied to open-wire routes in very much

the same manner. Wire sizes of 104 mil, 128 mil and 165 mil are commonly used in the Bell System plant, the particular size often being governed by mechanical rather than by transmission considera-

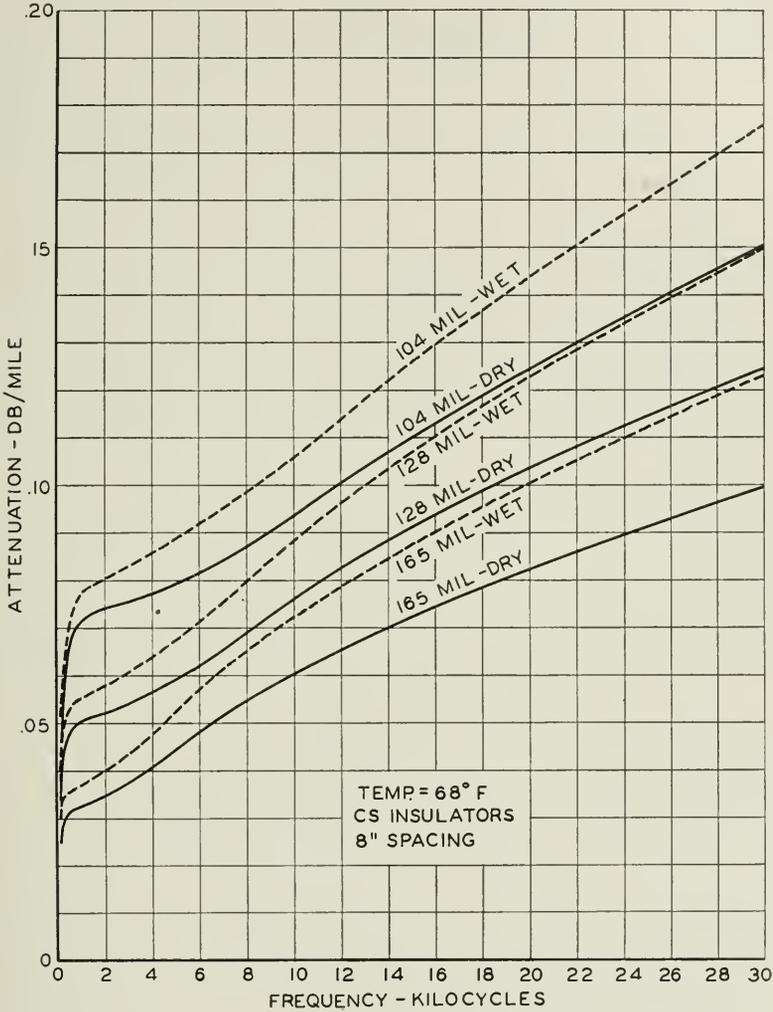


Fig. 18—Attenuation characteristics of 104, 128 and 165 mil open-wire lines.

tions. These lines are carried into the terminal and repeater offices through cables which are usually loaded to maintain smooth impedance relations and reduce the transmission losses.

The control of crosstalk⁷ is one of the major problems in the application of carrier to open-wire lines. On short lines where only a

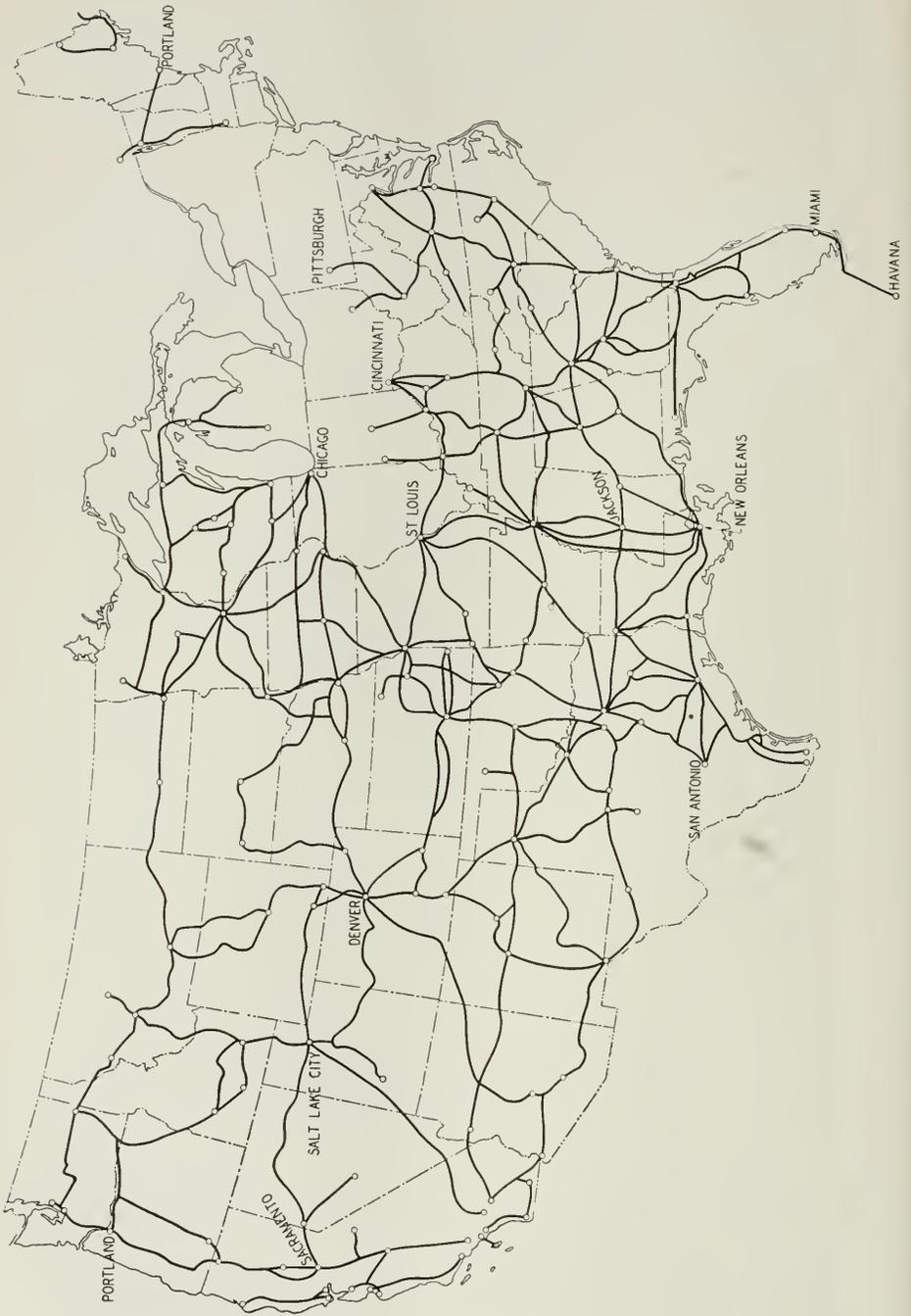


Fig. 19—Bell Telephone System routes on which Type C carrier telephone systems are operated.

few systems are involved, the engineering solution may be quite different from that on a long line on which it is desired to operate many systems. Several different plans for transposing wires have been developed which can be applied with various pole line configurations so as to meet the necessary requirements in any practical case. The more recently constructed carrier lines have, in general, employed a spacing of 8 inches between the wires of a pair, with from 16 to 26 inches between horizontally adjacent pairs. Besides contributing to the crosstalk reduction, the closer spacing is also less susceptible to interference from outside sources, which is an advantage from a noise standpoint.

Attenuation characteristics⁸ for typical eight-inch spaced pairs using the CS type of insulator are given in Fig. 18. Normal or dry weather characteristics are shown, together with the losses that are assumed for ordinary wet weather. Temperature changes also result in sizable transmission changes. It is also important to note that the losses when the wires are coated with sleet or frost may go far beyond those indicated for the wet weather condition.

CONCLUSION

The widespread use that has been made of the Type C system up to this time is pictured in Fig. 19, which shows the routes in the Bell System over which systems are now operating. The new design with its lower costs, improved performance and greater flexibility should find increased application not only on these routes but on shorter lines on which the system has hitherto not been economical.

REFERENCES

1. "Carrier Systems on Long Distance Telephone Lines," H. A. Affel, C. S. Demarest and C. W. Green, *Bell Sys. Tech. Jour.*, volume 7, July 1928, pages 564-629; *A.I.E.E. Transactions*, volume 47, October 1928, pages 1360-1386.
2. "Cable Carrier Telephone Terminals," R. W. Chesnut, L. M. Ilgenfritz and A. Kenner, *Bell Sys. Tech. Jour.*, volume 17, Jan. 1938, pages 106-124; *Electrical Engineering (A.I.E.E. Transactions)*, volume 57, May 1938, pages 237-244.
3. "Carrier Telephone System for Short Toll Circuits," H. S. Black, M. L. Almquist and L. M. Ilgenfritz, *Electrical Engineering (A.I.E.E. Transactions)*, volume 48, January 1929, pages 117-140.
4. "A New Single-Channel Carrier Telephone System," H. J. Fisher, M. L. Almquist and R. H. Mills, *Bell Sys. Tech. Jour.*, volume 17, January 1938, pages 162-183; *Electrical Engineering (A.I.E.E. Transactions)*, volume 57, January 1938, pages 25-33.
5. "Copper-Oxide Modulators," R. S. Caruthers, Presented at Winter Convention of A.I.E.E., January 1939.
6. "Stabilized Feedback Amplifiers," H. S. Black, *Bell Sys. Tech. Jour.*, volume 13, January 1934, pages 1-18; *Electrical Engineering*, volume 53, January 1934, pages 114-120.
7. "Open-Wire Crosstalk," A. G. Chapman, *Bell Sys. Tech. Jour.*, volume 13, January 1934, pages 19-58; April 1934, pages 195-238.
8. "The Transmission Characteristics of Open-Wire Telephone Lines," E. I. Green, *Bell Sys. Tech. Jour.*, volume 9, October 1930, pages 730-759; *Electrical Engineering (A.I.E.E. Transactions)*, volume 49, October 1930, pages 1524-1535.

Crossbar Dial Telephone Switching System *

By F. J. SCUDDER and J. N. REYNOLDS

This paper describes the crossbar dial telephone switching system recently adopted by the Bell System for large cities where the panel system has been used for nearly twenty years. Central offices of the crossbar type can be introduced in panel areas without changes in existing offices and without changes in existing dial telephone instruments. Crossbar offices and panel offices in the same building will operate on a common power plant and utilize other equipment in common, such as "A" and "B" operator switchboards and outgoing trunks.

The precious metal contact crossbar switches are used for all switching purposes as contrasted with the base metal contact panel switches. The switches operate with relay-like movements under control of common control or marker circuits consisting primarily of multi-contact, *U* and *Y* type relays. The control and marker circuits, which are connected to the switch frames by means of multi-contact relays, perform their operations in a fraction of a second. The switches, the *U* and *Y* type relays and the multi-contact relays are equipped with twin contacts of precious metal. Senders similar to those of the panel system are employed.

The system will be used for new offices in larger cities as manufacturing and plant conditions permit.

INTRODUCTION

IT is the purpose of this paper to describe briefly the crossbar dial telephone central office switching system which has recently been developed by the Bell System for use in large cities. Sixteen years ago, in February 1923, a paper was read before the Institute, by Messrs. E. B. Craft, L. F. Morehouse and H. P. Charlesworth of the Bell System which outlined the history and the problems involved in telephone central office switching and described the panel dial central office system which had just been developed and was being introduced in the large cities. The first central office of this type was placed in service in December 1921, and since that time 456 panel dial offices serving nearly four and one-half million subscriber stations have been installed in 26 different cities throughout the country. During these years many improvements have been made in the panel system to make it more serviceable to the telephone public and to meet the new

* Presented at Winter Convention of A.I.E.E., New York, N. Y., January 23-27, 1939.

problems which have arisen, but in addition the engineers of the Bell System have continued their search to find new and better means for meeting telephone switching demands. This work has resulted in the adoption of the crossbar type central office switching equipment. Two offices of this type were placed in successful operation during 1938 and others are in process of manufacture and installation.

It will be appreciated that for large metropolitan areas, the development and economic introduction of a central office switching system which differs materially from the existing systems is a rather large undertaking. The system must fit into the existing plant as a whole without material change. Generally any important changes affecting the subscribers' use of the telephone or the methods used by switchboard operators should be avoided. Existing numbering plans should not be affected, existing classes of service should be continued, and the addition of others made feasible in case they should be required.

All of these and many other factors have been taken into account and all requirements have been met by the crossbar system which offers important improvements in telephone switching, both in operation and maintenance. Its introduction does not make any of the existing equipments obsolete in the sense that these equipments will be less serviceable nor will it cause their replacement. Central offices of the crossbar type can be installed in the same building with existing panel central offices without loss in operating economies in either type of office. Certain equipment, such as the existing and additional outgoing trunks to other central offices, manually operated switchboard positions, operating room and maintenance desks, power plant and alarms, can be used in common by the two types of offices in the same building.

GENERAL

Before describing the crossbar system it is desirable first to give a brief outline of the principal functions of a dial central office equipment. Such an office is capable of serving 10,000 subscriber line numbers, and is provided with a sufficient number of connecting switches, trunks and associated circuits so that under usual peak loads of traffic, calls will be completed promptly.

The central office circuits, in response to the lifting of the receiver by the calling subscriber, connect the subscriber line to the switching equipment. This equipment then extends the calling line, "link by link," through several switching stages to the called line as determined by the called office code and line number dialed by the calling subscriber. When the connection has been established to the called line, the subscriber bell is rung and, when the subscriber answers, the

talking connection is completed. During the conversational period the connection is held under control of the calling telephone, and when the telephone receivers are replaced, the central office equipment and the telephones are released for use on other calls. The equipment, of course, transmits the busy tone signal to the calling subscriber if the called line is found busy, and automatically routes a call for a discontinued or an unassigned line to an operator who informs the subscriber of the status of such a line.

Operators and associated switchboards are provided in the dial system to handle certain classes of calls and to render assistance to subscribers when required. Calls to these operators are established in response to the dialing of operators' codes in a manner similar to the establishment of calls to other subscribers.

Operators are usually provided to complete calls terminating in a dial office which are originated by subscribers served by manual offices.

Prior to the introduction of the crossbar system, the Bell System employed two general types of dial central offices. These are the well known step-by-step and panel systems.

The step-by-step system has been used generally in the smaller cities which are frequently served by a single central office or by a relatively small number of offices and where the trunking problems are consequently less complicated. The switches of the step-by-step system are controlled directly by the impulses from the subscriber dials and, necessarily in conformity with the dial, the system operates on a decimal basis. The selectors of this system are first moved under control of the dial to any one of ten vertical positions, corresponding to the numeral of the digit dialed, and in the case of trunk hunting switches is then automatically rotated over a row of ten trunk terminals to find an idle trunk during the interdigital time of the dial.

The step-by-step switch thus has access to ten different groups of trunk terminals with ten terminals each. The location of the trunk groups on the switches is governed by the digits dialed and consequently the relocation of a group necessitates directory changes. These limitations in trunk access and flexibility are not material handicaps in the smaller cities throughout the country where the system is giving excellent service.

The panel system meets the complex service requirements of the larger cities with their large volume of traffic and multiplicity of central offices. In these cities the number of trunk groups is large and the number of trunks in the groups varies widely. Further, the number of groups and their sizes are frequently changed by the introduction of

new offices and changes in the character or extent of existing central office areas.

In the panel system, senders are provided which record and store the dial pulses as they are dialed and then independently control the operation of the switching units. The large panel type switches provide access to large groups of trunks and to a large number of groups, and at the same time permit considerable variation in the sizes of the groups. The necessary flexibility in the size and location of the trunk groups is obtained by flexibly wired routing equipment provided in decoder circuits which are associated with the senders. These facilities permit trunk group locations on the switches as dictated by traffic regardless of the office codes listed in the directories and dialed by the subscribers. The panel system also readily provides for the routing of calls through intermediate or tandem offices where the traffic between offices can be more economically handled in this manner.

The crossbar system also makes use of the sender and decoder method of operation and provides a still greater flexibility in the trunking arrangements than is obtained by the panel system.

THE CROSSBAR SYSTEM

The two outstanding features of the crossbar system are the "crossbar switch" which is used for all major switching operations, and the "marker" system of control which is used in the establishment of all connections throughout the crossbar office.

The crossbar system is essentially a relay system employing simple forms of relays and relay type structures for all switching operations. The apparatus consists almost wholly of crossbar switches, multi-contact relays and the usual small relays similar to those generally employed in all telephone systems. The switching circuits are wired to the contacting springs of the switches, and the connections through the switches are made by pressing contacts together by means of simple electromagnetic structures instead of the moving brushes and associated fixed bank terminals of other systems.

The use of relay type apparatus with its small, pressure type contact surfaces economically permits the use of twin or double contacts with thin layers of precious metal for all contact points. Obviously, double precious metal contacts make for reliable operation, especially with the low speech and signaling currents inherent to a telephone system.

The short mechanical movements and the inherently small operating time intervals of the "relay-like" crossbar switch permit the use of

common circuits or "markers" to control the operation of the switches. This has permitted the use of large assemblies of switches and associated relays on unit frames which can be wired and completely tested for operation in the factory before the units are shipped.

In the design of the switching frames and associated control circuits, one of the objectives realized has been the standardization of a relatively small number of different types of equipment units, thereby simplifying manufacture and merchandizing. This also simplifies the engineering of the equipment by the Telephone Companies in the preparation of their specifications to meet the particular traffic requirements of the various central offices.

The marker system used for controlling the switching operations has many advantages, the more important of which will be disclosed later in the general description of the operation of the equipment. It might be mentioned here, however, that the marker is an equipment unit consisting almost entirely of relays, which completes its functional operations in the establishment of a call in a fraction of a second. This short operating time permits a few markers to handle the entire traffic in the largest office. The markers are connected momentarily by means of multi-contact relays to the various switching units of the office to control the establishment of the calls through the crossbar switches.

An outstanding advantage of the marker system of control is the "second trial" feature, by means of which two or more attempts can be made to establish a call over alternate switches and trunks when the normally used paths are all busy. The markers are arranged to detect short-circuited, crossed, grounded and open-circuit conditions at all vital points, and before releasing from a connection they make circuit checks to insure that the connection has been properly established. When trouble conditions are detected, they make a second attempt to complete the connection, after sounding an alarm and recording the location and nature of the trouble encountered. The marker system facilitates the introduction of new service features and changes in operation, which may be found desirable from time to time, due to the fact that the principal controlling features of the entire system are vested in a small number of markers.

APPARATUS

Crossbar Switch

The crossbar switch from which the system derives its name is the basic switching unit of the system. Figure 1 shows the front view of a 200-point crossbar switch.

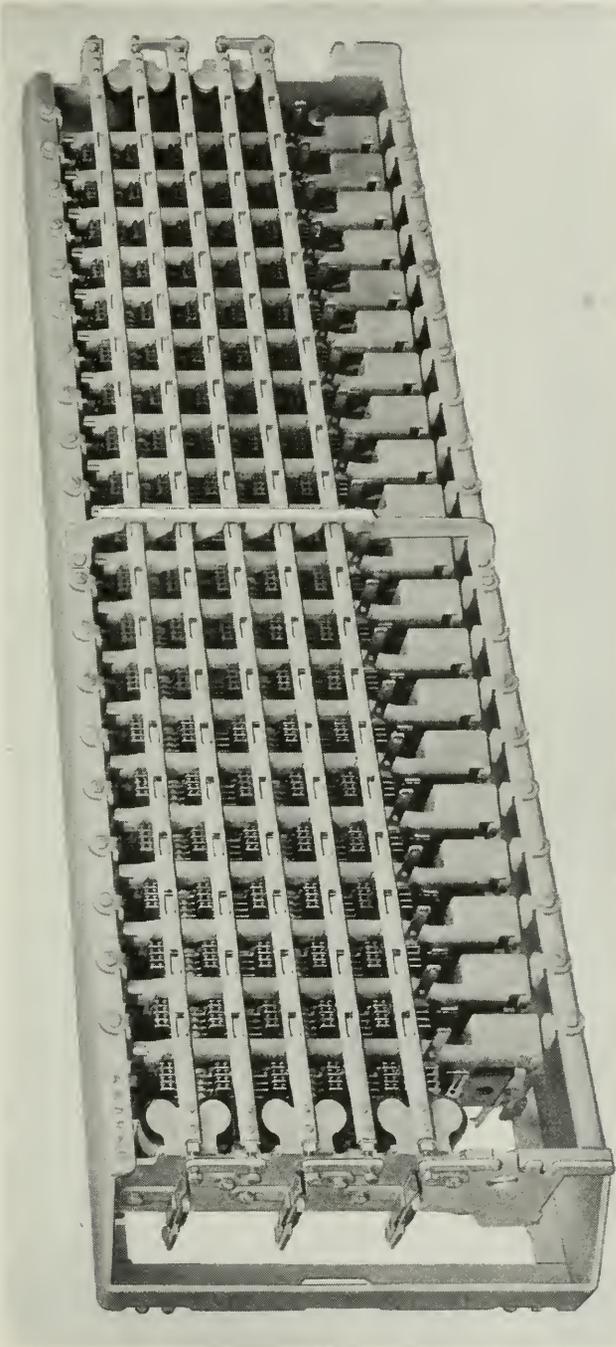


Fig. 1—Crossbar switch—front view.

Fundamentally this switch consists of three major functional parts: (a) twenty separate vertical circuit paths, (b) ten separate horizontal circuit paths, and (c) a mechanical means for connecting any one of the twenty vertical circuit paths to any one of the ten horizontal circuit paths by the operation of electromagnets. From a structural viewpoint the switch is comprised of a rectangular welded frame on which are mounted twenty vertical units and the selecting mechanism consisting of five horizontal bars operated by ten selecting magnets.

Primarily the switch is a multiple relay structure with twenty vertical relay-like units, each unit having an operating or "holding" magnet and ten sets of contacts in a vertical row. The switch arrangement provides a rectangular field of contacts in twenty vertical rows and ten horizontal rows or a total of 200 sets of contacts, one set at each "crosspoint." These crosspoint contacts are operated independently of each other by a coordinate operation of the horizontal and vertical bars. The horizontal bars are controlled by the ten horizontal or "selecting" magnets and the vertical bars by twenty vertical or "holding" magnets. Any set of contacts in any vertical row may be operated by first operating the selecting magnet corresponding to the horizontal row in which the set of contacts is located, and then by operating the holding magnet associated with the vertical row. Since the contacts are held operated by the holding magnet alone, the selecting magnet is operated but momentarily and is released as soon as the holding magnet is operated. After the selecting magnet is released, other connections may be established through the switch by the operation of other selecting and holding magnets. It is thus apparent that ten connections can be established through the switch, one for each of the horizontal paths.

From Fig. 2 the rather simple mechanical interlocking of the horizontal and vertical bars which causes the operation of a set of crosspoint contacts will be understood. The ten sets of contacts in a vertical row are associated with the vertical or "holding" bar of the row. Each horizontal or "selecting" bar is provided with twenty selecting fingers which are made of flexible wire. These fingers are mounted at right angles to the bar, one at each of the vertical rows of contacts. Thus when a selecting bar is rotated through a small arc by its magnet, the selecting fingers will move up or down into a position so that when a holding bar is operated by its magnet, it will engage the selecting finger at the crosspoint of the two bars and cause the corresponding set of contacts to operate. The selecting bar and the fingers not used will then be released when the selecting magnet is released, but the selecting finger used to operate the selected set of

crosspoint contacts will remain latched and the contacts held closed by the holding bar until the holding magnet is released at the end of the connection. The selecting fingers are each provided with a damping spring to reduce vibration on the operation and release of the fingers.

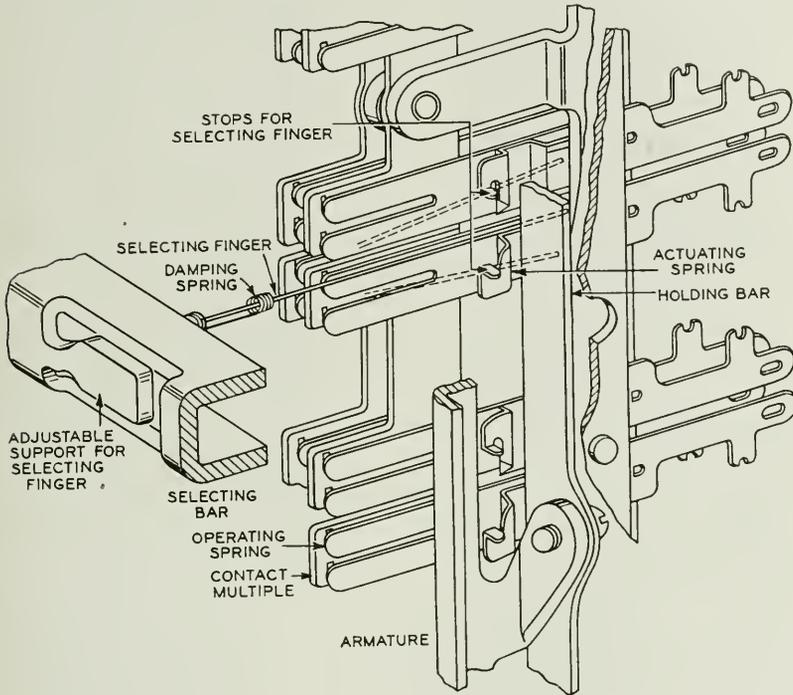


Fig. 2—Crossbar switch selecting mechanism.

It will be noted that the selection operation is performed by five horizontal bars although there are ten horizontal rows of contacts. This is accomplished by operating the bars in either of two directions. As shown in Fig. 1, two magnets are associated with each bar, one whose armature is on top of the bar, the operation of which causes the selecting fingers to move in a downward direction, and the other whose armature is below the bar causing the fingers to move upward. The selecting bars are restored to the normal or mid-position by the centering springs located on the end of the switch adjacent to the magnets.

Figure 3 shows the vertical unit of the crossbar switch with its ten sets of normally open "make" type contact springs, the holding magnet at the bottom, and the long vertical armature to which is attached

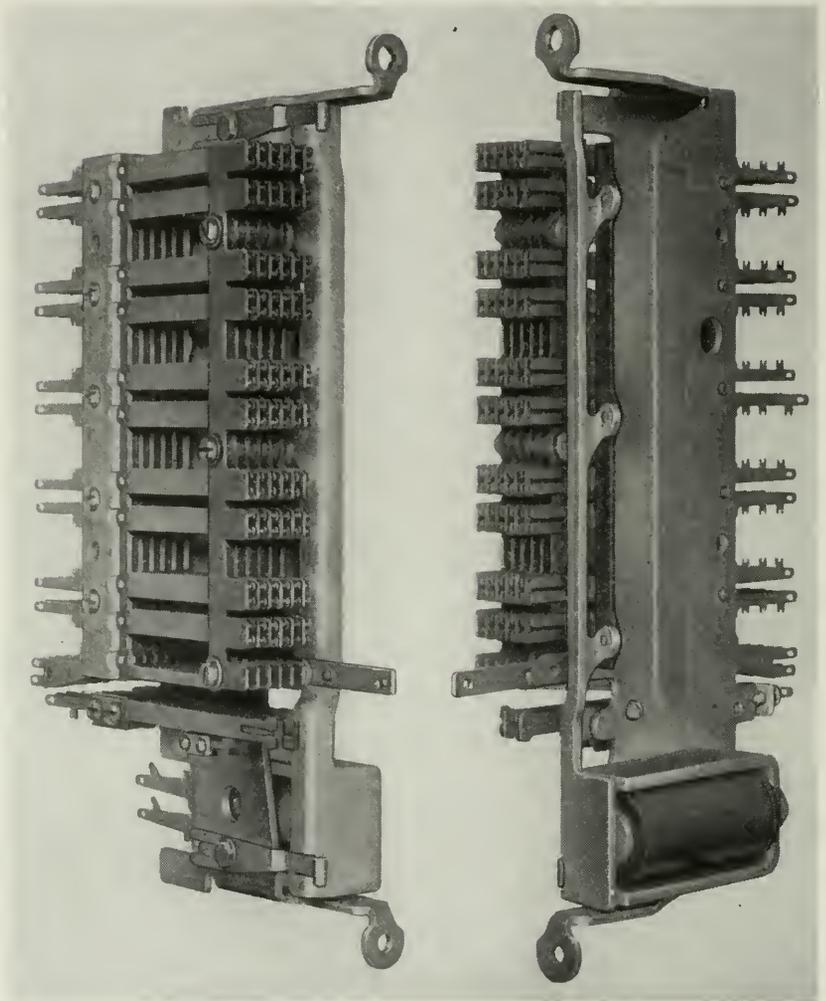


Fig. 3—Crossbar switch vertical unit.

the vertical holding bar. The vertical unit shown has six pairs of contacts at each of its ten crosspoints. Other vertical units are provided with ten sets of 3, 4 and 5 pairs of contacts per set. One spring of each pair as shown is a fixed spring consisting of a projection of an insulated vertical metal strip, made in the shape of a comb. This strip extends from the top to the bottom set of contacts of a vertical row. Wiring lugs are provided at the lower end of these vertical strips facing the rear to which are wired the lines or trunks

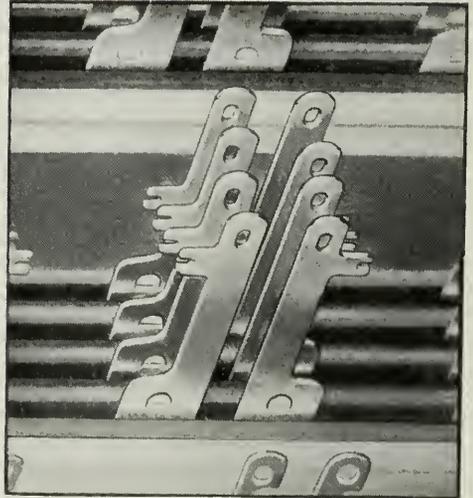
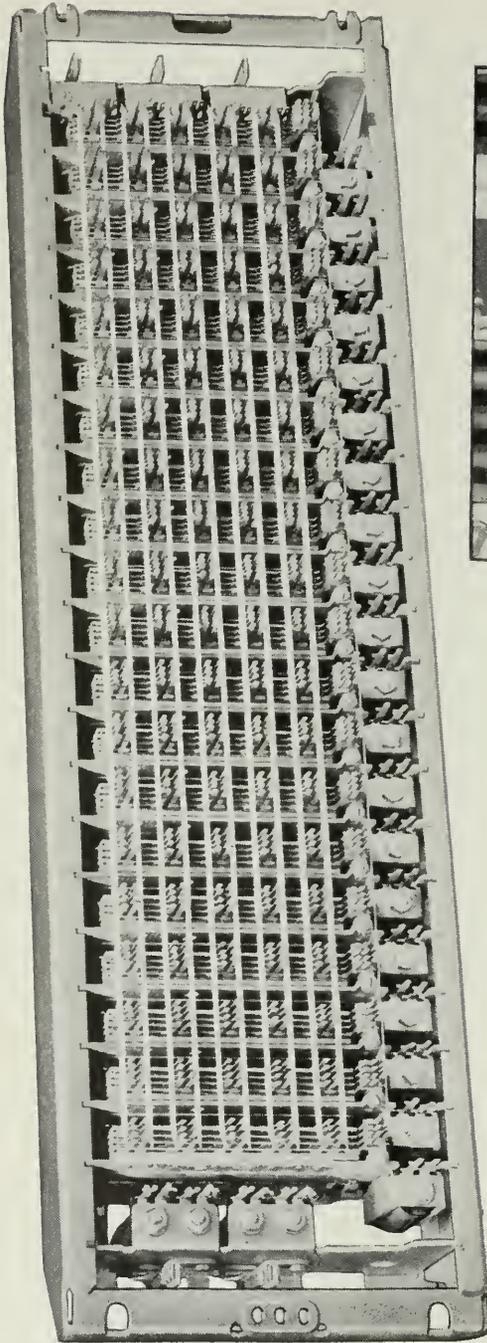
of the vertical circuit paths. At the lower end of these strips and facing the front is another projection used by the maintenance force for testing purposes. The mate or movable spring of each pair is individually insulated from all other springs. These springs extend to the rear of the switch for wiring purposes and may be strapped horizontally to the corresponding springs of adjoining vertical units to extend the horizontal circuit path through the switch.

The contacting ends of the thin movable contacting springs are bifurcated to provide two flexible contacts in parallel. The contacting surfaces on these springs as well as the mating fixed springs are provided with a thin layer of palladium. The use of the double precious metal contacts is an important feature of the crossbar system in providing more reliable contacting surfaces. Experience has shown that the chance of simultaneous failures of both contacts of a pair is extremely small. The actual contacting surfaces of each pair of springs consist of small bars of contact metal located at right angles to each other. These bars are composed of a ribbon of nickel capped with a thin layer of palladium. This crossbar arrangement of contacts provides a rather large area over which the two springs can make contact with each other, and thereby permits considerable tolerance in the manufacture and adjustment of the contact spring assemblies.

The switch may be equipped with "off normal" contact spring assemblies. When these are furnished they are associated with each selecting or holding magnet and are operated like relay contacts when the associated magnet operates, regardless of which crosspoint contact is closed. They are used to perform circuit functions as required in the various uses of the switch.

In the design of the switch special attention was given to the problem of wiring and cabling. Figure 4 shows the wiring terminals on the rear of the switch. These terminals are arranged for individual wiring and also have staggered, notched projections so that the terminals can be readily strapped together horizontally with bare wire as shown. This is an important feature of the switch since it permits a multiple of terminals to be easily soldered together and reduces the wire congestion on the switch.

The 200-point crossbar switch is $9\frac{1}{4}$ inches in height and $30\frac{1}{2}$ inches in length. In addition a 100-point switch $20\frac{1}{2}$ inches in length is provided. This switch is similar to the 200-point switch but is equipped with 10 vertical units.



Multi-Contact Relay

The multi-contact relay used in the crossbar system is shown in Fig. 5. It resembles in design the vertical unit of a crossbar switch.

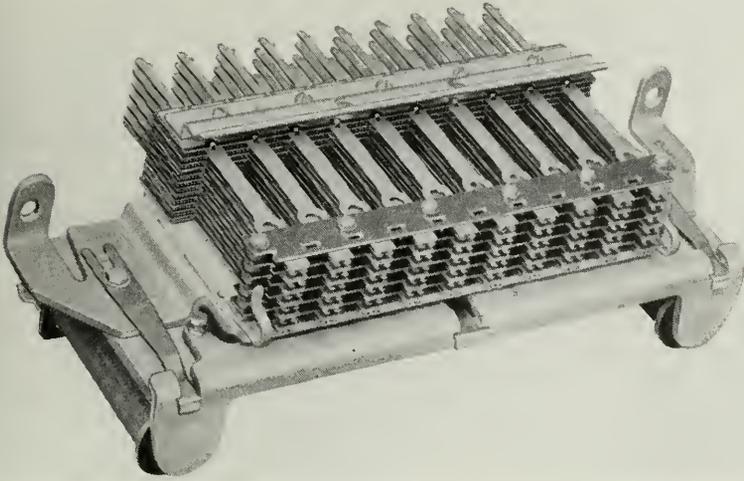


Fig. 5—Multi-contact relay.

The relay is provided in four sizes in respect to the number of contacts, namely, in 30, 40, 50 and 60 sets of individually insulated contacts, all of which are of the normally open type which are closed when the magnets of the relay are operated. Each relay is provided with two separate magnets, armatures and associated groups of springs, and both magnets are energized in parallel in order to close all of the contacts. By operating the two magnets independently the structures can be used as two separate relays, each equipped with 15, 20, 25 or 30 sets of contacts. The relay occupies a mounting space approximately 2" x 11" and is provided with a cover.

All contact springs are equipped with twin contacting surfaces similar to the contacts used on the crossbar switch except that they are composed of solid bars of precious metal due to the heavy duty requirements. To facilitate wiring, these relays are manufactured with two types of wiring terminals. In one type the movable springs are of graduated lengths and are provided with notched lugs for bare wire strapping to permit the multiplying of springs horizontally to corresponding springs on other relays mounted adjacent. In the second type the strapping lugs are omitted and all springs are of the

same length and are provided with soldering eyelets for individual or non-multiple wiring.

The multi-contact relay finds its chief use in the common connector circuits where a large number of leads must be connected simultaneously to a common circuit.

U and Y Type Relays

New and improved general purpose small relays which have been coded the "U" and "Y" type are used in this system. Figure 6 shows a typical "U" type relay. Although somewhat similar to the E and R type relays which have been in a common use in the telephone systems for many years, it differs from them principally in that it has a heavier and more efficient magnetic structure which permits the use of a greater number of contact springs. These relays permit the use of spring assemblies up to a maximum of 24 springs in various combinations of springs, including transfer contacts, simple make-and-break contacts. The relays are constructed of relatively simple parts, most of which are blanked and formed in the desired shapes in the same manner as the earlier E and R type relays. The cores are made from round stock and are welded to the mounting bracket of the relay. The structures of all of these relays are similar and differ principally in their spring assemblies and windings.

In order to insure more reliable contact closures, the relays are equipped with twin contacts. Various types of contact metal and sizes of contacts are provided, depending upon the characteristics of the circuit controlled by the contacts.

Improved methods of clamping the springs in their assemblies, together with the design of the springs, provide stability and minimize manufacturing and maintenance adjusting effort.

Contacts practically free from chatter on both the operation and release of the relay have been obtained by the use of relatively heavy stationary springs, short thin movable springs, and a pivoted arrangement of the armature suspension. By reference to Fig. 6 it will be seen that the rear ends of the armature are pivoted by two pins which project through holes in the hinge bracket mounted on the rear spring assembly. In the earlier flat type relays of the E and R type, the armature was suspended at the rear by means of a reed type armature hinge.

The Y type relays make use of the same manufacturing tools and processes as the U type. Copper or aluminum sleeves are provided over the cores beneath the windings to secure the slow-release characteristics required on these relays. The relay armature is embossed

so that when the relay is operated satisfactory contact is made between the metal surfaces of the magnetic circuit which insures uniform time characteristics.

In both the U and Y type relays the cylindrical cores permit the use of form wound coils which are wound on special machines and slipped

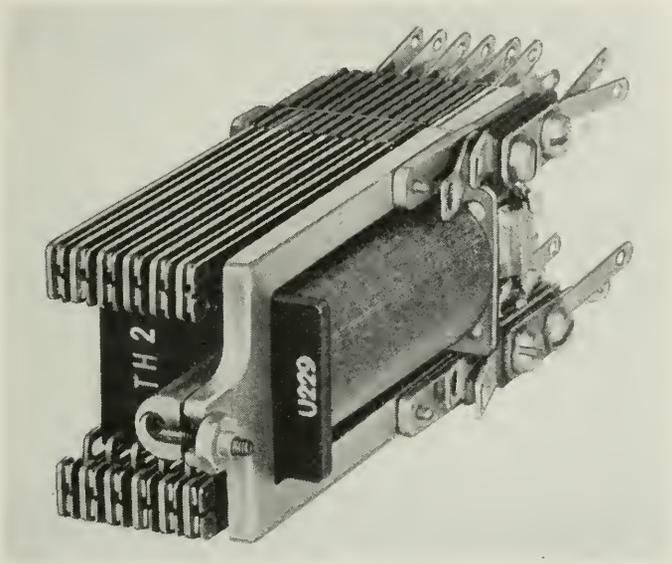


Fig. 6—"U" type relay.

over the cores when completed. In the manufacture of these coils a removable mandrel is used. It is covered with a layer of sheet cellulose acetate and accommodates several coils. These coils are then automatically wound on the mandrel from different spools of insulated wire. Separations are left between adjacent coils so that when the winding operation has been completed the individual coils can be separated. A very thin sheet of cellulose acetate is automatically interleaved between successive layers of wire to hold the wire in place and to provide insulation between layers. This general method of winding coils also is used for the magnets of the crossbar switches and multi-contact relays.

FUNCTIONS OF THE EQUIPMENT UNITS

The general operation of the system as a whole may be more easily understood by first describing the principal equipment units in the system and their functions before proceeding with a description of the

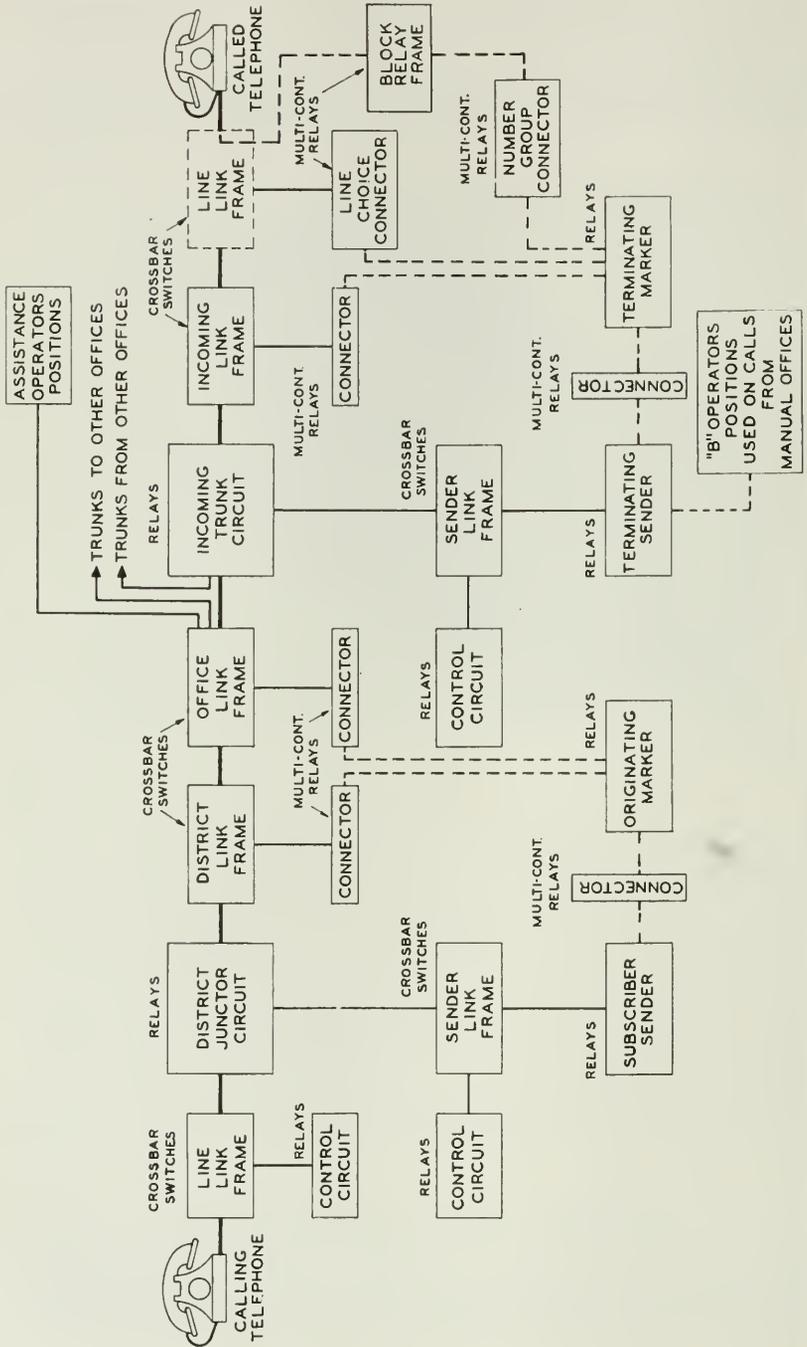


Fig. 7—Functional arrangement of equipment units.

operation of the circuits. A simplified block diagram of the principal equipment units of the system is shown in Fig. 7. It will be noted that in general there are three types of equipment units:

1. The transmission battery supply and supervisory circuits consisting of the "district junctors" and the "incoming trunks."
2. The crossbar switch frames.
3. The common "control" circuits, the "senders" and the "markers."

The "district junctor" and the "incoming trunk" circuits are composed principally of small relays. The district junctors furnish the talking battery for the calling subscribers and supervise the originating end of connections. The incoming trunks control the ringing of the called subscriber bells, furnish talking battery for the called subscribers, and supervise the terminating end of connections.

The switch frames, which consist almost entirely of crossbar switches, provide the means for switching between the subscriber lines, the district junctors and the incoming trunks. Switch frames also are used for switching the district junctors and the incoming trunks to the senders.

The "senders" consist principally of small relays and their functions are similar to those of the operators at a manual switchboard. The "subscriber senders" register the called numbers from the subscriber dials and transmit the necessary information to the "markers," to the "terminating senders" and to the manual operator positions in manual offices for completing connections to the called lines. The subscriber senders also control the operation of the selectors in distant panel offices. The "terminating senders" in the terminating end of the crossbar office receive the numerical digits of the called numbers from the subscriber senders of any dial office and transmit the required information to the "terminating markers" for setting up the connections to the called lines.

The "markers" are the most important control circuits in the system. They are composed of both small and multi-contact relays. There are two types, one for originating traffic and one for terminating traffic. The operating time of the markers is short, considerably less than one second, and consequently only three or four markers of each type are required in the average office.

The "originating markers" determine the proper trunk routes to the called office. They have access to all outgoing trunk circuits and all the crossbar switch frames that are used for establishing the connections to the called office trunks. They test the trunk group to find an idle trunk to the called office, and also test and find an idle

channel through the switch frames, and finally operate the proper selecting and holding magnets of the crossbar switches to establish the connections from the subscriber line to the trunk circuit.

The "terminating markers" perform similar functions in the terminating end of the office to set up the connection from the incoming trunk circuit to the called subscriber line. They have access to all of the subscriber lines terminating in the office, and to all crossbar switch frames used for connecting to subscriber lines. They test the called line to determine whether it is idle, and also test for and find an idle channel through the switch frames and finally operate the proper magnets of the crossbar switches and establish the connection to the called subscriber line.

In addition there are common "control" circuits associated with the "line link" and the "sender link" frames for controlling the operation of the switches on these frames. There are also the common "connector" circuits, consisting mainly of multi-contact relays, which are used for connecting the markers to the senders, to the switch frames and to the test terminals of the called subscriber lines.

It should be noted that the line link frames, although shown separately, are used for both originating and terminating traffic.

After the talking connection has been established between two subscribers, all of the common control units, such as the senders, markers, connectors, line link control circuit, and the sender link frames and their associated control circuits, will have been released, and the talking connection will be maintained in this condition by the holding magnets of the crossbar switches used on the line link, district, office and incoming link switch frames. These switch magnets are held operated under control of the supervisory relays in the district junctor and the incoming trunk circuits and are released when the subscribers replace the receivers.

TRUNKING ARRANGEMENTS

The fundamental method of using the crossbar switch for setting up connections is illustrated in Fig. 8. This figure shows a 200-point crossbar switch with twenty vertical units each wired to a subscriber line and ten trunks strapped horizontally across the switch. With such an arrangement, any one of the twenty lines may be connected to any one of the ten trunks. The number of lines which can be connected to the same ten trunks may be increased to forty by adding a second 200-point crossbar switch with twenty different lines connected to its verticals and by wiring the horizontal contact multiple of this second switch to the horizontal multiple of the switch shown

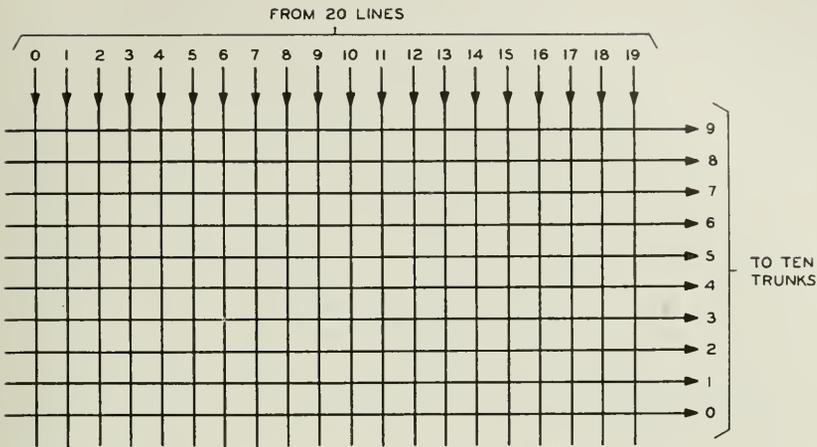


Fig. 8—Simple trunking arrangement with a single 200-point crossbar switch.

in Fig. 8. By adding other switches in this manner, any number of lines may be given access to the ten horizontal trunks.

To obtain greater trunking access, two groups of switches known as “primary” and “secondary” are used. Figure 9 illustrates this primary and secondary switch arrangement as used in the “line link” switch frames and in various forms throughout the crossbar office.

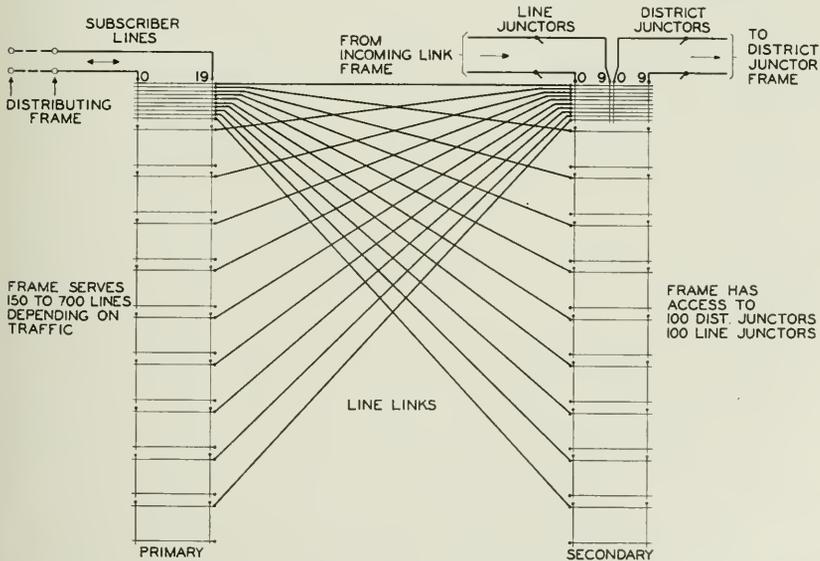


Fig. 9—Primary-secondary trunking arrangement.

The switches are arranged in two vertical files of ten primary switches and ten secondary switches. There are twenty subscriber lines connected to the verticals of each of the ten primary switches and twenty trunk circuits are connected to the twenty verticals on each secondary switch. The horizontal multiples on the primary switches are connected to the horizontal terminals of the secondary switches, each primary switch having one horizontal path connected to each of the ten secondary switches. With this arrangement, the twenty lines of any primary switch have access to all 200 trunks connected to the secondary switches. Since all of the primary switches are wired in this manner, that is, with their ten horizontal paths distributed over the ten secondary switches, then all of the 200 lines on the primary switches have access to the 200 trunks on the secondary switches. It is evident that another vertical file of ten primary switches may be added with twenty subscriber lines connected to the verticals of each switch, and with the horizontal paths strapped and connected to the horizontal paths of the primary switches shown. This would give 400 lines access to the 200 trunks on the secondary switches. In actual practice on a line link frame, several files of primary switches may be connected together in this manner depending upon the traffic volume of the subscriber lines. The circuit paths connecting the horizontal rows of terminals of the primary switches to the horizontal rows of terminals of the secondary switches are called "line links."

To establish a path from a line circuit on a primary switch to a trunk circuit on a secondary switch, the common "control" circuit serving this line link frame locates the subscriber line to be served and then simultaneously selects an idle "line link" on the primary switch on which the subscriber line appears and a group of trunks wired to a secondary switch in which there are one or more idle trunks. Thus the selection of the line link is made contingent upon the availability of trunks, and by means of this together with the primary-secondary distribution of the links a very efficient usage of the links and trunks is obtained.

In the "line link" frame shown in Fig. 9, it will be seen that the trunks on the verticals of the secondary switches are split into groups of 100 trunks each, one group being connected to the "district junctions" and used for originating traffic and the other group of 100 trunks being connected to "line junctions" and used for terminating traffic.

It will be noticed that there is but one crossbar switch appearance of a subscriber line in the office. This is on a vertical unit of a primary crossbar switch where both the originating and terminating calls are

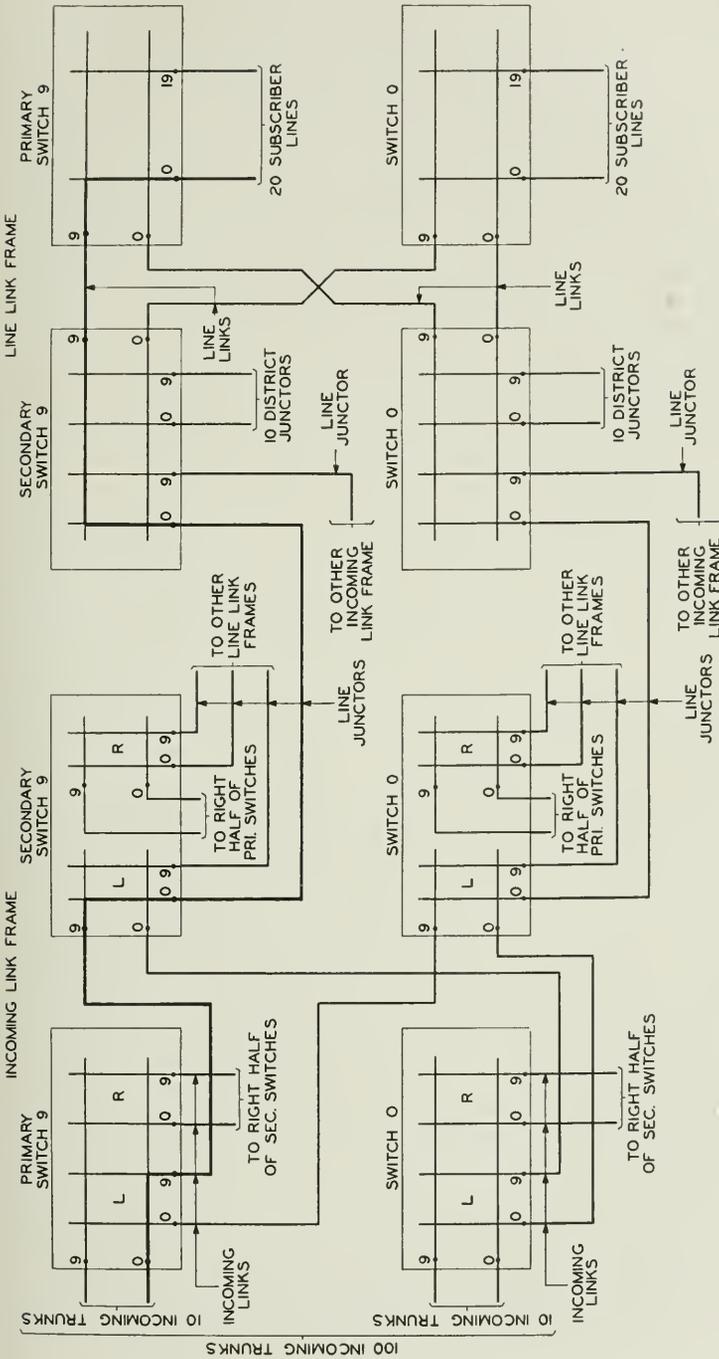


Fig. 10—Double primary-secondary trunking arrangement.

completed by means of the same line link circuits. Thus all originating traffic from any of the twenty lines on any primary switch flows through the associated ten line links to the 100 district junctors and all terminating traffic to these twenty lines flows through the same ten line links from the 100 line junctors.

This single "primary and secondary" trunking arrangement also is used at other points in the system, such as in the originating and terminating sender link switch frames, where the circuits reached are non-directional, that is, where any one of the selectable circuits wired to the frame can be used for setting up a connection.

For the switch frames where the circuits reached are directional, that is, where a particular called line or a particular group of trunks must be used in order to complete a connection, the problem of trunking becomes more complex and it is necessary to provide a trunking arrangement using two "primary and secondary" switch frames arranged in tandem.

Figure 10 shows a typical arrangement of this kind which is necessary to secure the required trunking flexibility and efficiency. This figure shows an "incoming link" frame to which incoming trunks are connected and a "line link" frame to which subscriber lines are connected as described above. These two frames are used in tandem for establishing the terminating connections between the incoming trunks and the called subscriber lines. As is indicated, 100 incoming trunks are connected to the 100 horizontal paths of the ten incoming link frame primary switches, there being ten incoming trunks connected to each of the primary switches. A total of 150 to 700 subscriber lines may appear on the verticals of the primary switches of the line link frame; however, only 200 lines or twenty on the verticals of each of the ten primary switches are shown in the figure.

In order to connect a particular incoming trunk to a particular called line, an idle channel is selected through these two switch frames, consisting of an "incoming link" on the incoming link frame, a "line junctor" between the two frames and a "line link" on the line link frame, and all are connected in series as shown in the figure. It will be noted that the incoming trunks on each of the primary switches have access to twenty incoming links appearing on the twenty verticals of the switch. These twenty incoming links are distributed over the ten secondary switches of the frame, two links being connected to each switch, one to each half switch. It will be observed that in order to provide for the distribution of the twenty incoming links over the ten secondary switches, the horizontal paths of the secondary switches are separated between the tenth and eleventh verticals,

thus taking advantage of the flexibility of the crossbar switch by providing twenty horizontal paths instead of ten on each switch. The incoming links, on each half of these secondary switches, have access to "line junctors" appearing on the verticals of these switches. These line junctors are in turn distributed over the secondary switches of all the line link frames in the office. There will be at least one line junctor as shown, from each secondary switch on an incoming link frame to a secondary switch on every line link frame in the office, or a minimum of ten line junctor paths between any incoming link frame and any line link frame. The number of the line junctors between these frames will vary depending upon the number of frames required in an office. The line junctors on the verticals of each of the line link frame secondary switches in turn have access to ten line links on the horizontal paths. These ten line links are, as described above, distributed over the primary switches of the line link frame, one to each primary switch. These line links then have access to the called subscriber lines which appear on the verticals of the primary switches. With this arrangement of switches and the three groups of interconnecting link paths, any incoming trunk can be connected to any called line on the line link frame shown, or by means of other groups of line junctors, to a called line on any other line link frame in the office.

Terminating markers are employed for selecting the paths through these switches to connect an incoming trunk to a called subscriber line. The marker, as will be explained later, records information which permits it to connect to the test wire and holding magnet of a called line and to the test wires and switch magnets of the groups of incoming links, line junctors and line links through which the incoming trunk may be connected to the called line. The marker simultaneously tests these three groups of paths and "marks" an incoming link, a line junctor and a line link which are idle and are accessible to one another, and then operates the switch magnets to connect these three paths and the incoming trunk and the called line together. The paths are selected in an ordered arrangement, so that the lowest numbered incoming links, line junctors and line links are preferred and are used as long as they are available. This increases the efficiency of the paths as compared with a random selection, since it reduces the chance that one or two of them although idle cannot be used because the third one is busy.

A double primary and secondary trunk arrangement similar to the one shown in Fig. 10 is employed for connecting district junctors to outgoing trunks in the originating end of the office.

Brief Description of Circuit Operation

The operation of the system will be described by tracing the progress of a call through the system. The establishment of a call from one crossbar subscriber to another crossbar subscriber may be divided into four stages: two in the originating end of a connection and two in the terminating end.

1. The calling subscriber is connected to a sender for the purpose of registering the called number which is dialed.

2. The subscriber sender is connected to an originating marker and the marker selects the switch frames for establishing the connection to an outgoing trunk.

3. The outgoing trunk circuit is connected to a sender in the terminating end to register the called number.

4. The terminating sender is connected to a terminating marker and the marker selects the switch frames for establishing the connection to the called subscriber line.

The first stage in the progress of a call is illustrated in Fig. 11. It will be seen that the line of a calling subscriber terminates on a vertical unit of a primary crossbar switch located on a line link switch frame. When the subscriber receiver is lifted from the telephone preparatory to dialing, a line relay is operated, as in other systems, and the circuits proceed with the establishment of the connection to an idle subscriber sender which will register the called number when it is dialed.

The circuit functions on this stage of the call are as follows:

1. The subscriber line is located by the "line link control" circuit which is common to the line link frame, by a coordinate method of testing. That is, the control circuit determines the primary crossbar switch in which the line is located and the particular vertical unit in the switch on which the line is terminated. This operation is similar to the line finder operation in other dial systems, except that the operation is accomplished by relay operations instead of by a mechanically traveling brush.

2. The line link control circuit then simultaneously selects an idle line link between the primary switch in which the line appears, and a secondary switch on which a group of district junctions appears which has at least one idle district junctor in the group and which has access to idle senders and an idle sender link.

3. This will bring into operation the common "sender link control" circuit of the sender link switch frame to which the selected group of district junctions is connected. This control circuit will select an idle district junctor in this group which appears on a primary switch on

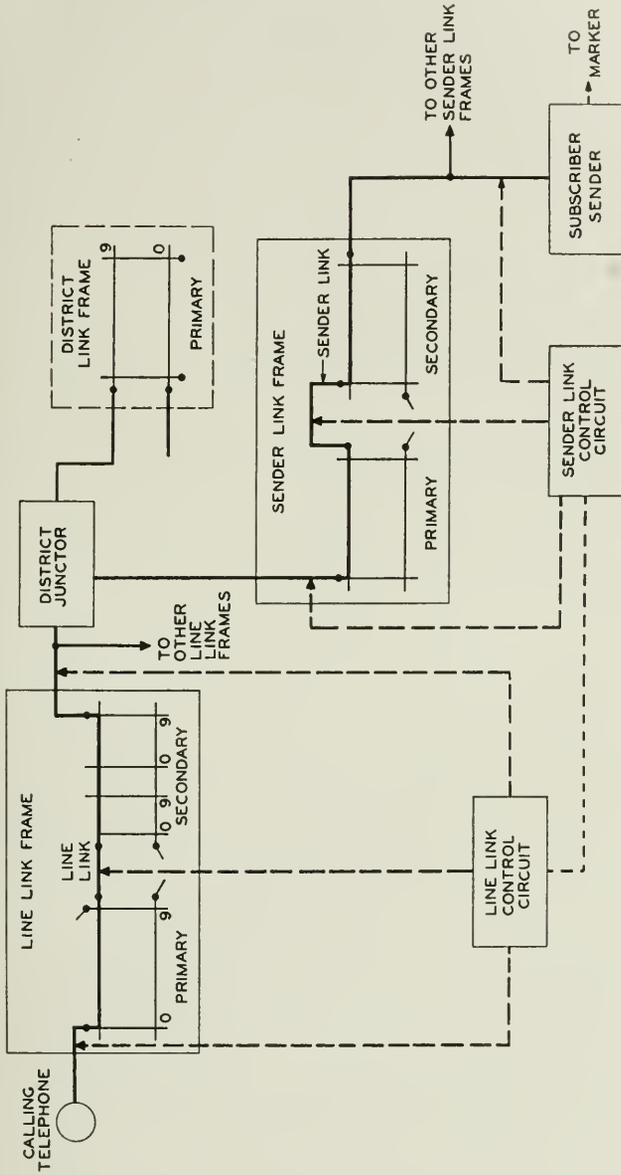


Fig. 11—Calling line connected to district junction and subscriber sender.

the sender link frame. There are ten sender links serving the selected district junctor. These ten sender links and the ten sender groups to which they have access on the secondary switches are then tested simultaneously to find an idle sender link with access to a group of senders in which there are one or more idle senders. When this choice has been made an idle sender in the group is then selected.

4. The two control circuits in cooperation with each other first operate the selecting magnets and then the holding magnets associated with the paths selected on the switches of both the line link and sender link frames, and thereby establish the connection from the calling subscriber to an idle subscriber sender.

This connection may be traced by referring to Fig. 11, from the calling line on the vertical unit on a primary switch of the line link frame, through a line link and a secondary switch, through a district junctor circuit, to a vertical unit on a primary switch of the sender link frame, through a sender link and a secondary switch to a subscriber sender which is connected to a horizontal circuit path on the secondary switch.

5. The two control circuits are then released and made available for use on other calls. The connections through the switches to the sender are held established by means of the holding magnets which are held operated over a signal control lead, called the "sleeve" lead, under control of the relays in the sender, which in turn are under control of the subscriber telephone.

Upon completion of these operations which take but a fraction of a second the subscriber sender transmits the dial tone to the calling subscriber as an indication to dial the number. When the subscriber dials, electrical impulses are transmitted to the sender, which receives and registers them. When the sender has registered the office code, which in New York City for example is contained in the first three digits dialed, the sender will connect itself to an idle originating marker by means of multi-contact relays of a marker connector circuit.

Before proceeding further it is desirable to mention several other functions of the two common control circuits used for setting up this part of the connection.

1. The control circuits signal to the sender the class of the calling line, that is, for example, whether the line is a coin line or a non-coin line.

2. The sender link control circuit signals to the sender the number of the district link switch frame on which the selected district junctor appears, since this identification will be used later in the establishment of the connection.

3. The sender link control circuit tests the circuit paths chosen from the line circuit to the sender before disconnecting from the connection, in order to insure the proper establishment of the connection. In case of a failure the control circuits will make repeated trials to establish the connection over different paths and give an alarm to the maintenance force.

4. Emergency control circuits are provided for use in case the regular control circuits are removed from service for maintenance reasons.

The next stage in the progress of the call is illustrated in Fig. 12. In this stage of the call the principal control unit is the originating marker. Its major function is to control the switches in the establishment of the connection to an idle outgoing trunk circuit to the called office, which may terminate in a distant office or in the same office as the calling subscriber.

When the subscriber sender connects to the originating marker through the connector circuit, the sender transfers the called office code indication and the district link frame identification to the marker circuit. The called office code indication causes the operation of a "route" relay in the marker corresponding to the particular office called.

There are a number of "route" relays in each marker and one is assigned to each called office routing. The route relay is connected as required by the office code to which it is assigned, so that it will direct the marker to the trunks of the called office and to the office link switch frame on which these trunks appear and indicate the number of trunks in the group. The route relay also is connected to determine the type of the called office, such as Crossbar, Panel, or Manual, and to set up the corresponding circuit conditions in the subscriber sender to enable the sender to handle the connection properly after the marker has been released. The connections of the route relay contacts to the control relays in the marker are made flexible so as to permit the assignment of any route relay to any office code and to permit changes to be made from time to time in the route information, changes in trunk group sizes and location, changes in the type of terminating office, etc. The route relays and associated flexible connection facilities represent a considerable portion of the marker equipment, especially in large metropolitan offices where several hundred central offices are involved.

When the route relay is operated, the marker proceeds with the establishment of the connection as follows:

1. It connects to the office link frame on which the trunks to the

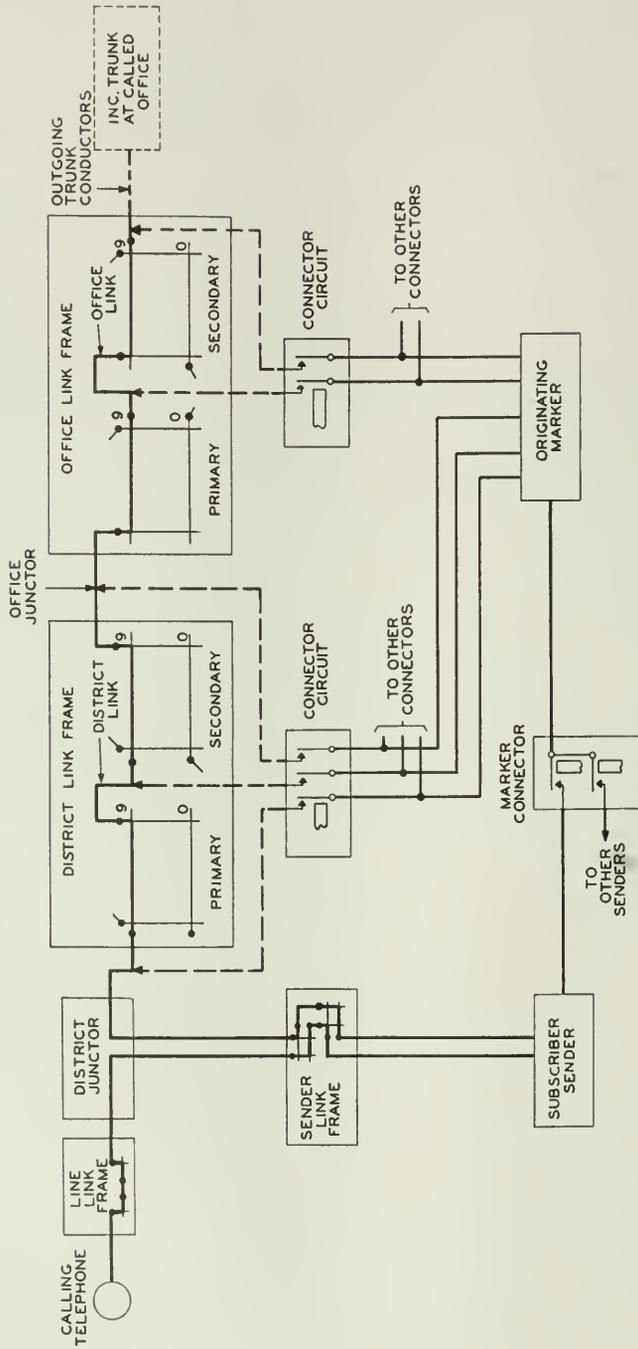


Fig. 12—District juncture and subscriber sender connected to the outgoing trunk.

called office appear. This connection is made through the office link frame connector circuit, one of which is provided for each office link frame. Through this connector the marker is extended to the test leads of any desired trunk group on the office link frame and to the crossbar switches of the frame. When so connected the marker has exclusive control of the trunks and switches of the frame and other markers which desire connection to the same frame are deprived of access until the connected marker releases.

2. The marker next tests the outgoing trunks to the called office and selects an idle one. If, as determined from the route relay, the trunks are divided over more than one frame, the marker will connect to the second group of trunks on the second office link frame in case the first group of trunks is found to be busy..

3. The marker also connects by means of a connector circuit to the district link frame associated with the district junctor to which the calling line is connected. The identification of this frame was obtained from the sender and the sender link control circuit as previously mentioned. Through the connector circuit of the district link frame the marker is extended to the control leads of the district junctor circuit and the crossbar switches of the district link frame. As in the case of the office link frame, only one marker is connected to a frame at a time.

4. The marker after selecting an idle trunk circuit which appears in a horizontal circuit path on one of the secondary switches of the office link frame, then proceeds with the selection of an idle channel through the switches of the two switch frames. A number of these connecting channels is provided between the district junctor and the outgoing trunk. Each channel consists of a "district link" on the district link frame, of an "office link" on the office link frame, and an "office junctor" connecting the district link frame to the office link frame. The marker tests a group of these channels simultaneously and selects an idle one. It then operates the switch magnets which will connect these three paths of a channel, and the district junctor and the outgoing trunk together, thereby establishing a connection from the district junctor to the outgoing trunk.

5. When the marker has completed this operation it checks the connection to insure that it has been properly established and that it is capable of being held under control of the district junctor, before releasing itself from the connection.

6. The marker performs these functions in approximately .5 second, then releases and becomes available for use on other calls.

It will be observed that the three links involved in establishing the

connections between the district junctors and the outgoing trunks are used in series and are chosen simultaneously. Generally in other systems the establishment of a connection involving three such paths, is made in three successive stages with a possibility that after a selection has been made at one stage it will be found that the paths accessible to it are all busy and, therefore, the connection cannot be completed

Before describing the next stage in the establishment of a call, it is desirable to point out other features and functions of the originating marker.

1. The marker permits wide variations in the sizes of trunk groups, permitting trunk groups as small as two and as large groups as may be required. This makes for an efficient use of the office link frame terminals and thereby tends to reduce the office link frame equipment.

2. The marker makes a second trial to establish connections over alternate trunk routes in case calls cannot be completed over the normally used groups because of busy conditions.

3. The marker makes a continuity test of the circuits over which the switches are controlled and tests them for short-circuits, crosses, opens and grounds which would interfere with the proper establishment of a call and where troubles are detected, it signals this condition to a common "trouble indicator" where an indication of the trouble and its location is recorded and a maintenance alarm given. The call is then completed over another group of circuits.

The first stage in the progress of the call through the terminating end of a crossbar office is illustrated in Fig. 13. It consists of connecting the incoming end of the selected trunk to a terminating sender for the purpose of receiving the number of the called line from the subscriber sender.

When the incoming trunk is selected by the originating end of the office equipment, the "sender link control" circuit associated with the terminating sender link frame on which the incoming trunk appears, is called into action. The control circuit then proceeds with the following functions:

1. To locate the incoming trunk circuit, which appears on one of the ten horizontal paths of a primary switch.

2. It selects an idle sender link between this primary switch and a secondary switch on which there is an idle terminating sender.

3. The control circuit selects one of the idle terminating senders reached through the secondary switch and then operates the selecting and holding magnets associated with the selected circuits, which will establish the connection from the incoming trunk to the terminating sender.

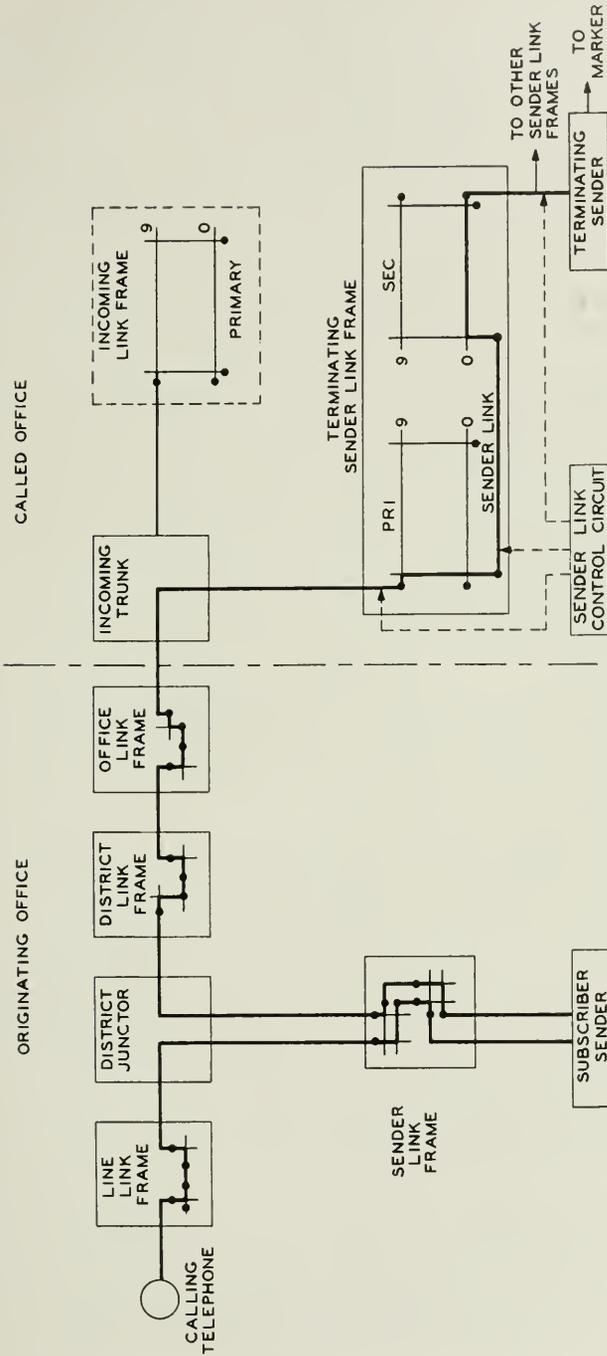


Fig. 13—Connection established to terminating sender.

4. The control circuit will signal to the terminating sender the number of the incoming link frame in which the incoming trunk appears. This frame identification will be used later in establishing the connection to the called line.

5. The control circuit will then disconnect after checking to insure that the connection to the sender has been properly established and that it will be held under control of the trunk and sender circuits after the control circuit leaves the connection.

As soon as this operation has been completed, which takes but a fraction of a second, the terminating sender will be in direct connection with the subscriber sender in the originating end of the connection. This path may be traced, by referring to Fig. 13, from the subscriber sender through the sender link frame, through the district junctor, through the district link and office link frames, over the outgoing trunk to the incoming trunk, and through the terminating sender link frame to the terminating sender.

At this stage of the connection the calling subscriber is still connected with the subscriber sender, and dialing may be still in progress. As the subscriber proceeds with the dialing of the digits of the called number, the subscriber sender will transfer them to the terminating sender. This is done by means of impulses transmitted over the circuit paths between the two senders. When the subscriber sender has completed the transfer of the called number to the terminating sender, the subscriber sender will be released and the calling line will then be connected through the district junctor to the incoming trunk.

When the terminating sender has secured the record of the called line number, the sender then connects to an idle terminating marker by means of multi-contact relays of a connector circuit.

The next stage in the progress of the call is shown in Fig. 14. The terminating marker is the principal control unit at this point in the connection. Its principal function is to provide means for establishing the connection from the incoming trunk to the called subscriber line.

When the terminating sender has connected to the terminating marker, the sender will transfer both the called line number and the incoming link frame identification to the marker. The terminating marker then proceeds to establish the connection to the called line as follows:

1. It connects itself to the particular "number group connector" circuit including the "block relay" frame in which the called line appears in its numerical sequence. All subscriber lines are provided with a set of three test terminals which appear on the block relay

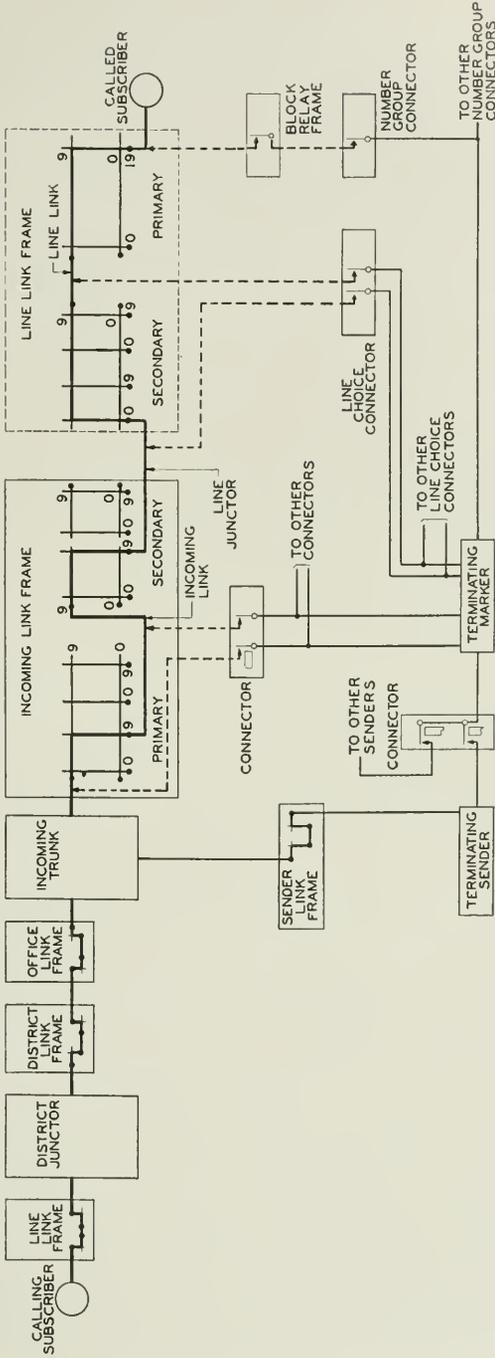


Fig. 14—Connection established to called line.

frame. These terminals correspond to the director number of the subscriber line. A number group connector generally has access to the test terminals of several hundred line numbers depending on the terminating traffic to the lines.

2. The marker will obtain a connection through the number group connector to the busy test terminal of the particular called line and determine whether the line is busy or idle.

3. It will determine from the two other test terminals, the identification of the line link frame where the called line appears, and the horizontal group of line links which has access to the called line. In addition, it determines the type of ringing to be applied to the called line from the circuit conditions on the test terminal.

4. Assuming that the called line is idle, the marker will connect, through the line choice connector circuit, to the line link frame and to the ten line links which have access to the called line.

5. It will then connect, through the connector circuit, to the incoming link frame associated with the incoming trunk to which the calling subscriber line is now connected. The incoming link frame identification was obtained from the sender link control circuit through the sender as previously mentioned.

6. The marker will then select an idle channel through the incoming link and line link frames as previously described. This channel will consist of an "incoming link," a "line junctor," and a "line link" all to be connected in series. The marker then operates the proper selecting and holding magnets of the crossbar switches in each frame which establishes the connection from the incoming trunk to the called subscriber line.

7. The marker will then cause the incoming trunk to start the proper ringing over the called subscriber line and to transmit the ringing tone signal over the trunk to the calling subscriber.

8. At this point the terminating marker and the terminating sender will have completed their functions and, together with the terminating sender link frame, will be released. The complete connection will then be established from the calling line to the called line and the conversational circuit completed when the called subscriber answers.

If the terminating marker finds the called line busy it will cause the incoming trunk circuit to transmit a busy tone to the called subscriber.

The terminating marker has the following other important functions:

1. If the call is for a PBX (Private Branch Exchange) the condition on one of the test terminals of the called line in the number group connector will inform the marker that the line is one of a group of lines. The marker will test all of the lines in the group, testing up to as many as twenty simultaneously, and will select an idle one.

The lines of a PBX may be assigned to non-consecutive numbers within the usual 10,000 series, and with the exception of the numbers dialed, they may be assigned to line numbers in a special group of 2500 outside of the 10,000 series. These features reduce the necessity for number changes due to the growth of private branch exchanges, and conserve subscriber line numbers in the office. The lines of a PBX group can be distributed over several line link frames and over two number group connectors to equalize the terminating traffic load in the case of busy private branch exchanges.

2. The marker recognizes numbers dialed which are unassigned, disconnected or changed numbers, and automatically routes such calls to an operator who will inform the calling subscribers as to the status of the numbers called.

3. In case the called number is on a party line, the marker determines from one of the test terminals which station of the line is to be rung, and signals the incoming trunk to provide the proper ringing.

4. The marker tests the continuity of the circuit paths to be used to the called line before establishing the connection, to insure that the connection is properly set up and that it will be held under control of the subscriber telephone after the marker disconnects. The marker also tests for short-circuits, crosses and grounds, and in case of a failure due to any inoperative condition it will connect itself to the common trouble indicator and leave a record of the trouble and its location and give an alarm to the maintenance force.

Figure 15 shows the complete talking connection through the various trunks and switch frames as finally established after all of the common control circuits have been released.

On a call to a subscriber served by a panel dial office, the connection is routed through the district link and the office link frames in the same manner as on a call terminating in a crossbar office, but in this case the idle trunk chosen on the office link frame terminates in an incoming panel switch in the distant panel dial office. The subscriber sender of the crossbar office causes the incoming and final selectors in the terminating panel office to select the called subscriber line without the aid of any terminating senders in either office. When the subscriber sender has completed these functions it will be released and the connection will be established from the calling line over the inter-office trunk circuit and through the terminating panel incoming and final selectors to the called line. On this type of call the subscriber sender operates in the same manner as though the called line were in a crossbar office, and the selectors in the panel office operate in the same manner as though the call had originated in another panel office.

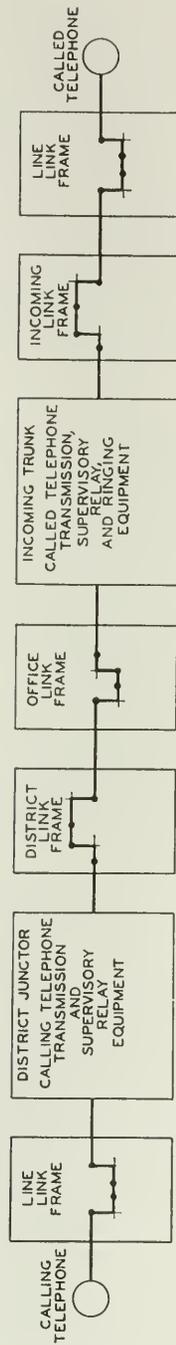


Fig. 15—Completed talking connection.

No changes are required in the panel selectors to function with the crossbar office.

On a call for a subscriber in a manual office, the call would be routed through the switches of the district link and the office link frames as previously described and connected to a trunk circuit on the office link frame which terminates in the "B" switchboard in the manual office. The subscriber sender of the crossbar office then transfers the called number by impulses transmitted over the interoffice trunk circuit to the operator's position equipment in the manual office. The called number appears in the form of visible numbers on the operator's keyshelf. The operator completes the connection by "plugging" the associated trunk circuit, which terminates on a cord and plug, into the called subscriber line jack.

A call originating in a panel dial office for a subscriber line in a crossbar office reaches the crossbar office through an incoming trunk circuit as in the case where the call originated in a crossbar office. The call from the panel office is then handled by the crossbar office terminating sender and marker in exactly the same manner as described for calls originating in the same crossbar office.

A call originating in a manual office for a line connected to the crossbar office reaches the crossbar office over an incoming trunk circuit from an "A" operator's position in the manual office. These incoming trunks in the crossbar office are similar to the incoming trunks previously described. In this case, however, the incoming trunk is connected to a terminating "B" sender and by means of this sender to a "B" board operator in the crossbar office. The "B" operator will obtain the called number verbally from the distant "A" operator and then, by means of the keyset on her position, register the called number in the terminating sender. The terminating sender will then select a terminating marker and the connection will be established in exactly the same manner as described for a call originating in a crossbar office.

MAINTENANCE FACILITIES

Automatic routine testing circuits are provided for testing all the principal circuit units, such as the district junctors, incoming trunks and senders. These test circuits automatically put each circuit, one after the other, through all of its functions on all classes of calls to insure that it performs satisfactorily. It tests the important relays of the circuits to insure that they have the proper adjustment to handle the worst circuit conditions. In case any circuit fails to meet the test conditions, the test is stopped and an alarm given to the maintenance force.

Trouble indicator circuits are provided for use in connection with the test and maintenance of the marker circuits. These circuits are arranged so that when trouble is encountered by a marker, the marker will seize the trouble indicator and operate combinations of relays and light small lamps which indicate the nature and the location of the failure and give an alarm to the maintenance force.

EQUIPMENT

Figure 16 shows a typical switch frame used in the crossbar system. This particular frame is a "line link" frame which serves a group of subscribers for both originating and terminating traffic. The frameworks on which the equipment is mounted are constructed of rolled

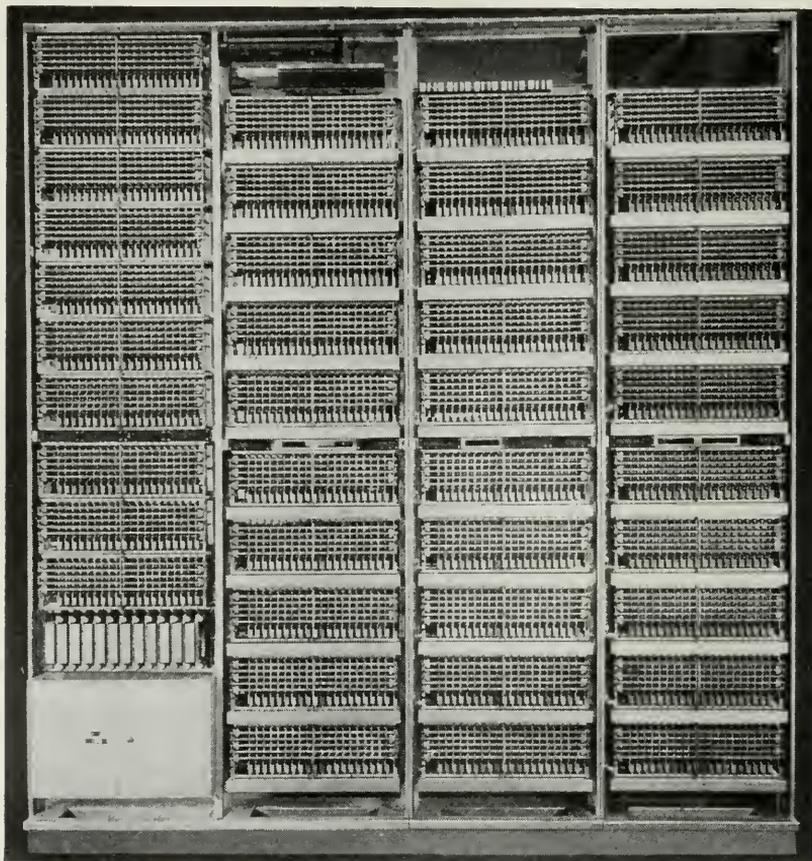


Fig. 16—Line link frame.

bulb angle iron sections with a sheet metal base. The bulb angle construction provides a framework which is light in weight and has the required strength, and permits an equipment mounting arrangement which conserves space and facilitates the wiring of the apparatus. The frames are welded and incorporate such features as sanitary base construction, guards to protect the apparatus and wiring against damage from the rolling ladders located between the rows of frames, and a cable duct or runway for the A.-C. power service cables with plug receptacle outlets for use with electric soldering irons, portable lights, etc.

These frame equipments are built in standardized units, which provide the required flexibility to satisfy the variations in telephone traffic and classes of service encountered in the different telephone areas. Where it has been necessary to divide an equipment assembly into several units, due to the limitations of handling, shipping and to care for different classes of service, the equipments have been designed so that the installation effort required for interconnecting such units has been reduced to a minimum.

The bays of equipment located at the right, in Fig. 16, equipped with crossbar switches, are the primary line link bays. The vertical units of these crossbar switches are wired to the subscriber lines. These primary bays are made available in units of 100 and 200-line capacities. As discussed previously the number of primary bays provided in a line link frame may be varied to fit the traffic load of the subscriber lines. The left-hand bay of this frame contains the vertical file of crossbar switches, known as the secondary switches and the vertical units of these switches are wired to district junctions and line junctions. The line link control circuit apparatus, which is common to the frame, is located at the bottom of this bay.

Figure 17 shows a group of three frame units, namely, the subscriber sender link, the district junctor and the district link frames, which are closely associated in the trunking network and have been designed as a fixed equipment group. However, for shipping reasons the group is divided into three separate equipment units. The district junctor circuits, consisting primarily of relays, are mounted in groups on the middle frame. These groups are provided in standardized units of various types, such as those required to serve coin and non-coin subscribers lines. A similar arrangement of frames is used for the combination of terminating sender link, the incoming trunk, and the incoming link frames.

Figure 18 shows a row of subscriber sender frames and a frame of "A" operator senders located at the extreme right. These frames

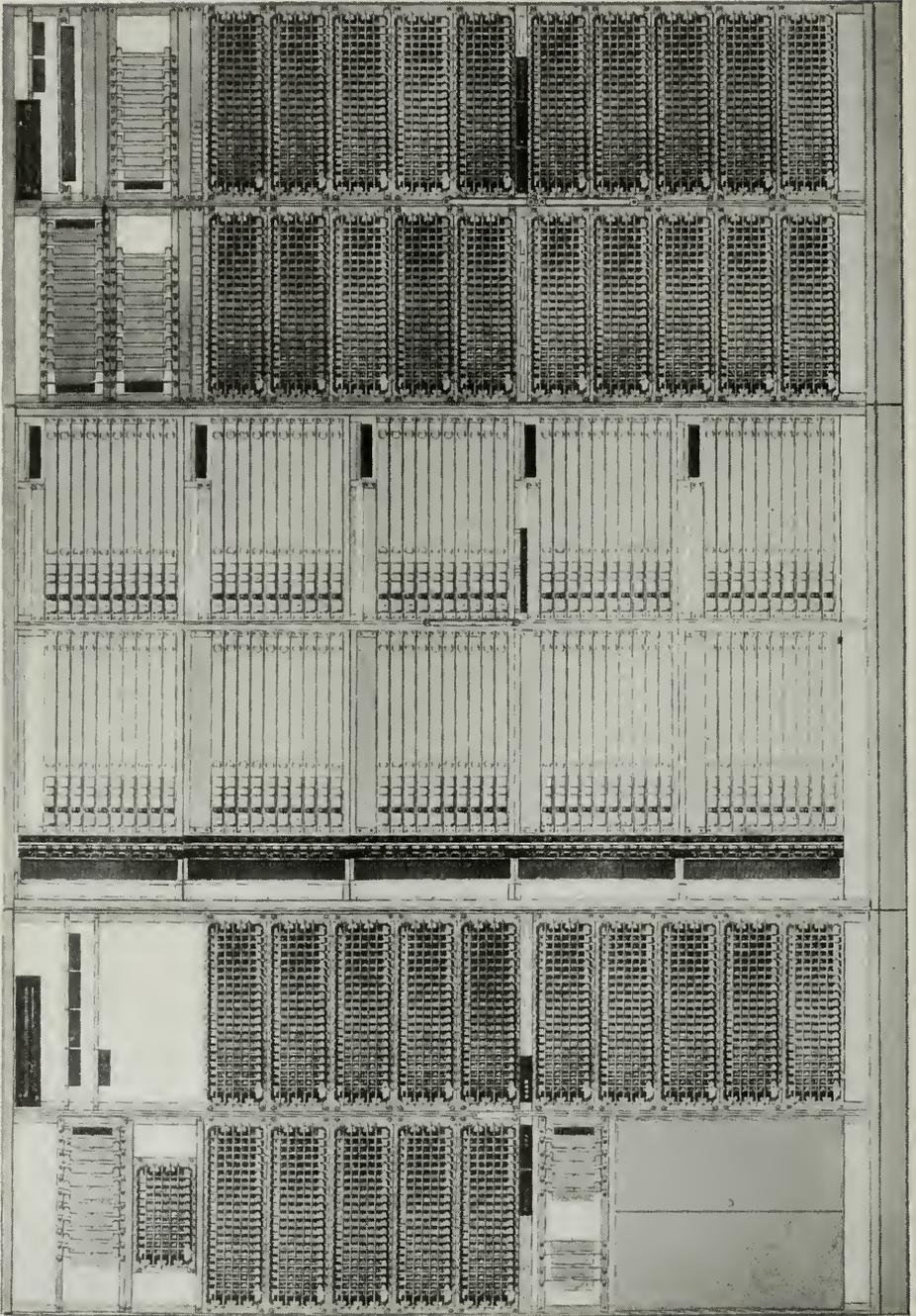


Fig. 1. A typical frame of a distributor-connector distributor line and corder line frames.



Fig. 18—Sender frames.

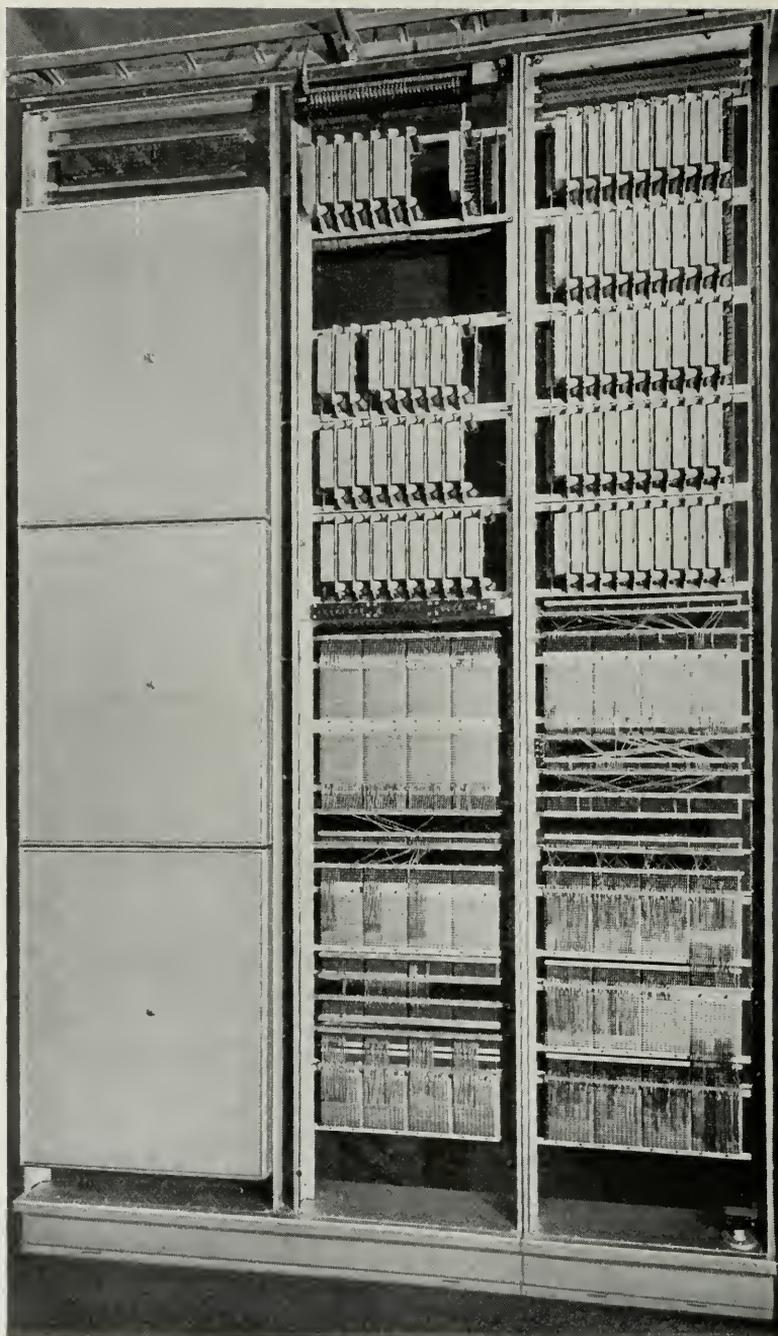


Fig. 19—Originating marker frame.



Fig. 20—Battery supply feeders, power wiring, fusing, etc.

accommodate five senders which may be of one type, or a combination of both types. The crossbar switch shown on the right of each subscriber sender unit, is a part of the sender circuit and is employed for the purpose of registering the called numbers dialed by the subscribers.

The "A" operator senders are associated with the "A" operator switchboard equipment and are used for the completion of certain classes of calls such as toll and assistance calls.

A view of the originating marker frame is shown in Fig. 19. There will be a variation in the equipment on this frame for different cities due to the variation in the number of route relays required, the number depending upon the number of central office codes that may be dialed by subscribers and operators. This variable feature is cared for by providing the route relay equipment in bays of 100 codes as shown in the right-hand bay. The terminal fields shown below the route relays on the frame provide the flexible connecting facilities which permit the use of any route relay for any office code and which readily permit changes in routings, variations in trunk group sizes and other features which are subject to change from time to time.

The power plant equipment provided for the crossbar offices is similar to the equipment now being furnished for all large dial central offices. The principal power supply arrangements provide 48-volt direct current for the operation of practically all the signaling and the telephone transmission circuits. Also several other sources of direct current are provided for miscellaneous purposes as in other standard dial systems. A new distribution scheme for the battery feeders on the frames is employed which reduces the amount of copper required. A common set of 48-volt battery feeders supplies the signaling and talking current for all frames. Individual frame filters are connected across the battery supply leads at the frames where a noise-free battery supply is required for talking circuits. Figure 20 shows a view of the overhead battery cables, conduits for the A.-C. power leads, and the fuse cabinets for the fusing of the battery supply to a row of frames.

APPLICATION

As mentioned in the first part of this paper, two crossbar dial central offices were cut into service in 1938 and these have now been in commercial operation for several months. One of these offices serves a residential area in Brooklyn, while the other serves a congested business area in the midtown Manhattan district of New York City. The operation of these offices under actual service conditions has been highly satisfactory and our expectations in regard to performance have been fully realized.

This type of system will be used for new offices in large cities instead of the panel system as rapidly as manufacturing and plant conditions permit and the apparatus which was designed for this system will be used in other fields of the telephone system.

A Twelve-Channel Carrier Telephone System for Open-Wire Lines

By B. W. KENDALL and H. A. AFFEL

A new carrier telephone system is described, together with its application in the long distance telephone plant. By its use, an open-wire pair which already furnishes one voice circuit and three carrier circuits may have twelve more telephone circuits added. Thus in all sixteen telephone circuits are obtained on a single pair. Several such systems may be operated on a pole line.

Various problems incident to the extension of the frequency range, from about 30 kilocycles, the highest frequency previously used, to above 140 kilocycles, are discussed. Among the more important of these are the control of crosstalk between several systems on a pole line, arrangements for taking care of intermediate and terminal cables, and automatic means for compensating for the effects of weather variations on the transmission over this wide frequency range.

INTRODUCTION

BARE wires supported on insulators, stretched between poles, make up the pioneer electrical communication circuit, the open-wire line. Although great advances have been made in the application of cable structures, the open-wire lines still hold their own in some sections of the country. This is because, to offset their physical vulnerability, they have several unique virtues. They are flexible and permit adding one pair of wires at a time. They are also comparatively economical where conditions favor their use. Furthermore, they are low-attenuation circuits and for this reason were the first to be used for high-frequency carrier systems.

The first carrier systems, beginning in 1918, added three or four channels to the existing voice circuit on a pair. To keep pace with this development, improvements in transposition systems were devised so that many such carrier systems might be operated on the same pole line. Such carrier systems, typified by the three-channel type C¹ system, have seen continuous growth in use in the long distance plant. Now a twelve-channel system, the type J, is being made

¹ "Carrier Systems on Long Distance Telephone Lines," H. A. Affel, C. S. Demarest and C. W. Green, *Bell System Technical Journal*, July 1928, and *A. I. E. E. Transactions*, Oct. 1928, pp. 1360-1387. "A New Three-Channel Carrier Telephone System," J. T. O'Leary, E. C. Blessing and J. W. Beyer, *Bell System Technical Journal*, this issue.

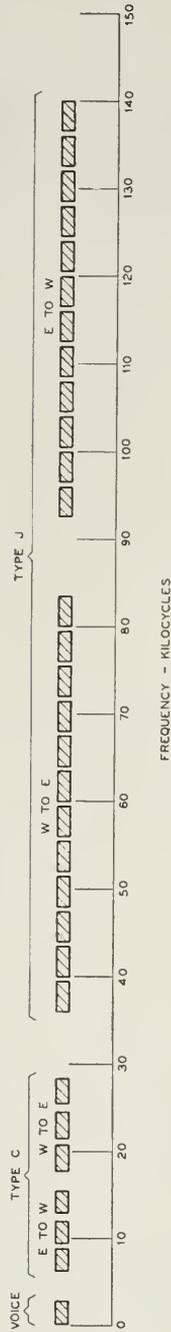


Fig. 1—Frequency allocations.

available to add to the type C system, thus giving sixteen telephone circuits on an open-wire pair in addition to the two telegraph circuits. Since there are already about 60,000 miles of pole line equipped with type C systems, the new type J system was developed to go in the frequency range above the type C system rather than to supersede it with more channels (Fig. 1).

The new system has been designed to meet high standards of transmission and reliability for distances up to several thousand miles. The frequency band transmitted by the individual derived circuits is exceptionally wide, from about 100 to 3600 cycles for a single system and has been previously discussed² in relation to the channel spacing in this and other new broad-band developments.

An important feature of the work on the type J system has naturally been that of making the line circuits suitable for carrying the higher frequencies. The tendency of circuits to crosstalk into one another increases rapidly with frequency. Advances in transposition design and structural improvements have now made it possible to extend the frequency range from 30,000 cycles to 140,000 cycles, which is about the upper frequency of the type J system. The problem of incidental cables in open-wire lines has also been serious, since the losses increase with frequency, and what is usually more important, there may be substantial reflection effects at junctions of the open-wire line and cable. These are serious, not only from the standpoint of the transmission loss which they entail, but from their effect on crosstalk. The increase in attenuation at the higher frequencies has also brought other problems into the picture. For example, repeaters are needed at more frequent intervals than with the lower frequency systems. Attenuation variation with frequency due to weather changes is greater than at the lower frequencies.

Figure 2 shows schematically the complete type J system, with its different major circuit elements, resulting at the terminals in the division of the single line circuit effectively into sixteen talking circuits. In no recent development is the function of the wave filter in providing essential units in a frequency dividing plan more forcefully illustrated than in the application of this new system, in combination with the type C and other facilities which exist. There are about sixty different designs of filters and networks in the terminals and repeaters. Their functions are varied,—as, for example, separating the individual channel bands, separating the opposite directional groups of channels, separating the type J frequency range as shown in Fig. 1 from the type C and other ranges, separating the different carrier frequencies

² "Transmitted Frequency Range for Circuits in Broad Band Systems," H. A. Affel, *Bell System Technical Journal*, October 1937.

of a carrier supply system in which the carriers are all derived from a common 4000-cycle source, etc.

The new system, as in the case of the type C, uses single sideband transmission with carrier elimination. Copper-oxide units are employed as translator elements of various kinds,—modulators, demodulators, and harmonic producers. Methods of mounting the equipment, and methods and apparatus for testing follow lines already worked out for the type K cable carrier system, which was described a year ago in two A. I. E. E. papers.³

CHANNEL TERMINALS

A terminal of the type J system changes twelve independent voice channels into a compact block of twelve carrier channels properly allocated in frequency for transmission over the open-wire line. Inversely, such a block received from the open-wire line is separated and transformed into twelve independent voice channels. The first step in transmitting the twelve voice channels is to modulate them on twelve carrier frequencies 4 kilocycles apart from 64 to 108 kilocycles and to select the lower sidebands by means of quartz crystal channel band filters. The last step in the conversion from a received twelve-channel block to the twelve independent voice channels consists in the division of the block by twelve quartz crystal channel filters and the demodulation of these messages to produce voice frequency transmissions. These two frequency changes and separations are performed by the same equipment that is used in the type K cable carrier system terminals.

Figure 3 shows the circuit of a modulator and a demodulator for the opposite directions of a single conversation with indicated connections for the eleven others which make up this fundamental twelve-channel block. The modulator consists of a bridge assembly of copper-oxide varistors and is supplied with about 0.5 milliwatt of carrier power from the carrier supply system which is described later. Of the two resulting sidebands, the lower is selected by the crystal band filter following the modulator. The line sides of twelve modulator band filters are joined in parallel and a compensating network is connected to preserve the band characteristics of the upper and lower channels.

On the receiving side, after separation by one of the twelve parallel filters the sideband is applied to a demodulator supplied with the

³ "A Carrier Telephone System for Toll Cables," C. W. Green and E. I. Green, *Bell System Technical Journal*, January 1938 and *Electrical Engineering*, May 1938. "Cable Carrier Telephone Terminals," R. W. Chesnut, L. M. Ilgenfritz and A. Kenner, *Bell System Technical Journal*, January 1938 and *Electrical Engineering*, May 1938.

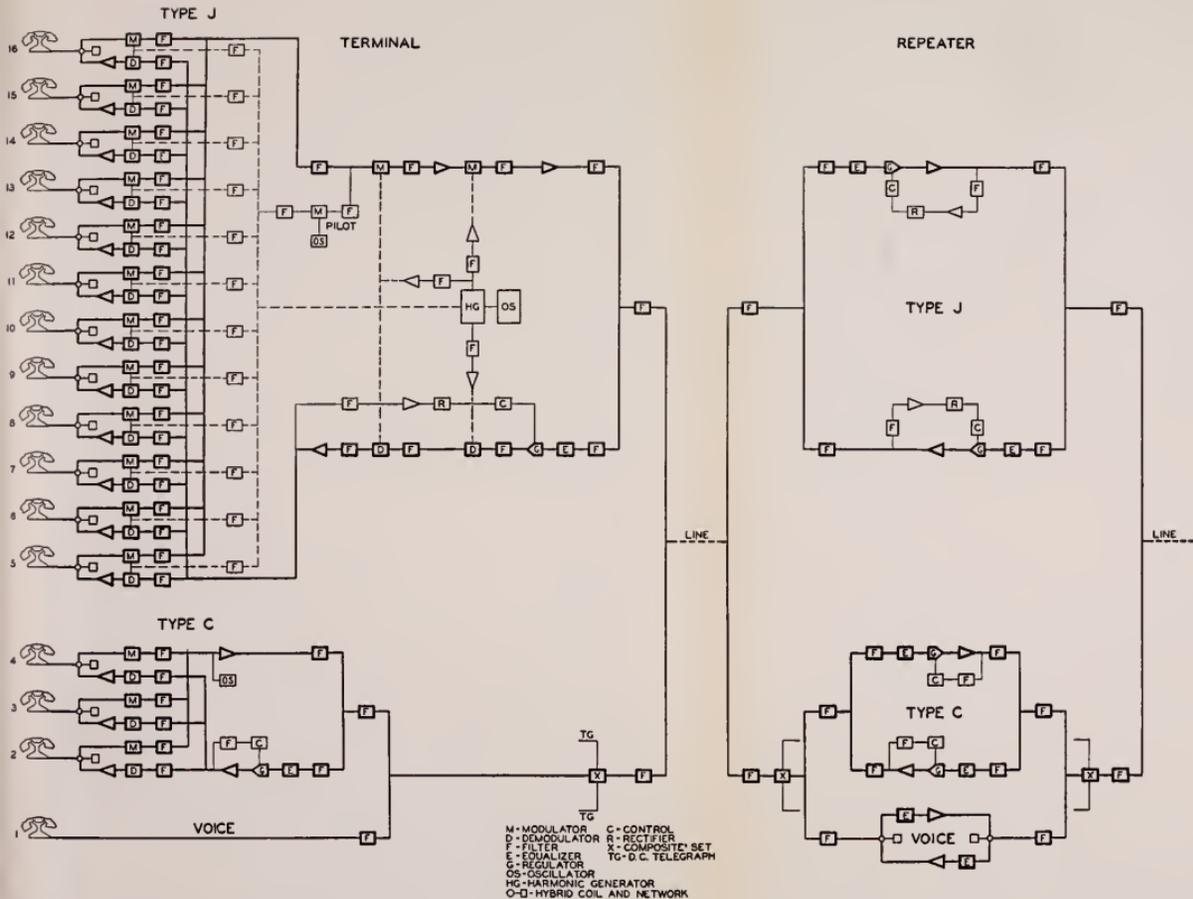


Fig. 2—Terminal and repeater layout.

of a carrier supply system in which the carriers are all derived from a common 4000-cycle source, etc.

The new system, as in the case of the type C, uses single sideband transmission with carrier elimination. Copper-oxide units are employed as translator elements of various kinds,—modulators, demodulators, and harmonic producers. Methods of mounting the equipment, and methods and apparatus for testing follow lines already worked out for the type K cable carrier system, which was described a year ago in two A. I. E. E. papers.³

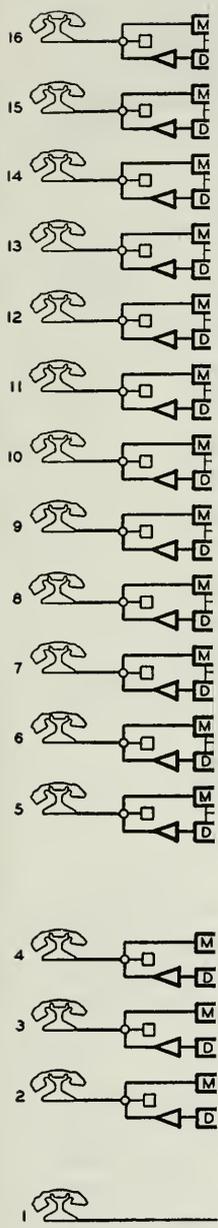
CHANNEL TERMINALS

A terminal of the type J system changes twelve independent voice channels into a compact block of twelve carrier channels properly allocated in frequency for transmission over the open-wire line. Inversely, such a block received from the open-wire line is separated and transformed into twelve independent voice channels. The first step in transmitting the twelve voice channels is to modulate them on twelve carrier frequencies 4 kilocycles apart from 64 to 108 kilocycles and to select the lower sidebands by means of quartz crystal channel band filters. The last step in the conversion from a received twelve-channel block to the twelve independent voice channels consists in the division of the block by twelve quartz crystal channel filters and the demodulation of these messages to produce voice frequency transmissions. These two frequency changes and separations are performed by the same equipment that is used in the type K cable carrier system terminals.

Figure 3 shows the circuit of a modulator and a demodulator for the opposite directions of a single conversation with indicated connections for the eleven others which make up this fundamental twelve-channel block. The modulator consists of a bridge assembly of copper-oxide varistors and is supplied with about 0.5 milliwatt of carrier power from the carrier supply system which is described later. Of the two resulting sidebands, the lower is selected by the crystal band filter following the modulator. The line sides of twelve modulator band filters are joined in parallel and a compensating network is connected to preserve the band characteristics of the upper and lower channels.

On the receiving side, after separation by one of the twelve parallel filters the sideband is applied to a demodulator supplied with the

³ "A Carrier Telephone System for Toll Cables," C. W. Green and E. I. Green, *Bell System Technical Journal*, January 1938 and *Electrical Engineering*, May 1938. "Cable Carrier Telephone Terminals," R. W. Chesnut, L. M. Ilgenfritz and A. Kenner, *Bell System Technical Journal*, January 1938 and *Electrical Engineering*, May 1938.



proper carrier frequency to restore the voice frequency message. Because of the low level at which demodulation takes place, the demodulator is followed by a single-stage amplifier to produce the level desired in the voice frequency circuit. The gain of this amplifier is adjustable over a moderate range.

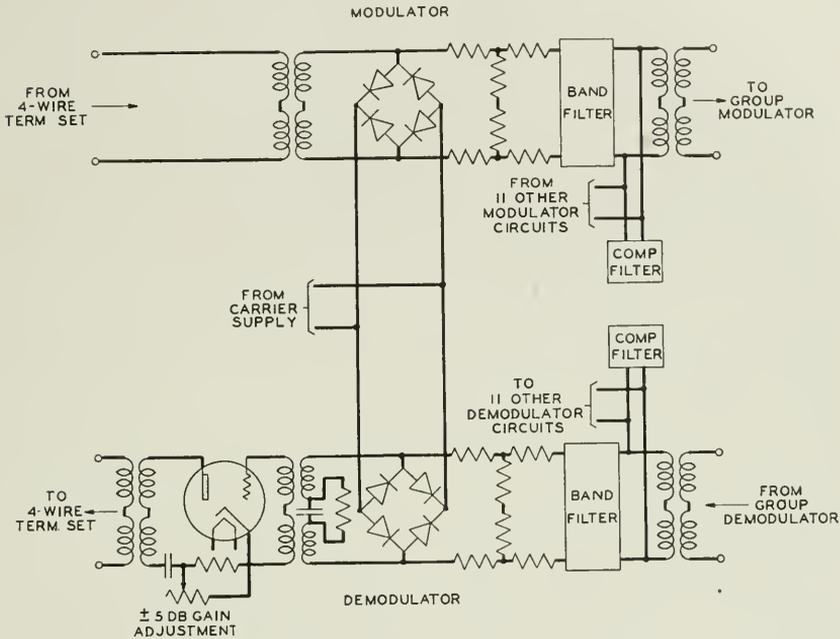


Fig. 3—Channel modulator and demodulator.

The combination of a single modulator and a single demodulator and associated equipment shown in Fig. 3 is called a “Modem” and two of these are mounted on a single equipment panel. Nine of these panels, sufficient for one and a half type J systems, or eighteen conversations, mount in a single relay rack bay of standard height.

CARRIER SUPPLY

The carrier frequencies 64–108 kilocycles are all derived as harmonics of a 4-kc frequency produced by a tuning fork controlled oscillator. This frequency is applied to an easily saturated coil to produce a sharply peaked wave which is rich in odd harmonics. Even harmonics of 4 kilocycles are obtained by rectification in a copper-oxide unit of part of the odd harmonic output. Odd and even harmonics appear in separate circuits from which each frequency desired is separated by a quartz crystal filter. Frequencies as high as the 121st

harmonic, that is, 484 kilocycles, are obtained in this way from the carrier supply system. Because of the importance of the carrier supply two sources are provided, with automatic equipment to transfer rapidly from the regular to the emergency source.

GROUP MODULATION

As shown in Fig. 1, the type J system uses a band of 36 to 84 kilocycles for the west to east direction of transmission and 92 to 140 kilocycles for the east to west direction. The output of the fundamental twelve-channel unit consists of twelve lower sidebands from carriers of 64-108 kilocycles. This must, therefore, be translated to the two type J directional groups for line transmission. Since the frequencies in the fundamental unit overlap those in both directions of line transmission, this transfer must be made in two steps. Figure 4 shows these frequency translations. The first group modulation is the same for both directions of transmission. By modulating the fundamental unit with a carrier of 340 kilocycles there is obtained a block of lower sidebands extending from 400 to 448 kilocycles. A second modulation with a 484-kc carrier then gives, for transmission from west to east, a twelve-channel block of upper sidebands extending from 36 to 84 kilocycles. For the east to west transmission the second modulation uses a 308-kc carrier, producing a twelve-channel block of lower sidebands between 92 and 140 kilocycles.

Frequencies as high as 308, 340 and 484 kilocycles are chosen for group modulation in order that undesired products shall be well separated from desired products to permit their elimination by simple filter structures.

The same group modulation processes that have been described above for adapting the twelve-channel group for line transmission are used in the opposite sequence for receiving the block from the line and preparing it for separation by the channel band filters of the receiving terminal; thus, for instance, at an east terminal the block of upper sidebands, extending from 36 to 84 kilocycles as received from the line, is first modulated with 484 kilocycles producing lower sidebands between 400 and 448 kilocycles. These are next modulated with 340 kilocycles, which produces a block of twelve lower sidebands extending from 60 to 108 kilocycles, which is the group that the fundamental twelve-channel terminal unit is designed to handle.

Figure 5 shows the essential features of the group modulating and group demodulating circuits. As in the type K system, group modulation is performed at a very low level of the message material and with a high level, about 25 milliwatts, of the group carrier supply, in order

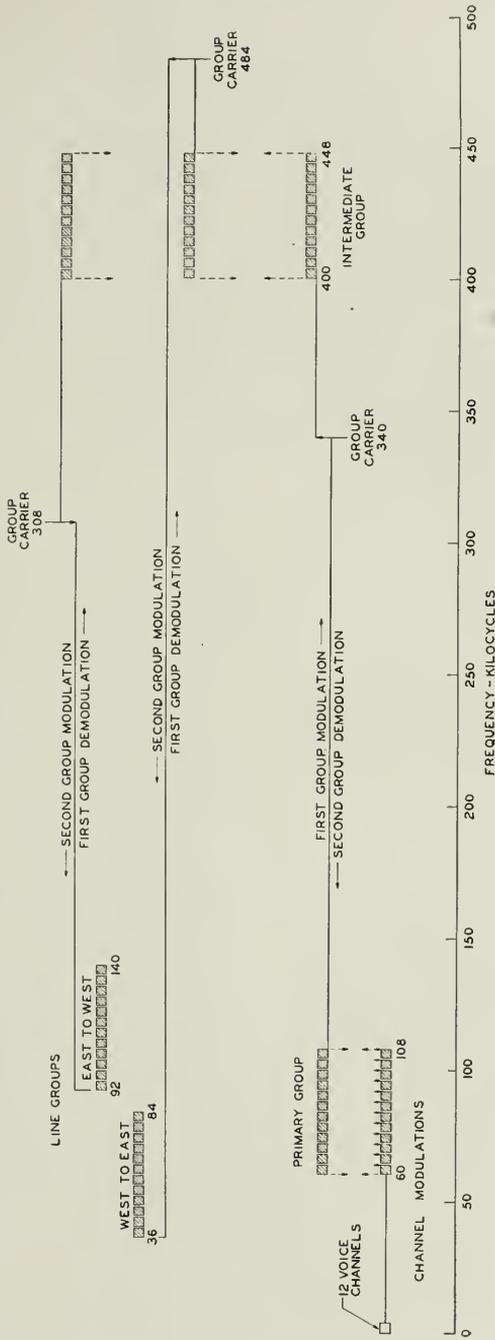


Fig. 4—Frequency translations.

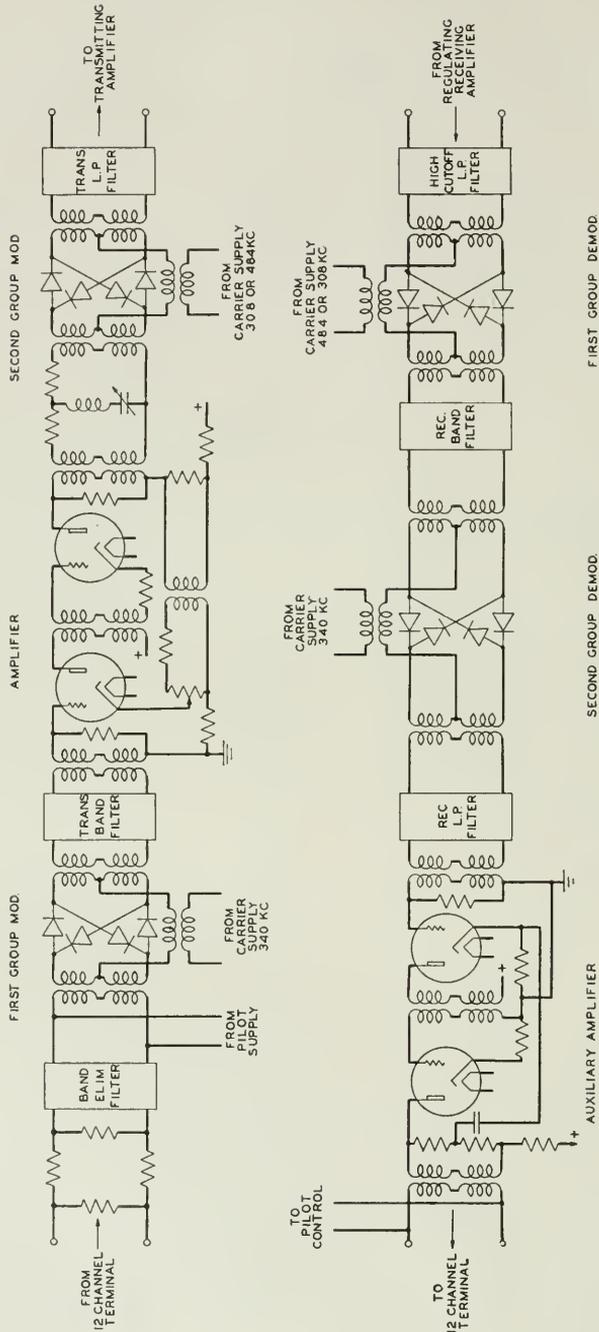


Fig. 5—Group modulator and demodulator.

to minimize interchannel crosstalk. The group modulators are of the doubly balanced bridge type which aids in suppressing some of the unwanted modulation products. Following the first group modulator and also following the first group demodulator are coil and condenser type 400–448 kc band filters which reject the unwanted products and pass the band of frequencies containing the twelve channels. Between this filter and the second group modulator on the transmitting side of the terminal, an intermediate amplifier is used in order to keep the level of the group transmission above danger of noise. Following the second group modulator and also following the second group demodulator are low-pass filters which cut off frequencies above about 160 kilocycles, to suppress unwanted modulation products. From the output of the receiving low-pass filter the twelve-channel group, 60–108 kilocycles, passes through a two-stage “auxiliary” amplifier to bring it to the desired level.

The carrier frequencies for group modulation and for group demodulation are derived from the same 4-kc tuning fork controlled oscillator that supplies carriers for the twelve-channel unit. From the circuit in which appear the odd harmonics of 4 kilocycles, the 77th, 85th and 121st harmonics, that is, 308, 340 and 484 kilocycles, are selected by carrier supply filters and separately amplified by two-stage amplifiers to produce the powers required for group modulation. Outputs from these amplifiers are fed to individual frequency busses capable of supplying the group modulators and demodulators for ten systems. An emergency carrier supply for these frequencies is also provided, with arrangements for switching rapidly from the regular to the emergency circuits.

TERMINAL AMPLIFIERS

As indicated on Fig. 5, the transmitted twelve-channel group, now transferred to the proper frequency range for line transmission, goes from the low-pass filter at the output of the second group modulator to a transmitting terminal amplifier which is similar in most essentials to the amplifiers of the line repeaters. The twelve-channel group, arriving from the line, passes through a regulating amplifier arranged and controlled to compensate for variations in equivalent of the adjacent line section before passing to the first group demodulator. Similar regulating amplifiers are used at all repeater points.

FILTERS

At terminals and also at repeater points, two kinds of filter sets are required. One kind is used in the line to separate the type J

frequency range 36 to 140 kilocycles from the type C and other lower frequencies on the line. The second kind is the directional filters of the type J system itself. These separate a twelve-channel band of frequencies lying below 84 kilocycles used for west to east transmission from the twelve-channel group lying above 92 kilocycles which is transmitted from east to west. These directional filter sets are carefully designed to equalize any non-uniformity of loss in both the directional and the line filters. As this equalization involves a considerable loss over a large part of the filter band it is provided entirely in the receiving directional filters where the transmission is at a low level and the loss can readily be made up by amplification. In this way nearly the full energy output of the transmitting or repeater amplifier is available for line transmission.

LINE CROSSTALK PROBLEMS

As noted previously, type J systems will, in general, be applied on pairs on which type C systems are already operating. Such pairs have already been arranged to transmit frequencies up to 30,000 cycles, and transposed in such a manner as to perform satisfactorily as regards crosstalk to and from nearby pairs on which similar carrier systems are operating. In addition, on most modern lines the spacing between wires of a pair has been reduced from twelve to eight inches; and, on many of the lines, in order further to reduce crosstalk by increasing the spacing between pairs, the number of pairs on a cross-arm has been limited to four instead of five, omitting the pole pair. Now, by applying a new transposition system designed for type J operation up to 140,000 cycles, an eight-inch spaced four-crossarm line may be arranged to transmit type J frequencies on at least ten pairs out of sixteen. Type C systems may, of course, be used on all of the pairs. Finally by using the most advanced transposition design methods, and increasing the crossarm spacing, in addition to the features noted above, a new line may be constructed to permit the operation of sixteen channels on all pairs.

To make the pairs of wires good for type J systems, more than a four-fold increase in frequency range, was difficult. The natural tendency of the circuits to crosstalk is increased even more than the frequency ratio, so that in addition to applying a new transposition design it is necessary that the transposition poles be more accurately located, and that the sags of the two wires of each pair be kept more nearly alike. On lines which already have eight-inch spaced wires, no major structural changes are necessary. However, on lines which have only twelve-inch spaced wires and where it is desired to make available a

number of pairs for type J transmission, structural changes, such as respacing the wires of the pairs concerned to six inches, are necessary in order to reduce the coupling.

One factor of extreme importance is that of reflected near-end crosstalk. In the application of transposition systems it is usually not possible to reduce the near-end crosstalk to a magnitude approximating the far-end crosstalk. It is the latter with which the carrier systems are chiefly concerned, since similar types of systems on different pairs all transmit the same frequency range in the same direction. If, however, the lines concerned do not have smooth impedance characteristics, i.e., a high degree of freedom from reflection effects, near-end crosstalk may be converted by reflection into far-end crosstalk of sufficient magnitude to be controlling over the true far-end crosstalk.

This means that lines to be used for several type J systems must be made unusually smooth electrically—impedance variations kept within a few per cent. The achievement of such smoothness consists chiefly in:

- (1) Reducing the electromagnetic and electrostatic couplings to other pairs so that there are no large energy interactions, with corresponding impedance irregularities. Generally speaking, when the pairs concerned have been transposed for reduced far-end crosstalk up to the maximum frequency transmitted, this condition is also satisfied.
- (2) Minimizing the effect of intermediate and terminal cables. This latter problem has caused considerable concern and is responsible for the development of several new techniques in the design and treatment of such cables, where they appear in a long line otherwise consisting chiefly of open wire.

CABLE TREATMENT

As a means of overcoming the reflection and attenuation effects of short pieces of terminal or intermediate cable, loading naturally suggests itself, as applied in type C systems, where the cable pairs involved are commonly equipped with carrier loading coils, spaced at about 700-foot intervals. This compares with the 3000-foot or 6000-foot spacings which are standard for voice-frequency loading. However, loading pairs in existing cables satisfactorily up to 140,000 cycles would mean coils at approximately 200-foot intervals. Because of physical limitations, existing manhole spacings, etc., this is highly impractical. A reasonable solution has, however, been found in the creation of a new form of low-capacitance high-frequency cable,—a disc-insulated unit which has constructional features in common with the coaxial cables and a capacitance of only .025 microfarad per

mile as compared with about .062 microfarad for conventional cable pairs. This permits more practical loading coil spacings. These disc-insulated units are made up as spiral-fours, that is, two pairs (.051" diameter wire) which form the diagonals of a square. When these cables are loaded with small coils at intervals of approximately



Fig. 6—Disc-insulated cable. Sheath diameter 2.3 inches.

600 feet, they present impedance characteristics substantially equivalent to that of an open-wire pair over the desired frequency range. Accordingly, they form a desirable, although somewhat expensive, solution of the problem of intermediate or entrance cables. As shown in Fig. 6, the spiral-four units are bound together in complements of seven or less under a lead cable sheath similar to standard toll cables. It should be noted that the low-capacity disc-insulated loaded cables not only provide a satisfactory solution of the impedance matching

problem, but they also give a cable circuit of low attenuation,—approximately 1.2 db per mile at 140 kilocycles.

Nevertheless, where spare pairs exist in cables, it has often been found economical to use them for type J transmission. It is possible to use them only non-loaded, in which case the attenuation is very high—4 to 6 db per mile, depending on the gauge, at 140 kilocycles, and impedance matching transformers are, of course, required at the junction of the open wire and cable. There are cases where this higher attenuation may be permitted and these pairs are used by separating the type J range from the lower frequency range, which is transmitted through pairs equipped with the older type C carrier loading. The separation is accomplished by filters which are usually housed in small filter huts at the junction of the open-wire line and cable.

In other cases it has been found economical to use the frequency separation method with filters and to install new non-loaded cables of lower attenuation to lead in the type J frequency band alone. Paper insulated 10-gauge pairs or the disc-insulated spiral-four cable of the type described above may be used for this purpose. In either case transformers are used to match the cable impedance to that of the open-wire line over the type J frequency range.

The reflection requirements are so severe and the effects of even short lengths of cable at the high frequencies so serious, that even short lead-in cables, where the open-wire line actually extends to the repeater or terminal building,—cables which are only 100 or 200 feet long, must receive special treatment. This has also been accomplished by the use of the disc-insulated spiral-four cables, loaded.

INTERACTION CROSSTALK

Because of the higher attenuation there will be many repeater points on a long line at which the type J system will be amplified but at which the other systems and wires on the line will pass through the station without amplification. In this case, even though the type J pairs are properly transposed to keep down crosstalk between themselves, there still remains the crosstalk between them and the other pairs on the line, not only pair-to-pair crosstalk but crosstalk from the type J pair to various circuit paths consisting of irregular wire combinations.

Two difficulties arise in this case: The first is that the crosstalk from the output of one J system into an irregular path may be retransferred into the input of a repeater on another type J system. The second is that the crosstalk from the irregular path may be returned to the input of the same repeater and either influence the overall transmis-

sion characteristic or, if sufficiently severe, actually cause the repeater to sing. This general situation has made it necessary to introduce in the circuits at such points "crosstalk suppression" filters in the non-J pairs and longitudinal choke coils in all pairs.

STAGGERING

In addition to the various steps which are taken in order to reduce crosstalk by improving the line conditions, the type J system may include a feature which has been used in the type C system,—the staggering of the channel bands used on neighboring pairs. The advantage of staggering results from the facts that (a) the sensitivity of the ear and the power of the voice vary over the audible range, (b) the efficiencies of transmitter and receiver also tend to vary over the frequency range, (c) part of a channel band may lie opposite "dead" frequency range on an adjacent pair, and (d) by controlling the arrangement of the sidebands the crosstalk may be made unintelligible even if not inaudible. The staggering feature is readily provided in the type J system by a suitable choice of carrier frequency for the second group modulator and first group demodulator. With the staggered systems the highest frequency used would be about 143 kilocycles.

ATTENUATION PROBLEM

In what has preceded in the discussion of line problems, the emphasis has been confined chiefly to the question of the smoothness of a line from an impedance standpoint in order to keep down reflection effects and, correspondingly, to improve the operation from a system-to-system crosstalk standpoint. There is also the problem of the higher attenuation incident to the use of higher frequencies. Between 30,000 cycles and 140,000 cycles, the normal wet weather attenuation for a 165-mil open-wire pair, for example, rises from about 0.13 to 0.28 db per mile,—an increase of approximately 2 : 1. Repeaters on the type J system, if applied on the basis of approximately the same output level and minimum level requirements, must be spaced at about one-half the interval of the type C systems. Normal spacings for type J systems would therefore be expected to range from 75 to perhaps 100 miles where no large amount of intermediate cable existed.

However, another problem, not present to a similar degree at the lower frequencies, tends in many cases to have a controlling effect on this spacing, that is, sleet or ice on the wires. With ice, frost, or snow on the wires, the wet weather attenuation may be exceeded by very large amounts. The additional attenuation is due primarily to the coating on the wires themselves rather than the coating on the

insulators. It arises from the potential gradient through the ice deposit in combination with the high dielectric loss characteristic of the ice or snow coating. Figure 7 gives examples of the attenuation

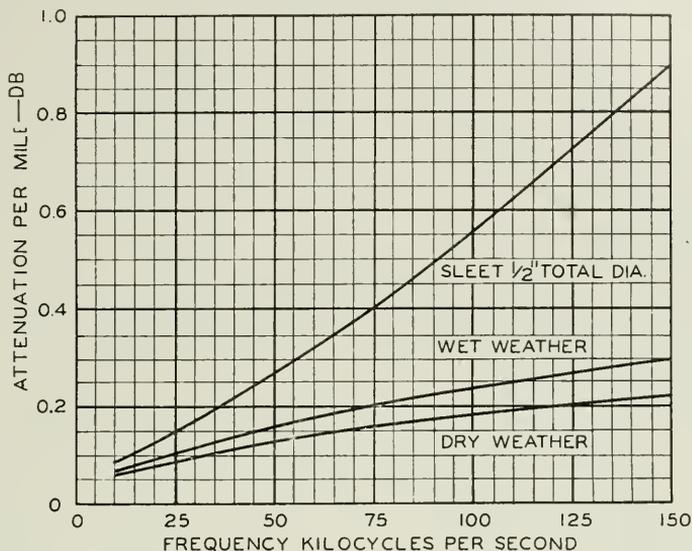


Fig. 7—Attenuation frequency characteristics of open wire lines.

frequency characteristics of open-wire lines, including certain measurements with ice coating. The exact increase in attenuation due to snow and ice naturally depends on the thickness and other characteristics of the coating. Even very thin coatings of ice on the wires tend to raise the attenuation at 140 kc from the normal wet weather figure of about 0.28 db to about 1 db a mile, i.e., an increase of three or four to one. Extremes up to 5 db per mile have been measured for short lengths of line with ice nearly two inches in diameter. Such heavy ice obviously approaches the mechanical breakdown conditions for the line.

Where ice and sleet occur the repeater spacings may be reduced to about fifty miles or less. The repeaters now being provided for the type J systems have gains of approximately 45 db. Repeaters are under development which are expected to raise the maximum available gain to something like 75 db. The normal dry or wet weather operation of such repeaters would be limited to gains of perhaps 10 to 25 db depending upon the amounts of cable included. The problem of obtaining automatic gain control over the extra wide range required by the high sleet attenuations is a difficult one.

REPEATERS

At each repeater point line filters and directional filters are required on both sides of the amplifying equipment to separate type J currents from those of lower-frequency services on the line and to separate oppositely directed groups for separate amplification in one-way line amplifiers. These filters have been described in connection with the terminals where they perform similar functions. Two regulating amplifiers, one for each direction of transmission, properly controlled to compensate for variations in the attenuation of the preceding line section, are also needed at each repeater point. These are described later under "Regulation."

Figure 8 shows the circuit of one of the line repeaters and indicates the location of the directional filters, and certain supplementary filters for suppressing frequencies outside the transmitted range; also the regulating amplifier circuit, and the pick-off of the pilot channel which controls the gain.

The line amplifier has three stages of pentodes. The first two stages use single tubes of high voltage amplification and low power capacity while the third stage has four power pentodes in parallel to increase the output capacity. Because of considerable heat developed by these power tubes, special precautions are necessary to dissipate the heat and to protect condensers and other elements mounted near them.

Negative feedback to improve the operation of the amplifier is supplied over two paths. The inner feedback, from the plates of the output tubes over a properly designed network to the grid of the input tube, reduces the gain at frequencies outside the transmitted band and so prevents singing at those frequencies. It has little effect at frequencies within the type J range. The outer feedback path includes the input and output transformers, which are made as hybrid coils. In each of these one pair of the conjugate windings is connected to the incoming or outgoing circuit of the amplifier while the other pair is used for the feedback connection. By feeding back through the transformers in this way, they benefit by feedback in much the same way as the tubes, and the overall characteristic of the amplifier is practically independent of the transformer characteristics. This feedback reduces the amplifier gain by over 40 db and correspondingly reduces modulation effects within the amplifier, and gives exceptionally stable transmission with respect to tube and voltage changes. It is also designed to improve and stabilize the input and output impedances.

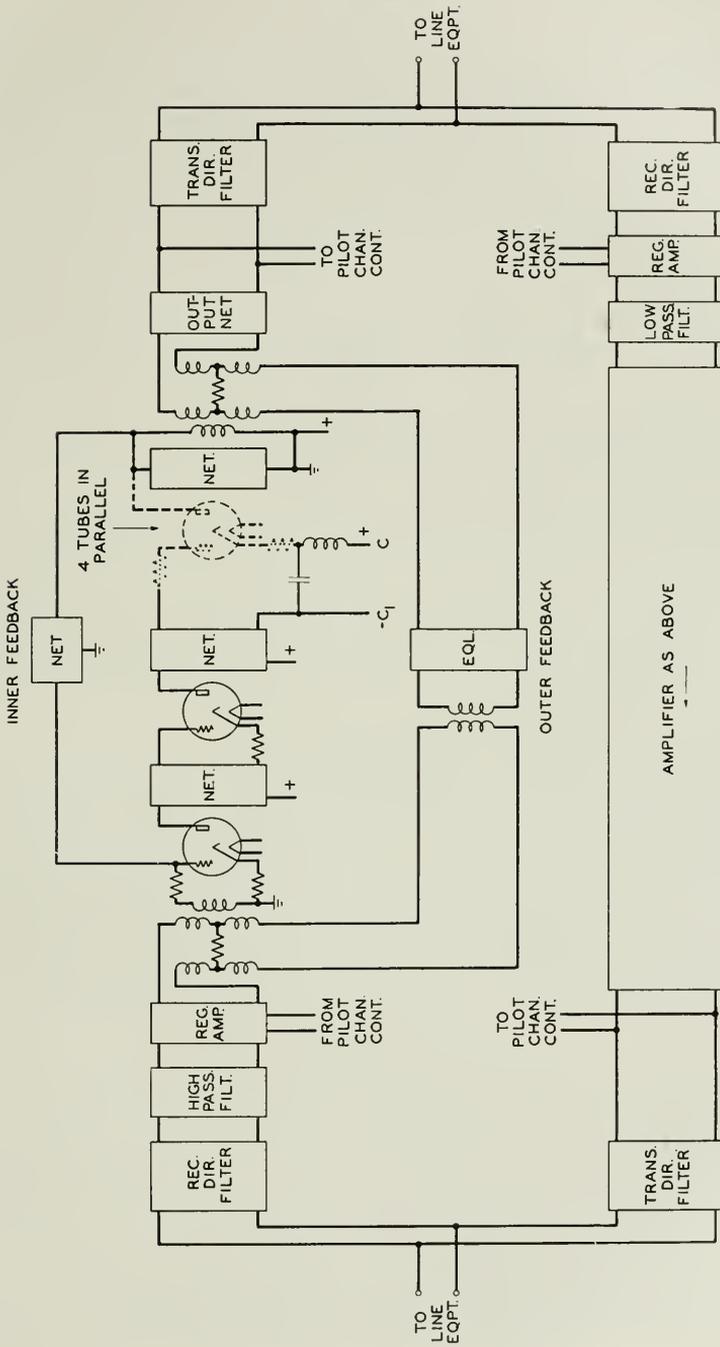


Fig. 8—Line repeater circuit.

EQUALIZATION

Equalization is necessary in each direction of transmission at a repeater point and in the receiving direction at a terminal, to compensate for frequency distortion produced by the preceding section of line. Fortunately, the attenuation frequency curves for the usual open-wire circuits, that is, 104, 128 and 165-mil wire, have nearly the same shapes for section lengths giving the same attenuation at the maximum frequencies for the two directions of transmission, so that these various circuits can be equalized alike.

As is well known, the transmission frequency characteristic of an amplifier with large feedback is almost the inverse of that of the feedback circuit itself, so that the insertion in the feedback circuit of a network having the same characteristics as a line section will provide equalized transmission over the amplifier and section combined. In the outer feedback circuit of the line repeater is included an equalizer which has a characteristic sloping with respect to frequency in the same way as the variation in loss under wet weather conditions of the longest open-wire section likely to be used. Thus, there is provided in the repeater a basic equalization for this longest wet weather line. At a receiving terminal a basic equalizer is provided which performs this same compensation, but in this case the slope of the curve must necessarily be opposite to that of the line attenuation and of the equalizer in the feedback path of the line repeater.

Line sections, however, vary in length and in the amount of entrance cable included. In order that they may be properly corrected by this basic equalization, they must be built out to equal this longest wet weather section. For this purpose there are provided flat loss pads and building-out networks whose losses have the same frequency shapes as the losses of short lengths of open-wire circuit. These pads and networks can be inserted or omitted by simple changes in strapping. They suffice to build out the shortest section which is expected to be used.

PILOT CURRENTS

For a satisfactory system, arrangements must be provided to correct automatically for the effects on line attenuation due to changes in weather, by adjusting the amplification at each repeater point and in the receiving terminal circuit. To permit measuring these effects a pilot current of fixed frequency, near the middle of the transmitted band, and of constant amplitude, is supplied from each terminal. This is applied to the transmitting side of the terminal circuit between the twelve-channel terminal and the first group modulator, where the

message band lies between 60 and 108 kilocycles. The pilot frequency is 84.1 kilocycles which is obtained by modulation of 88 kilocycles, from one of the output taps of the channel supply of that frequency, with 3.9 kilocycles derived from a tuning fork oscillator. This modulation is performed in a copper-oxide bridge similar to the channel modulators and the desired product is selected by an 84-kc. carrier supply filter. The output of 84.1 kilocycles is sufficient to supply pilot current for ten terminals in the office. A sharply selective crystal band elimination filter is inserted between the output of the twelve-channel terminal and the point where the pilot source is bridged on the circuit to eliminate any current near the pilot frequency which would interfere with the small pilot current that is sent out to control the system.

The two group modulation processes alter this pilot frequency of 84.1 kilocycles so that it appears on the line as 59.9 kilocycles in the west to east directional band, and as 116.1 kilocycles in the east to west band. Correction in accordance with the magnitudes of these mid-group currents in the two directions is satisfactory over all twelve channels under ordinary conditions. In the case of ice or snow the channels at the edges of the directional frequency groups may not be properly adjusted. Additional pilot frequencies will probably be needed ultimately to care for such unusual conditions.

REGULATING AMPLIFIER

Figure 9 shows the circuit of the regulating amplifier, and above this, the circuit of the pilot channel receiving equipment which controls it. Current enters the regulating amplifier circuit from the left, coming from the receiving directional filter through a shielded transformer and the pads and building-out networks used for equalization. At the terminals the circuit includes also the basic equalizer. Last in the circuit leading from the line to the regulating amplifier is the regulating network which consists of a series of three equal networks having a total loss of 20 db at 140 kilocycles in the east to west direction and 15 db at 84 kilocycles in the west to east direction. The network loss increases with frequency in the same way as the difference between dry and wet weather attenuation of the line. The two terminals of the regulating network and the two junction points between the three networks are brought to four sets of stator plates on an adjustable condenser. The rotor of this condenser, which has about the same area as one set of stator plates, is connected to the grid in the first stage of the regulating amplifier. Rotation of the condenser therefore applies, to the grid of the first tube, a voltage which decreases continuously as the condenser rotates from left to right.

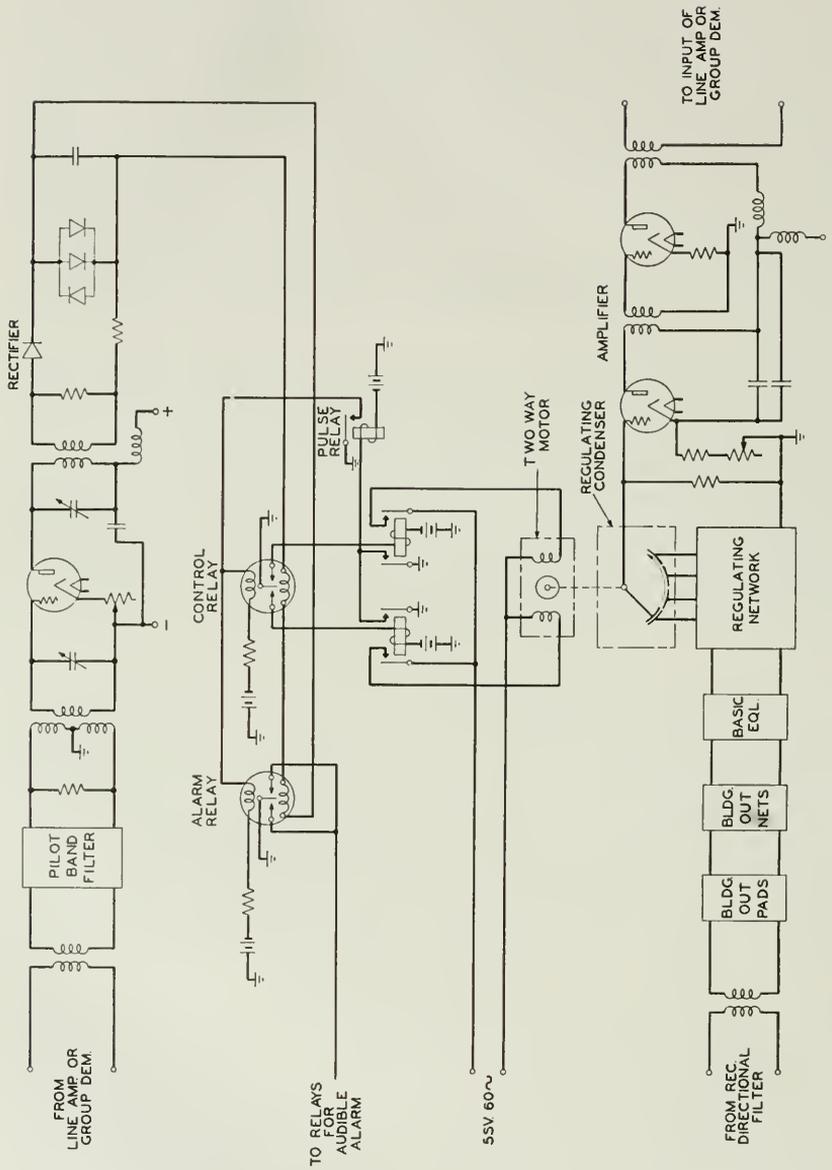


Fig. 9—Regulating amplifier and pilot control.

The regulating amplifier has two stages of pentode tubes, a high input impedance necessary for the proper operation of the condenser potentiometer, and feedback to stabilize the gain and to prevent intermodulation of the channels. Its output goes to the line amplifier at repeater stations, and to the first group demodulator at the terminals. At a west terminal there is interposed a high cut-off filter to eliminate frequencies above the upper band which may have been picked up on the open-wire line.

PILOT CONTROL

The setting of the condenser which controls the regulating network is determined in accordance with the amount of pilot current flowing in the circuit in the direction concerned. At repeater stations the pilot current is picked off at the output of the line amplifier, being separated from the message transmissions by a quartz filter which has about a 30-cycle pass band. For control of transmission from west to east at the repeater stations, this filter selects 59.9 kilocycles and for control of the oppositely directed transmission, 116.1 kilocycles. At the terminals the pilot channel selecting filter is connected across the output of the auxiliary amplifier following the second group demodulator where the pilot frequency is 84.1 kilocycles. The pilot current from the pick-off filter is amplified in a single-stage amplifier which has feedback for constancy of operation and input and output circuits tuned to the pilot frequency. After amplification the pilot current is rectified by a temperature compensated copper-oxide rectifier.

The resulting direct current passes through the operating windings of the control and alarm relays. These Weston Sensitrol relays are, in fact, microammeters with high and low contacts made by the pointers. The mechanical bias of the moving system is adjusted so that with the normal pilot current the pointer will remain free in the middle between the two contacts. A change of about 0.5 db in this current will cause the pointer of the control relay to make contact with the terminal at the corresponding end of its swing. As the limiting contacts are magnetized and the pointer is of magnetic material, good contact is insured. When contact is made on one side a 60-cycle circuit is closed through the motor which controls the regulating condenser in such a direction as to cause the loss in the regulating network to be increased. Closure of the other contact similarly causes the loss in the regulating network to be decreased. Closure of either contact also closes a circuit containing a slow-operate "pulse" relay to release the Sensitrol relays after an interval of about four seconds. During this time the gain of the regulating amplifier will have been

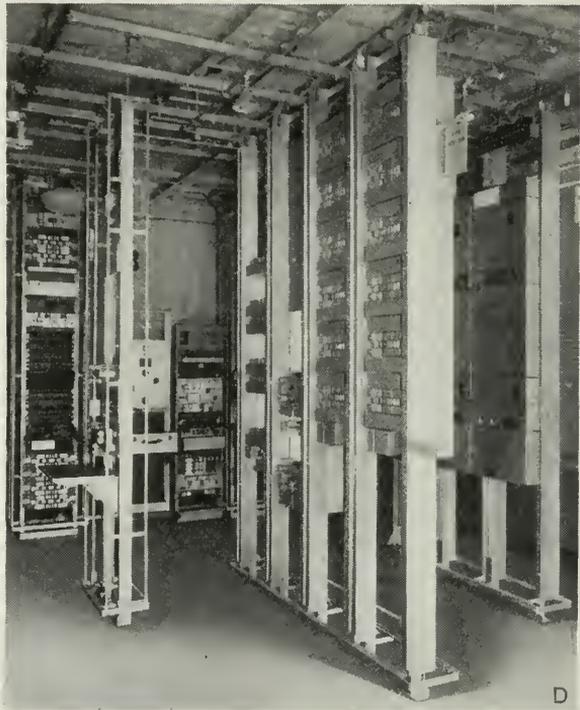
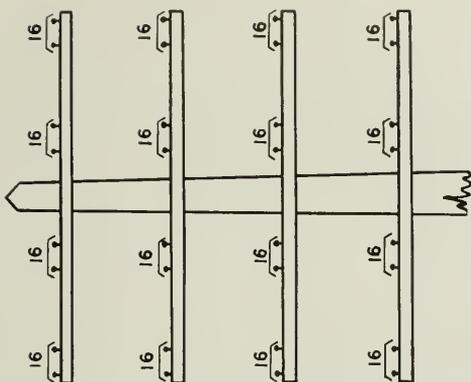
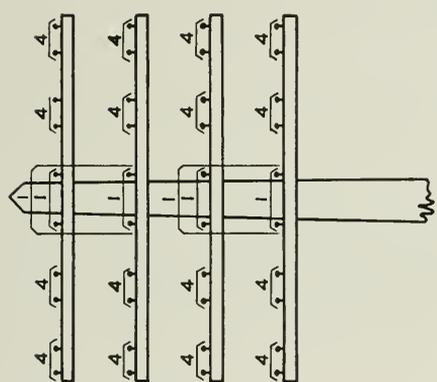


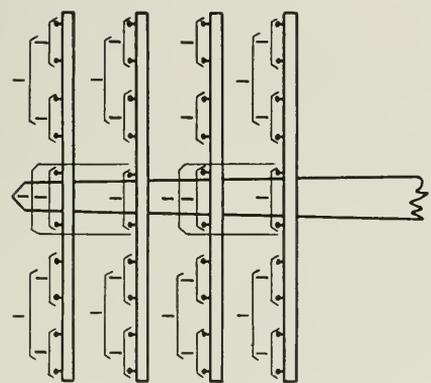
Fig. 10—Typical installations.
(A) Auxiliary repeater station.
(B) Cable hut.
(C) and (D) Terminal installations.



New line construction all 8" spaced pairs, no pole pairs, crossarms spaced 36 inches instead of 24 inches. No phantoms. Facilities—16 voice circuits, 240 carrier circuits, total, 256



Line construction with 8" spaced non-phantomed systems on all 8" spaced pole pairs. Type C facilities—22 voice circuits, 48 carrier circuits, total, 70



Facilities—30 voice circuits.

Fig. 11—Growth in line carrying capacity.

changed about 0.1 db. If now the pilot current level is within 0.5 db of normal the operation is complete. If not, it is repeated and the device keeps periodically testing the circuit so long as it is away from satisfactory compensation. There are also alarm circuits for attracting attention in cases of wide variations in equivalent. In severe ice conditions where a single regulating repeater has not sufficient gain to make up for the great loss in the line, the next succeeding repeater will do its utmost to make up the deficiency.

CONCLUSION

In what has preceded, developments have been described which are making it possible to provide a very substantial increase in circuits on open-wire pole lines without additional wire stringing. Photographs showing typical office installations of type J carrier equipment, unattended repeater stations, and filter huts are shown in Fig. 10.

Three stages in the development of the open-wire line over the past twenty years, giving successive increases in circuit capacity, are shown in Fig. 11. Prior to the application of carrier systems, a four-crossarm pole line would yield thirty voice circuits. Now, on a new line 256 circuits are potentially obtainable. Thus it is probable that the open-wire line will continue as an important factor in furnishing facilities in moderate numbers, particularly in the less densely populated sections of the country and where climatic conditions are not unfavorable. Installations of type J systems have already been made in various parts of the country.

Recent Developments in the Measurement of Telegraph Transmission

By R. B. SHANCK, F. A. COWAN and S. I. CORY

This paper describes the progress which has been made in recent years in the development of methods and apparatus for the measurement of telegraph transmission in the Bell System. Such measurements play an important part in transmission maintenance work in the field and are also necessary in development work. The changes which have occurred in service requirements, particularly the large commercial development of start-stop teletypewriter service and the effect of these changes on the technique of telegraph transmission measurement, are first discussed; then a description is given of several new measuring devices and their use.

IN keeping with advances in the telegraph transmission art, noteworthy improvement has been made in measuring devices and methods in the past few years. The faster, more accurate, and generally more dependable telegraph service now available has been made practicable not only by improvements in the telegraph systems but also by the use of improved measuring apparatus and techniques.

In the early stages of development most transmission-measuring systems were arranged to measure transmission on "looped" circuits, that is, with sending and receiving terminals at the same point, so that a comparison between the sent and received signals could be made. Although such an arrangement is quite useful for laboratory testing, it imposes serious limitations on field testing. Therefore, it is generally desirable to make tests on a straightaway basis.

For straightaway tests it is necessary to have at the receiving end certain information regarding the sent signals. This requires either the use of signals of certain known characteristics or the determination of the important characteristics of the sent signals and the transmission of this knowledge to the receiving end. The latter of these alternatives is not frequently used since it requires another communication channel, although in some instances, where tests on working circuits are desired, it represents the only practicable approach. Straightaway measurements of telegraph transmission have for this reason been generally confined to measurements in which the important characteristics of the transmitted signals were known.

Fortunately, either synchronous or stop-start teletypewriter signals

fall into the last-mentioned classification and the increased use in the Bell System of stop-start teletypewriter transmission has afforded the opportunity of making telegraph transmission tests on working circuits without the need of transmitting information regarding the character of the sent signals. Of course, when sectionalized transmission measurements are desired it is necessary for proper interpretation to transmit the results of the measurements to a single point for analysis, but the communication of this information is not at all burdensome.

Aside from the ability to measure on a straightaway basis, which is primarily a field-maintenance requirement, testing equipment should preferably be direct-reading without the need of measuring adjustments on the part of the tester. Direct-reading devices in general effect considerable reduction in the time required for measurement. This feature is especially important in the field where a rapid test of possible trouble conditions is desirable and in the laboratory where the large numbers of tests which are necessary for thoroughly checking a telegraph transmission system under all of the likely operating conditions become even at best tedious and time-consuming.

The major development in telegraph transmission testing within the past few years has been the provision of instrumentalities which possess the desirable properties indicated above. They permit the rapid and direct reading of signal distortion on working teletypewriter circuits. The same instrumentalities when used with selected or miscellaneous teletypewriter test signals also permit the rapid determination of the capabilities of a telegraph circuit in the field or in the laboratory.

A paper published in 1927¹ discussed fundamental concepts relating to signal distortion and described a number of measuring devices which had been employed in the Bell System. Another paper² treated the design of telegraph circuits for distortionless transmission from the standpoint of the steady-state characteristics. The fundamental ideas set forth in these papers have continued to form the basis for development of the technique of measuring telegraph transmission. However, it has been necessary better to adapt these ideas to start-stop teletypewriter operation and changing field requirements.

Operation of telegraph circuits by means of start-stop teletypewriters^{3, 4} using 7.4-unit code has become of much greater importance in the Bell System in the past dozen years. The majority of private-line service is now furnished on a teletypewriter basis; also teletypewriter exchange (TWX) service,⁵ inaugurated in 1931, has become an

¹ References are listed at end of paper.

important factor, having already grown to the point where the trunk-circuit mileage employed is a large part of the total telegraph mileage. Incidentally, there has been at the same time a general increase in operating speeds, so that the majority of the circuits now operate at a nominal speed of 60 words per minute (23 dots per second or 46 bauds).

As regards the requirements for measuring apparatus for field transmission maintenance, the desired precision and convenience have increased considerably in the last few years. This is due to several causes, chief of which are the continuing desire to give better service with greater freedom from interruptions and isolated errors, increase in speeds of operation, and the use of more complicated circuit layouts with more sections in tandem, particularly in Press and TWX service. For complicated circuits it is very advantageous to employ maintenance procedures in which each section is measured and adjusted separately to close limits, to avoid the more costly and otherwise less desirable overall line-up. Furthermore, a need has arisen for accurate transmission measuring devices for other uses such as checking the condition of receiving teletypewriters, transmitting keyboards and regenerative repeaters,⁴ and use in "equalizing" of telegraph circuits, that is, the application of wave-shaping arrangements for reducing distortion. Finally, in line with improvements in main-line circuits greater emphasis has been placed on maintaining loops and circuits to outlying points so that they introduce but little distortion.

ADAPTATION OF MEASURING TECHNIQUE TO TELETYPewriter BASIS

The earlier types of measuring sets were arranged to measure the total change in the duration of signal pulses, that is, the combination of the displacements at the beginning and end of any given pulse. This method of measuring gives results which are directly indicative of the impairment for Morse operation since the interpretation of the signals depends on the total duration of pulses. This method also gives a moderately good indication of the effect of distortion for teletypewriter operation.

In start-stop teletypewriter operation there are two ways in which circuit imperfections may cause the transmission to be impaired. In the first place, imperfections other than constant delay (known as line lag), which may be neglected, may cause the start transition of any character to be displaced with respect to the time at which it should occur. This causes the starting of the receiving mechanism to be advanced or retarded and effectively displaces the succeeding transitions of the character. Secondly, other imperfections may also cause any of the succeeding transitions to be displaced in either direction. The combination of these two effects determines the effective distortion.

It is of interest to consider the case of bias alone. With uniform bias the displacement of mark-to-space transitions is in one direction and that of the space-to-mark transitions is in the other direction with respect to their positions in undistorted signals. Since the start transition is mark-to-space the result is that the effective displacement of subsequent mark-to-space transitions is zero and the effective displacement of space-to-mark transitions is numerically equal to the bias. In practice bias is seldom uniform and may vary with the signal combinations. In these cases there is an effective displacement of mark-to-space as well as space-to-mark transitions.

From the foregoing, it will be seen that it is of considerable practical value to be able to measure teletypewriter circuits on a start-stop basis in terms of displacement of transitions with respect to the start transition. (For a more complete explanation of the effect of distortion on teletypewriter operation, reference should be made to published discussions.^{4, 5, 7}) In testing with miscellaneous signals, bias may for convenience be taken as the average effective displacement of space-to-mark transitions relative to mark-to-space transitions; characteristic distortion will have the appearance of a combination of fortuitous and bias effects; and the maximum total distortion will be the sum of the average effect and the variation therefrom which causes the greatest displacement.

OTHER CHANGES IN MEASURING TECHNIQUE

In measuring with normal and inverted signals¹ on circuits of the types commonly employed, the result obtained for the bias varies somewhat from pulse to pulse of the test signal. A case in which this variation is appreciable is that of carrier telegraph having level compensators (automatic devices which correct for changes in the magnitude of the received current). With these compensators, the response is fairly rapid, the result being that the bias is to some extent a function of the signal combinations of the transmitted material. This bias variation is also noticeable with open-and-close d-c. telegraph circuits having large bridged capacitance or series inductance.

On account of this bias variation, it is desirable to measure the algebraic average of distortion of the individual pulses of miscellaneous signals and take this as the bias. This may be conveniently done by measuring on the start-stop basis mentioned above. In such measurements the differences between the distortions of the individual pulses and the average distortion may be considered as due to the combination of characteristic and fortuitous effects; further measurement is necessary in order to separate these effects. A measure of character-

istic distortion may be obtained by determining the difference between the systematic distortion measured with selected recurring signals and the average distortion with miscellaneous signals. Fortuitous distortions may be taken as the difference between the total and systematic distortions obtained with recurring signals. In order to distinguish between the variable bias and the true characteristic distortion in the case of level-compensated circuits measurements may be made with the compensator disabled and with it functioning.

For tests in the field where it is desired to measure systematic distortion effects, as for instance in connection with equalizing, several simple signals corresponding to certain especially selected teletypewriter characters are employed. In each of these signal combinations there are only two transitions; therefore it is convenient to observe the effect of the remnants of one transition upon the next transition. As discussed in the Appendix, the characteristic distortion obtained for miscellaneous signals is a function of the distortion obtained with the simplest characters; if there were no distortion on the simple characters no characteristic distortion effects would be expected when miscellaneous signals were transmitted. The process of equalization, therefore, consists in adjusting the transmission characteristics of the line circuit to reduce the characteristic distortion measured on six special characters to a minimum. The teletypewriter characters which are used for this purpose are Blank, *T*, *O*, *M*, *V* and Letters; the corresponding signals are shown in Fig. 1. The distortions of these signals are observed at the receiving end on a measuring set operating on the start-stop principle or a portable systematic-distortion measuring set having an integrating meter, as will be described more fully below.

In equalization testing, each of the six signals is sent repeatedly for the time required to determine the total systematic distortion—generally about 50 repetitions. In analyzing the results the bias component is assumed to be substantially the same for all of these signals and whatever difference is observed from one signal to another is taken as being due to characteristic distortion. It is found generally that the result for the *O* signals, which are practically unbiased six-cycle reversals (see Fig. 1), is not radically different from the result obtained for bias with reversals at 6 or 23 d.p.s. (shown in the lower part of Fig. 1), the results being expressed, of course, in the same terms, as for instance in per cent of a 23-cycle dot. The largest distortion is usually found on either the Blank or Letters character, this being reasonable because usually the remnants of transients practically disappear within a few dot lengths. Sometimes it is

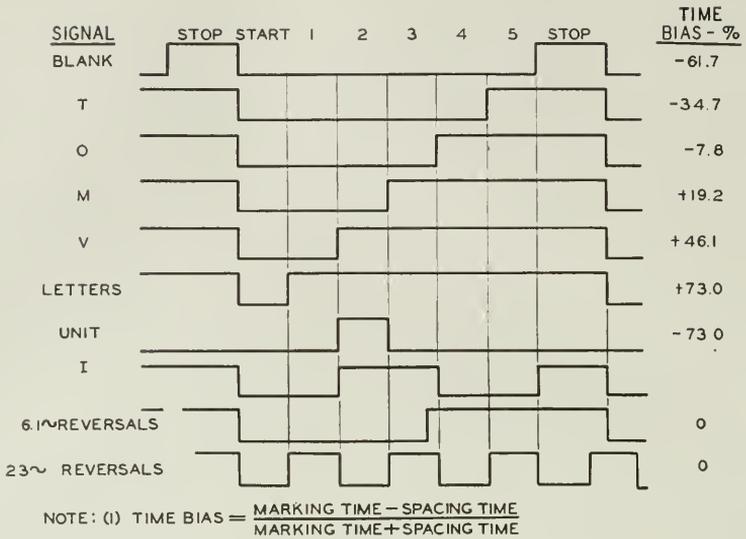


Fig. 1—Signals used in measuring systematic distortion. The first six signals and the eighth signal are teletypewriter characters.

desirable to extend the test to include biased test signals, as described below.

As an example, curves are given in Fig. 2 showing the results of tests with the six characters on a d-c. metallic telegraph⁶ circuit operating on 112 miles (180 km.) of composited 19-gauge cable pair. Curves 1 and 2 show the results before and after equalization respectively; it will be noted that considerable improvement was effected. On the basis of both experiment and theory, the slope of Curve 1 is known to indicate that the received direct current is larger than it

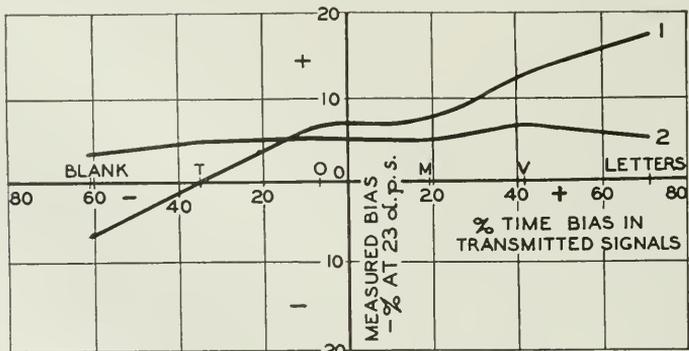


Fig. 2—Equalization test result for a 112-mile section of metallic telegraph; signalling speed 23 d.p.s. Curve 1, before equalization; curve 2, after equalization.

should be as compared to the higher-frequency components of the received current waves. The required type of equalizer in this case is one which adds loss at frequencies in the vicinity of zero, without a corresponding loss at the higher frequencies. If, however, the slope had been in the opposite direction, an equalizer which would discriminate against the higher frequencies would have been required. As a check of the equalizer setting, the total distortion and bias are usually measured using miscellaneous teletypewriter signals.

In addition to measuring with undistorted signals applied at the sending end of a circuit, the measuring technique has been expanded to include measuring with distorted signals and a device has been made available for field use by means of which reversals or teletypewriter characters may be distorted by known amounts. This kind of test furnishes additional information in that it affords an examination of the effect of signal combinations which are not included in perfect telegraph signals. It is of value because in actual operation a given telegraph section may not have perfect signals impressed at the sending end due to distortion occurring in previous sections or at the transmitter. Although such a test furnishes valuable information for line testing, it has been used in the field up to the present mainly in testing the distortion-tolerance of subscriber-station teletypewriters with signals from the adjacent central office and in maintaining regenerative repeaters.

It is necessary, of course, to make transmission measurements on the manual Morse circuits which still constitute a considerable part of the total mileage. Testing such circuits by the same methods as used for teletypewriter circuits has been found to give good results. However, due to improvement of telegraph circuits, the transmission-maintenance problem in this case consists mainly of keeping the bias within reasonable limits for which purpose simple tests with reversals can be used.

NEW MEASURING DEVICES

In the following is given a description of a number of testing methods and arrangements which have been found useful in recent years both in the field and in development work.

A. Start-Stop Distortion-Measuring Set for Central-Office Use

A start-stop type of measuring set for testing teletypewriter circuits has been developed and is now used generally in maintenance work and special testing in the field and in laboratory work. This represents an outstanding advance in that it provides a quick and convenient means for reading, directly from conventional-type milliameters as

illustrated by Fig. 3, the distortion of miscellaneous teletypewriter signals on working circuits. It has the distinct advantage of giving immediate indication of the occasional isolated peaks as well as the average distortion. In using this set, it is unnecessary to have a knowledge of the transmitted text.

This set employs the condenser-charging principle in the measurement of small time intervals corresponding to the distortion of the

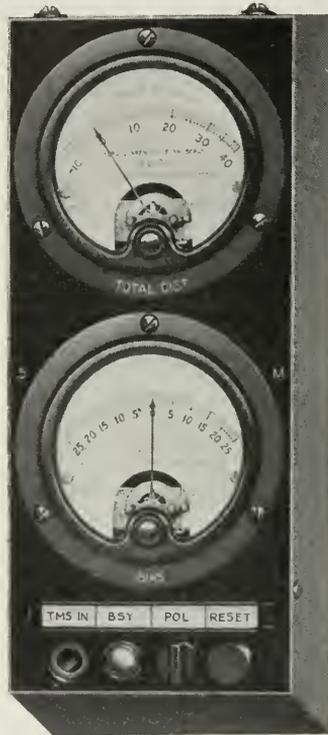


Fig. 3—Meters and control apparatus at telegraph board.

signals. The outstanding feature of this device is that the beginnings of the condenser-charging intervals are timed with relation to the start transition by a start-stop distributor which forms an integral part of the set. The charging intervals are terminated by the occurrence of transitions in the characters, at which times the operation of a receiving relay causes the condenser voltages to be compared to a reference voltage. The circuit is arranged to charge the condenser at a constant rate; hence the voltage attained is determined by the duration of the charging interval. In this way the displacements of the transitions

in the received teletypewriter characters from their proper positions are measured in terms of condenser voltages. Indications are afforded of the average distortion and the peak value of the total distortion, the latter being the sum of the bias, characteristic and fortuitous effects.

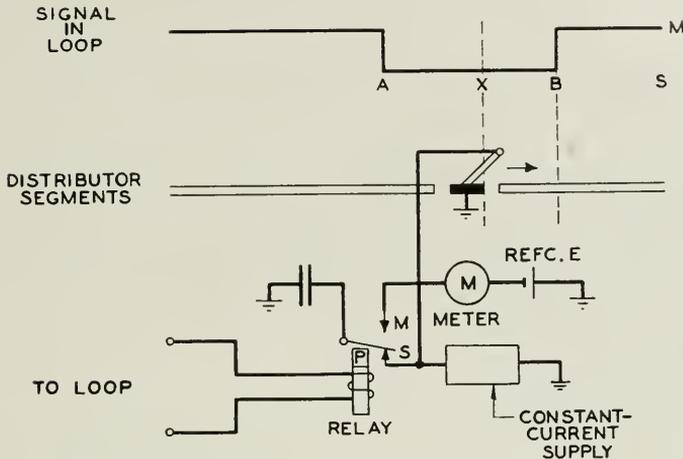


Fig. 4—Explanatory sketch of start-stop telegraph transmission measuring set.

The general features of this set will be described in the following. In Fig. 4 a condenser-charging circuit is shown with a distributor for the purpose of timing the charging intervals in the measurement of distortion occurring at transition *B*. Assume the distributor brush to be traveling in the direction of the arrow after being released by the mark-to-space transition at *A* by means of arrangements not shown. This same transition causes the relay armature to move to the spacing contact (*S*) and the condenser to begin to charge from the constant current supply, but as soon as the brush touches the grounded segment the condenser is discharged completely. After leaving the grounded segment the brush travels over an open segment and during this time the condenser accumulates a charge. At transition *B*, the armature of the relay moves to its marking contact (*M*) and the voltage of the condenser is compared with the reference voltage (REFC. *E*) which has been previously adjusted to such a value that if there is no distortion the condenser voltage and the reference voltage will be equal. However, if transition *B* does not occur at the proper time, because of distortion, the condenser voltage will differ from the reference voltage and a momentary current will flow through the indicating meter (*M*) in proportion to the amount of distortion.

Two condenser-charging circuits are provided, one for space-to-mark transitions and the other for mark-to-space transitions as is indicated in Fig. 5. Graph *A* of this figure shows an undistorted teletypewriter

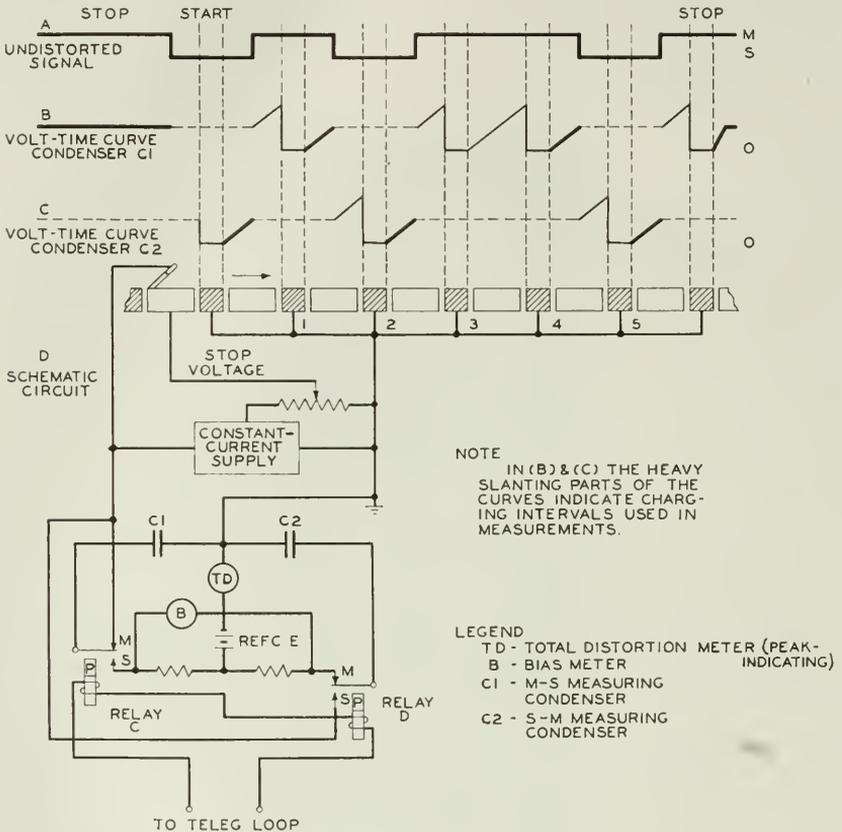


Fig. 5—Simplified diagram of start-stop telegraph transmission-measuring circuit.

character having a stop pulse, start pulse, and five selecting pulses. The segments of the distributor, laid out to show their relation to the received signal, are indicated at the top of the schematic circuit (*D*). It will be noted that there are seven short grounded segments, one for each of the pulses of the character, for initiating the condenser-charging intervals referred to above.

Assume the distributor brush to be at rest during the stop interval as shown. The measuring condenser *C1* is now charged to the reference voltage. As the brush leaves the stop position due to the mecha-

nism controlled by the start pulse (not shown), it travels in the direction of the arrow and the condenser charges vary as indicated by graphs *B* and *C*. It will be seen on graph *B* that condenser *C1* is charged during intervals between grounded segments and mark-to-space transitions. Graph *C* shows that condenser *C2* is charged between the grounded segments and space-to-mark transitions. If there is no transition while the brush is traveling between two adjacent grounded segments the condenser continues to be charged until the brush touches the second grounded segment at which time it is completely discharged. Therefore, the useful charging interval is that between a grounded segment and a transition occurring before the next grounded segment is traversed by the brush. These intervals are indicated by the heavy-lined portions of the graphs. They amount to 37.5 per cent of a unit pulse as a maximum, i.e., the maximum distortion which can be measured is about 37.5 per cent. The currents flowing as a result of the comparison of the condenser voltages with the reference voltage are indicated on a "Total-Distortion Meter" *TD* which is, in reality, a peak-indicating voltmeter, and on a "Bias Meter," *B* which is sufficiently sluggish to give an indication corresponding to the average distortion. These meters are calibrated to indicate directly the percentage distortion with miscellaneous teletypewriter characters.

Good accuracy is obtained with these sets; when measuring distortions of small or moderate amounts with a well-adjusted set, the indication is accurate to within about 2 per cent distortion at 60 words per minute. For occasional large distortions or for higher speeds, the accuracy is not quite as good, and there are certain possible mutilations of signals, such as the dropping out of pulses, which would not be readily detected.

In addition to measuring miscellaneous teletypewriter characters these sets may be used with recurring test signals in which the spacings of the transitions are such that the maximum characteristic effects will be obtained, and with signals which experience mostly bias and fortuitous effects. In this way a measure of the components of distortion may be obtained. Such tests are commonly made in adjusting variable networks to minimize characteristic distortion, i.e., making the equalization tests, referred to above, with selected teletypewriter signals.

In special testing where it is desired to separate the total distortion into its components, this may be done by measuring the systematic distortion with the first 6 signals of Fig. 1 and then measuring the bias and total distortion using the *I* character of Fig. 1 (which is sub-

stantially the same as unbiased reversals at about 11 d.p.s.). The bias component is simply the bias measured with I signals, the fortuitous component is taken as the difference between the total distortion and the bias of I signals, and the characteristic component is obtained by averaging the results for the 6 selected characters and then selecting the result which shows the maximum departure from the average.

These sets also furnish a convenient means for the measurement of mean-square values of distortion. The current impulses flowing through the total distortion meter are proportional to the distortion and it is practicable to insert in series a specially arranged meter which is calibrated to indicate the mean-square values of the distortion. The circuit used is shown by Fig. 6, the heavy lines showing

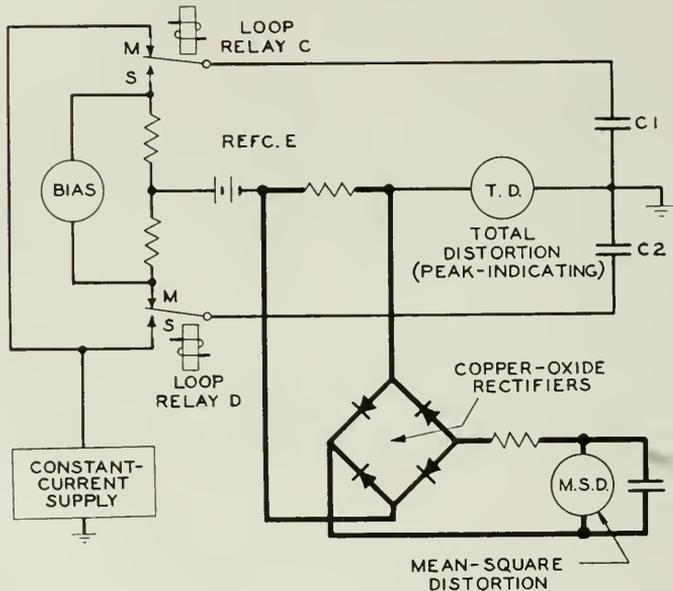


Fig. 6—Circuit of start-stop set for measurement of mean-square distortion.

the meter circuit which is inserted in series. It contains an integrating meter whose characteristic is modified by a full-wave copper-oxide rectifier, a large condenser for increasing the damping and resistances for adjusting the amplitude and response characteristics. Measurements of mean-square distortion have been found of value in connection with producing transmission ratings of telegraph circuits, these ratings being based on the assumption that the mean-square values of distortion of component parts of a circuit may be added directly to predict the total mean-square distortion.⁵

In certain other special tests, where it is desired to obtain a record of the variation in distortion, recording meters have been connected in series with the meters of the set and a continuous record made using either a recurring test message or the signals from the subscriber.

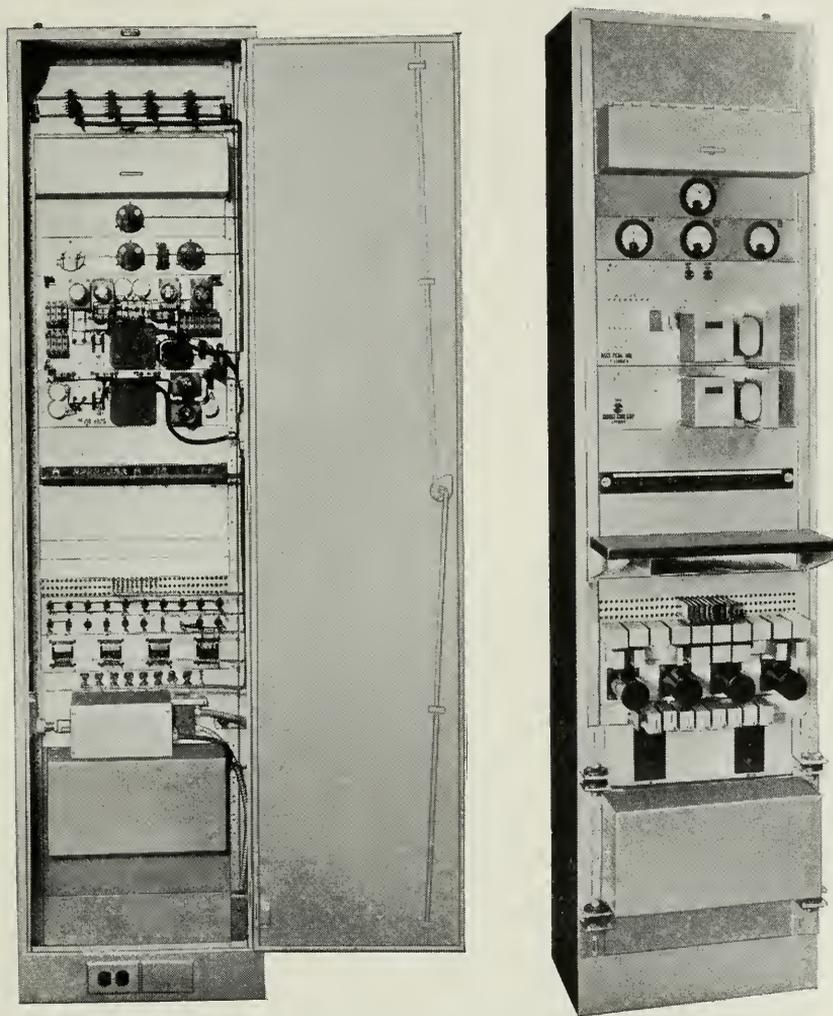


Fig. 7—Start-stop telegraph transmission measuring set.

An idea of the arrangement of the measuring set for central office use may be obtained from the front and rear views which are shown in Fig. 7. It is the practice to provide multiple appearances at telegraph boards in order that one set may be used at any one of a

number of positions. The unit containing the indicating meters and associated controls for mounting at the telegraph board is illustrated in Fig. 3. The set is also provided in portable form for temporary use in cases where a permanent installation is not justified.

B. Telegraph Stability Test Set

Recording meters have been used for a number of years in the measurement of the transmission stability of telegraph circuits.¹ In such a measurement a continuous graphic record is made over as long a period as desired of the variations in the bias and of the number and time of occurrence of fortuitous effects which would impair telegraph service. For this purpose telegraph reversals are impressed at the sending end and a recorder at the receiving end makes a record of the bias of these reversals. This type of test was found to be of such utility in field work that standard stability test sets were produced for this purpose.

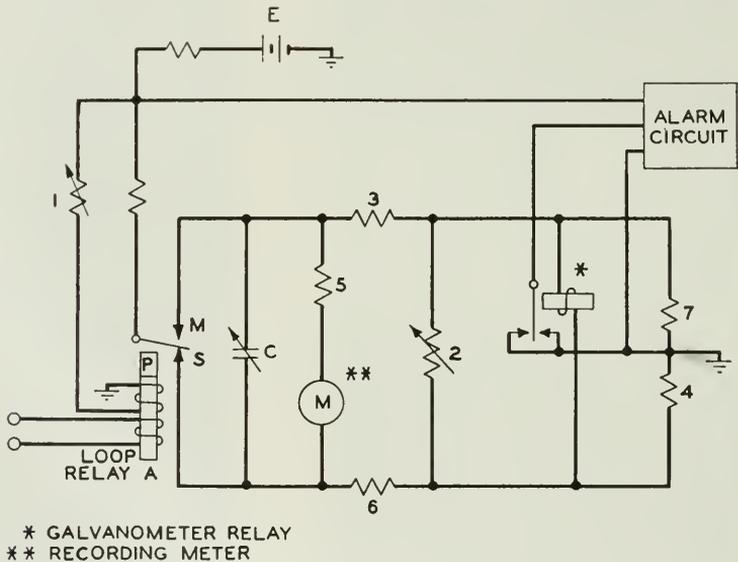


Fig. 8—Schematic circuit of telegraph stability test set.

The circuit of the telegraph stability test set is indicated in Fig. 8. This is essentially a simple bias-measuring circuit combined with an alarm circuit to indicate when given values of bias are exceeded. Movement of the armature of loop relay *A* from *M* to *S* contact or vice versa in response to the received signals causes a battery *E* to be connected first to one arm (resistances 6 and 4) and then to the other

arm (resistances 3 and 7) of a bridge type of circuit containing recording meter *M*. The two arms of the bridge are balanced and the meter, being bridged across them, receives positive and negative current pulses of equal magnitude in response to the armature movements. If these pulses are of equal duration, as for telegraph dots or reversals having zero bias, the average meter current will be zero. Biased reversals will cause the meter current to average at other than zero by an amount directly proportional to the percentage bias. A center-zero recording meter is used and this provides a running record of the variation in bias. A damping condenser *C* is used to reduce the width of the trace and the amount of unsteadiness of the indication due to fortuitous effects.

The alarm circuit contains a galvanometer-relay bridged across equal resistances 4 and 7. The needle of the galvanometer-relay moves to one contact or the other when excessive values of bias are experienced, the sensitivity being adjusted for response to different values of bias by means of adjustable resistance 2. The response is made somewhat sluggish to avoid alarms being given for interruptions of short duration which are not of interest in connection with an investigation of bias stability.

A sample chart obtained by means of one of these sets is shown in Fig. 9. This chart shows slow variations in bias in the upper part and in the lower part the change in the indication due to dropping out or adding a single dot and momentary failures. Such charts do not, of course, show the characteristic distortion, since this is not present in the case of unbiased reversals.

These sets are now generally used in the field in routine checks on telegraph circuits and in special checks on circuits which have developed faults in service. Usually these checks are made with the idea of locating the cause of hits or swings which it is difficult to locate otherwise; in some cases sets are used simultaneously at several repeater points to sectionalize trouble. Charts are run for long periods, sometimes for several weeks in such tests. The stability test sets are also used to obtain data for transmission ratings of circuits, in which case it is desired to know the extent of the bias variations over long periods and the number and frequency of occurrence of hits.

Figure 10 shows a view of the portable arrangement of the set including the recording meter. The alarm buzzer and the receiving relay are located on the panel along with a row of keys for adjusting the alarm for operation on given values of bias. Jacks are provided on the side of the box for connection to circuits, batteries and the

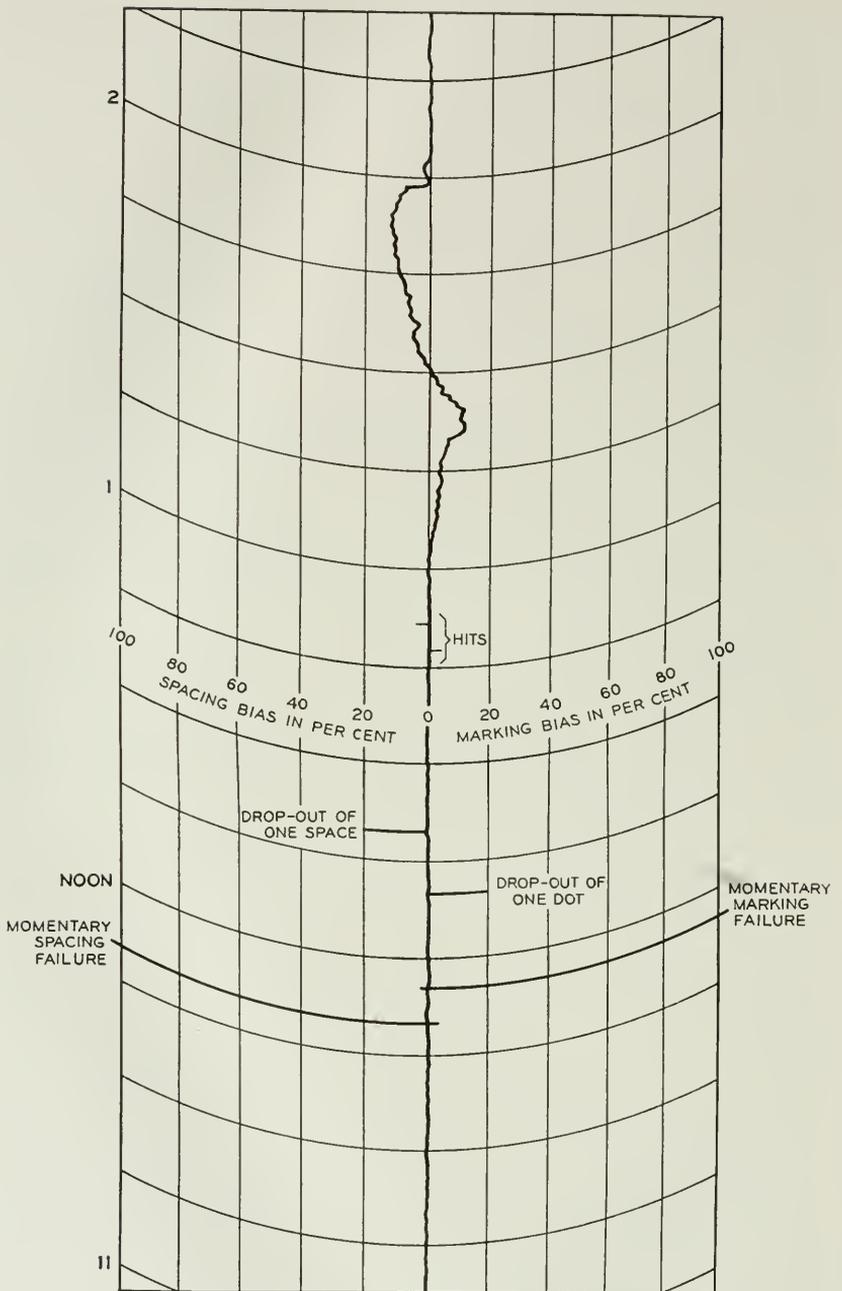


Fig. 9—Telegraph stability test—sample chart.

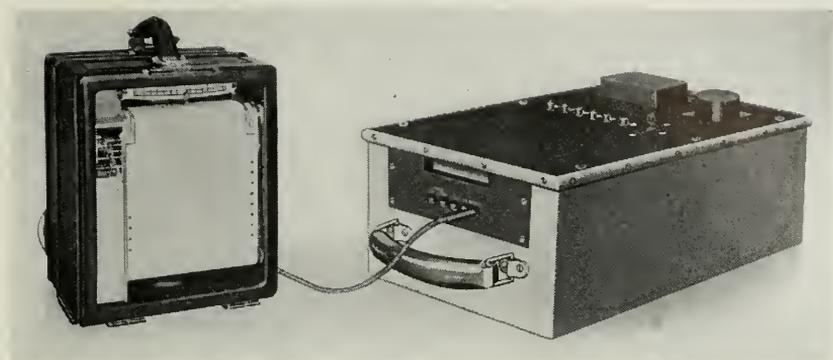


Fig. 10—Stability test set with recording meter.

recording meter. This set is also arranged for permanent mounting on a relay rack in a central office.

C. Portable Measuring Set for Use at Outlying Stations

A new measuring set has been made available for commercial use especially at outlying stations in either routine or special testing. This set is arranged for the accurate measurement of the systematic components (bias and characteristic) of distortion of recurring test signals. In addition measurements may be made of peak values of interference and of the effect of bias and variations in operating currents on the distortion. Such measurements are desirable in analyzing the causes of transmission troubles and in equalization work. By properly interpreting the results of these measurements a fair idea may be obtained of the maximum total distortion. In addition the set is arranged for the convenient measurement of the operating currents and voltages in various parts of the subscriber's circuit and in external circuits. This set may be used on either 110-volt a-c. or d-c. commercial power supply. It is mounted in an aluminum case, and weighs only about 28 lbs. (13 kg.).

The circuit employed for the measurement of systematic distortion is indicated schematically by Fig. 11. This is a simple bridge type of circuit similar to that generally used in a measurement of bias with reversals but especially arranged to indicate directly the percentage systematic distortion of the signals of Fig. 1 excepting "I" signal. The meter circuit is highly damped to prevent undesirable vibration of the meter needle in response to the 6-cycle fundamental frequency of the 60-speed teletypewriter characters. This entails the use of high resistances R , and large capacitance C . Since the voltage in the power supply is limited, a very sensitive meter M is required.

To obtain a zero reading on the meter for each undistorted test character the ratio of the marking to spacing currents is made inversely proportional to the ratio of the marking to spacing time intervals. This is accomplished by means of taps on a potentiometer as indicated in Fig. 11. With the connection made to the middle or Reversals

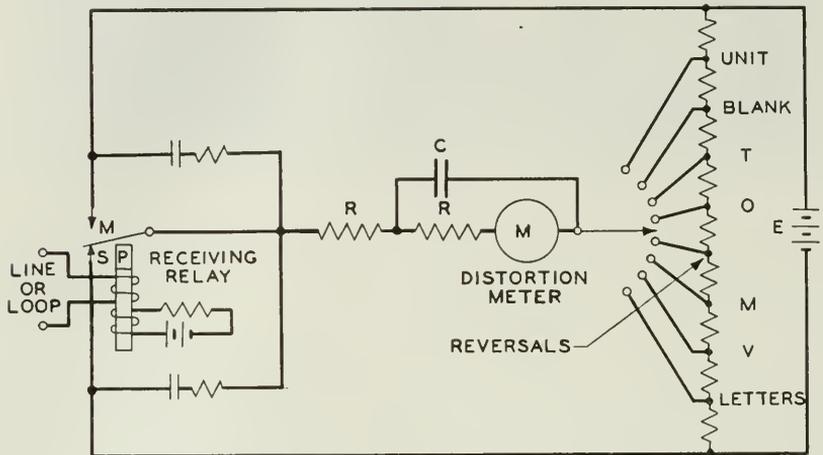


Fig. 11—Outlying-station test set. Schematic circuit for measurement of systematic distortion.

tap, equal and opposite currents flow through the meter when the relay armature rests first on its marking and then on its spacing contact. If the relay is repeating undistorted reversals, the time bias in the signals is zero and the average meter indication will be at zero. With a given undistorted recurring character such as Blank, and with the potentiometer set at the Blank tap, the meter indication will again average at zero. The meter circuit is arranged so that the systematic distortion is indicated directly in percentage based on the duration of a unit signal element of a teletypewriter character.

The circuit used in the measurement of interference is indicated by Fig. 12. The interfering effect is measured by noting on the meter *M* the biasing current which will just prevent the armature of the receiving relay from responding to the interfering currents. Movement of the armature from its contact is indicated by a response in a telephone receiver connected in the armature circuit when the switch is operated as indicated. The biasing current variation is effected by means of potentiometer *P* and the biasing current may be reversed by means of a switch (not shown).

The circuit of Fig. 12 may also be used to give an indication as to

the amount of transmission degradation to be expected due to the interference measured and to given changes in operating currents. In this case the switch is operated to connect meter $M1$ in circuit for the purpose of measuring bias. The effect on the bias of changes in the biasing current for a given operating condition may then be

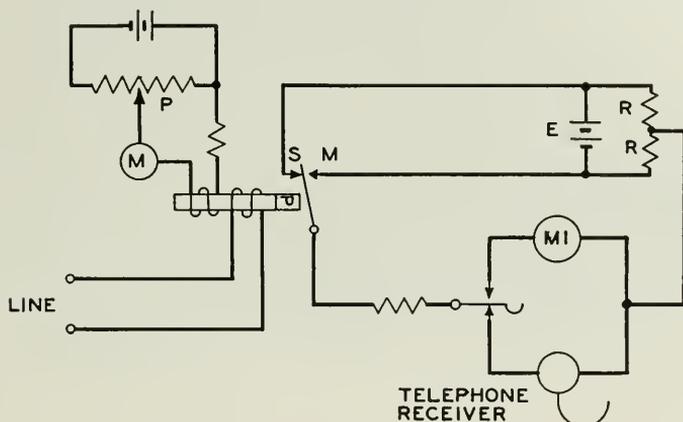


Fig. 12—Outlying-station test set. Schematic circuit for interference measurement.

noted on this meter when receiving reversals or any of the selected teletypewriter characters. The biasing current is varied by means of potentiometer P as before.

This set has several interesting operating features. By inspection of Fig. 13, which shows a view of the set, it will be seen that it contains its own receiving relay located in the lower left corner. This relay may be connected in series in a line circuit or in the local circuit of the subscriber set using the line jacks 1 and 2 (located in the upper right-hand corner of the set) or if desired convenient connection may be made by means of the special plug and adapter shown in the lower part of the figure. In the latter case the subscriber set relay is transferred to the measuring set and the special plug, and adapter if required, is inserted into the relay connecting block of the subscriber set in place of the relay. By operating the proper keys on the test set the currents in different circuits of the subscriber set may then be measured conveniently. Distortion may also be measured with this connection but in most cases it is desirable to measure with the receiving relay in its normal position in the subscriber set to obtain representative conditions.

Because of the advantages of the set, namely, the accuracy possible in the measurement of distortion, the convenience afforded in the



Fig. 13—Outlying-station test set.

measurement of operating and interfering currents, and the portability of the set, it is expected to be of considerable benefit in telegraph transmission maintenance work at outlying subscriber stations.

D. Distortion-Distribution Recorder

A short test to determine the peak value of the distortion existing on a circuit at any particular time is of great value but in development work it is sometimes desired to know the frequency of occurrence of different values of distortion; in other words, to obtain data from which a distribution curve of distortion may be plotted. Such distribution curves were referred to in the earlier paper and it was stated there that, in general, the distribution of distortion is in accordance with the normal law. To check this on different types of circuits in connection with work on transmission ratings of telegraph circuits and for other laboratory uses a "distortion-distribution recorder" has been devised.

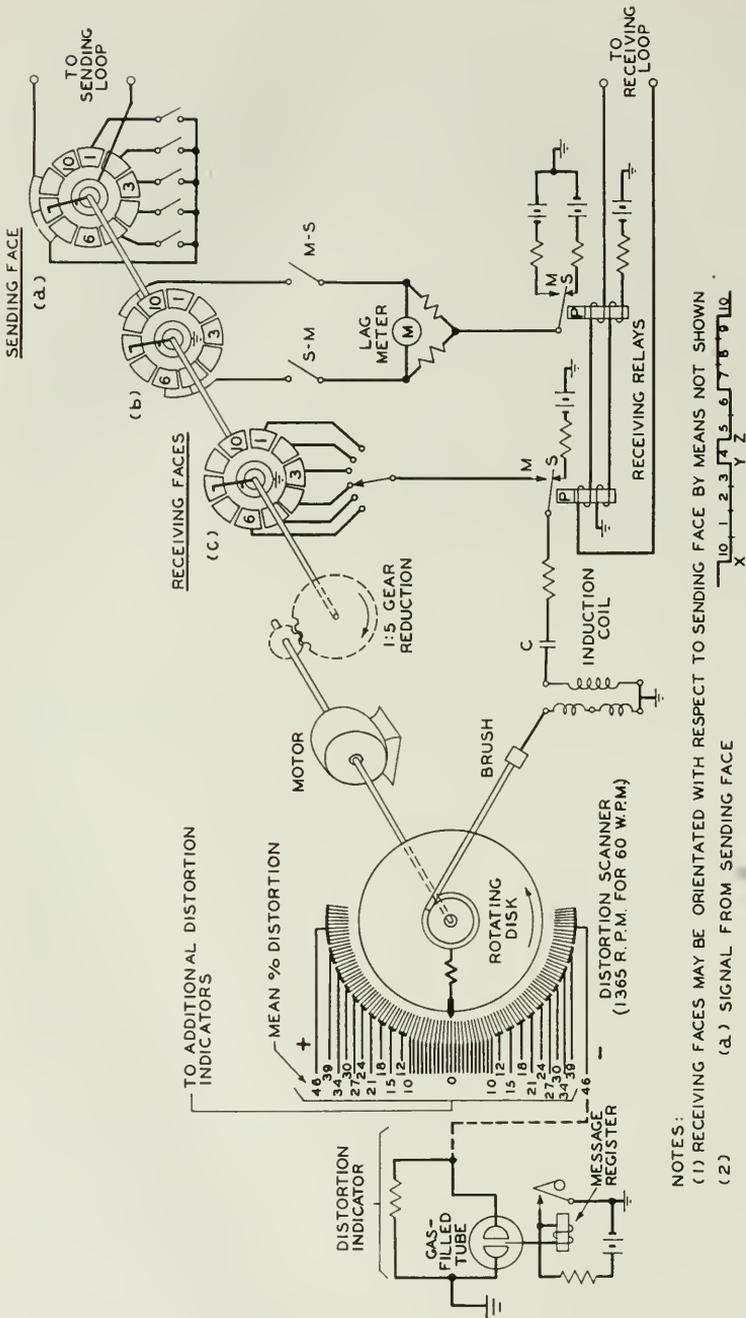
The distortion-distribution recorder operates on a synchronous basis and is suitable only for testing over looped-back circuits. It may be used at any speed up to about 75 words per minute (28.5 dots per second, 57 bauds). It indicates the distortion of the transition at either end of a signal pulse, the indication being in terms of the displacement of the transition from its correct time of occurrence in the

signal combination. Records of the distortion are made on message registers in ranges of one per cent for small distortions and greater ranges for larger distortions.

As shown in Fig. 14, this device contains a sending distributor face (*a*) to provide test signals and two receiving distributor faces (*b*) and (*c*). Coupled to these and running five times as fast is a large disc carrying a fine point from which a spark may be made to jump to any one of a series of stationary segments. This is referred to as "distortion-scanner" in the figure. Since the disc makes one-half revolution while the sending distributor brush is traversing one segment, one-half revolution of the disc requires the same time as one dot and is, therefore, equal to 100 per cent distortion. One hundred segments are provided as indicated, each being equivalent to 1 per cent distortion, so that this ring of segments forms a distortion scale covering the range of ± 50 per cent distortion. Distortion indicators are associated with these segments, each containing a gas-filled tube and a message register. Only the first ten segments on either side of zero are provided with individual indicating arrangements, the succeeding segments being combined in successively larger groups as shown in the figure; this affords adequate information for the usual case.

Assume the switches associated with the sending distributor face (*a*) to be operated to send the signal shown in note 2*a*. After traversing the circuit to be tested this signal operates the receiving relays, one of which is associated with a "lag-meter" and the other with a spark-producing circuit. If the time of occurrence of transition *X* (note 2-*a*) is to be used as a reference, the *M-S* switch of the lag circuit is closed and the sending segments oriented until transition *X* occurs while the brush of receiving distributor (*b*) is midway between segments 9 and 10, as is indicated by a lag-meter reading of zero. The positions of the segments of the receiving faces (*b*) and (*c*) with respect to the signal will now be as indicated in notes 2-*b* and 2-*c*, and the set is ready to measure the occurrence of a transition between any two segments of the sending face, for instance, transition *Y* or *Z*.

If for instance the displacement of transition *Y* with respect to transition *X* is to be measured, the switch associated with face (*c*) is set to connect to segment 4. When transition *Y* occurs the receiving relays operate to marking, condenser *C* is discharged through the primary of the induction coil and a spark jumps from the scanning point to a stationary segment. If transition *Y* is not distorted the spark will jump when the scanning point is opposite segment *O*. This will cause the gas tube associated with segment *O* to fire and



- NOTES:
- (1) RECEIVING FACES MAY BE ORIENTATED WITH RESPECT TO SENDING FACE BY MEANS NOT SHOWN
 - (2)
- (a.) SIGNAL FROM SENDING FACE
- (b) POSITION OF SEGMENTS OF REC. FACE (b)
- (c) POSITION OF SEGMENTS OF REC. FACE (c)

Fig. 14—Schematic circuit of distortion-distribution recorder.

operate the message register which in turn will extinguish the tube by momentarily short-circuiting the anode potential. Other transitions in the signal will not cause the condenser *C* to be discharged through the coil because the brush of face (*c*) will not be traversing the proper segment and there will be no path to ground.

The complete circuit of the device contains a selecting switch which automatically changes the signal combination for each revolution of the brush arm of face (*a*) and reversing switches to invert the signals and the relay connections. These have been omitted for the sake of brevity and clarity. It is thought, however, that the above description will give a good idea of how this device operates to make a record of the frequency of occurrence of distortions. A sample of such a record is plotted in Fig. 15, this being from data obtained over a

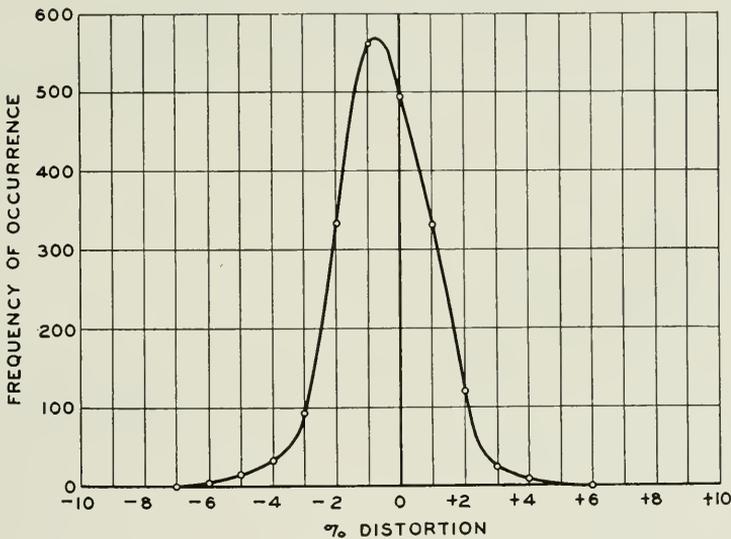


Fig. 15—Typical curve obtained with distortion-distribution recorder.

short-wave radio link. The measurements covered a 15-minute period and one transition of a repeated signal was observed continuously during this period. By inspection it is seen that this circuit has bias of -1 per cent and the r.m.s. value of distortion is about ± 3 per cent. With miscellaneous signals and with characteristic distortion present, the curve would have several peaks and would be somewhat irregular, but would have the general shape of a normal-law distribution curve.

E. Sources of Test Signals

Bell System telegraph circuits are tested at present with both substantially perfect and distorted test signals. The quality of the test signals is, of course, of prime importance because the measurements are generally made on a straightaway basis and it is not practicable to make correction for accidental distortion in the test signals at the sending end of the circuit. On high-grade circuits where the distortion is generally less than about 5 per cent, distortion in excess of a few per cent in the test signals is very undesirable.

UNDISTORTED TEST SIGNALS

Substantially perfect test signals are usually supplied from motor-driven commutators. One type supplies telegraph reversals. In this case an accurately governed motor drives two brush arms, each of which is associated with two rings of segments so that four sources of signals are provided by each machine. These reversals or dot signals are used in a number of ways, their principal advantage being that since the average of the marking and spacing intervals is zero, a simple integrating meter circuit may be used for measuring bias.

When it is desired to employ miscellaneous or recurring teletype-writer characters machines known as transmitter-distributors are used. Such a device, arranged to send a standard test sentence, is illustrated in Fig. 16. This consists of a motor-driven commutator with a continuously rotating brush arm and a direct-coupled cam transmitter (on the left) which changes the connections to the segments of the commutator in accordance with the code which is cut on the cams. In another form of this device a tape transmitter⁴ is used; the tape is usually of parchment although metal tapes and wheels drilled with the code combinations have been used.

When a number of sources of undistorted signals is required in a repeater station a device called a "multiple-sender" is used. This employs the distributor of Fig. 16 to operate a number of relays. The transmitting contacts of these relays are connected to jacks at convenient locations in the telegraph test board. The standard test sentence supplied by this device contains desirable signal combinations for testing transmission over lines and also for testing the operation of teletypewriters.

The signals from the commutator face traverse two groups of relay windings, a marking group and a spacing group, the circuit being as indicated in Fig. 17. This circuit effectively provides polar operation of the relays, the transmitter closing the circuit through one group of windings to operate the relays to their marking contacts and through

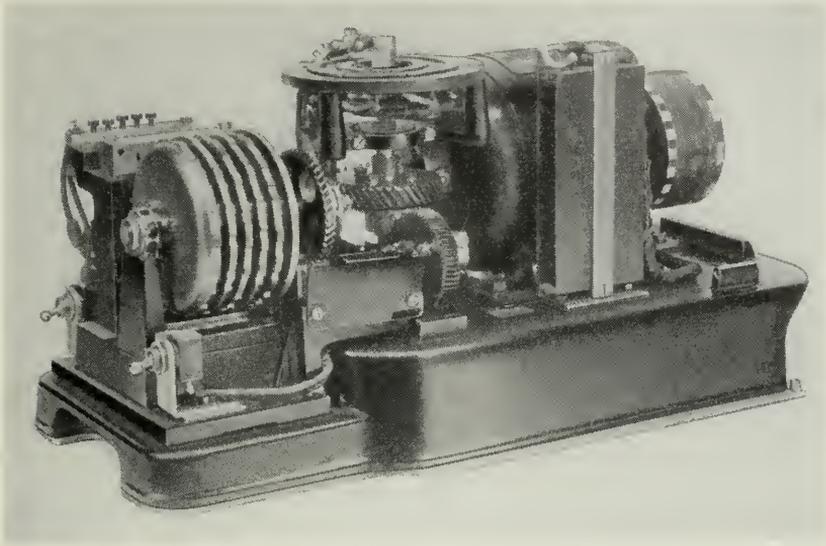


Fig. 16—Test-sentence transmitter-distributor.

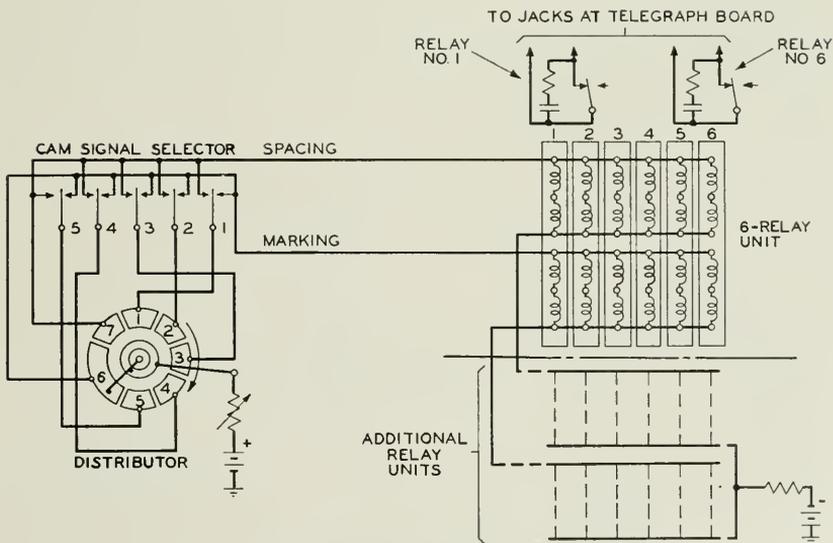


Fig. 17—Schematic circuit of automatic multiple sender.

the other group of windings to operate the relays to their spacing contacts. The series-parallel arrangement of the relay windings shown in the figure has the advantages that with the same number of relays in series and in parallel the combined inductance is only that of a single relay, and that any relay may be removed for inspection without materially affecting the operation of the others. Only one transmitting battery is used in this circuit and thus errors due to battery inequalities are avoided. As indicated in the figure, a spark-reducing circuit is associated with each output for the purpose of minimizing arcing and to neutralize the effect of travel-time of the relay armature which would otherwise cause the transmitted signals to be biased to spacing when opening and closing the circuit under test.

Each group contains 6 relays in parallel and any number of groups up to 8 may be used in series to provide a maximum of 48 outputs to meet the requirements for offices of various sizes.

The above-mentioned sources of signals as maintained in the field are usually accurate to within a few per cent distortion. For special uses it is possible to reduce this inaccuracy somewhat by additional maintenance.

DISTORTED TEST SIGNALS

A repeating device has been provided, primarily for transmission maintenance, by means of which the distortion of telegraph signals may be increased by predetermined amounts. The set is generally used with a source of undistorted signals to provide test signals having known amounts of distortion.

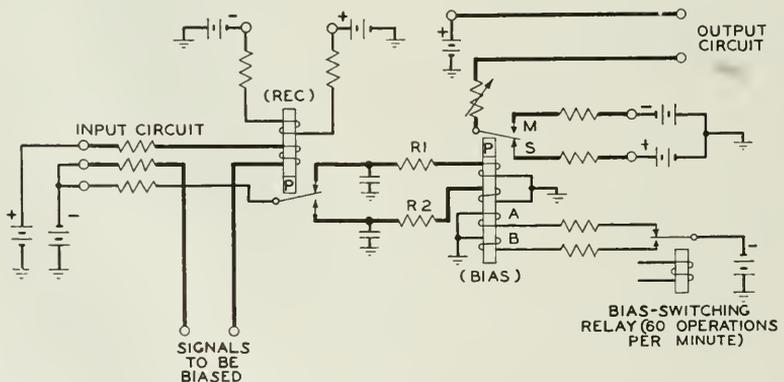


Fig. 18—Bias-producing circuit.

The schematic circuit of the signal-distorting device is shown by Fig. 18. The signals to be distorted are connected to the input and are repeated by the receiving relay (Rec) of the device into a biasing

relay (Bias) through a network which modifies the wave-shape of the signals and permits them to be biased easily by changing the current through the auxiliary windings *A* and *B* of the Bias relay.

The manner in which bias is produced will be understood more fully by reference to Fig. 19 which shows graphs of the currents in the

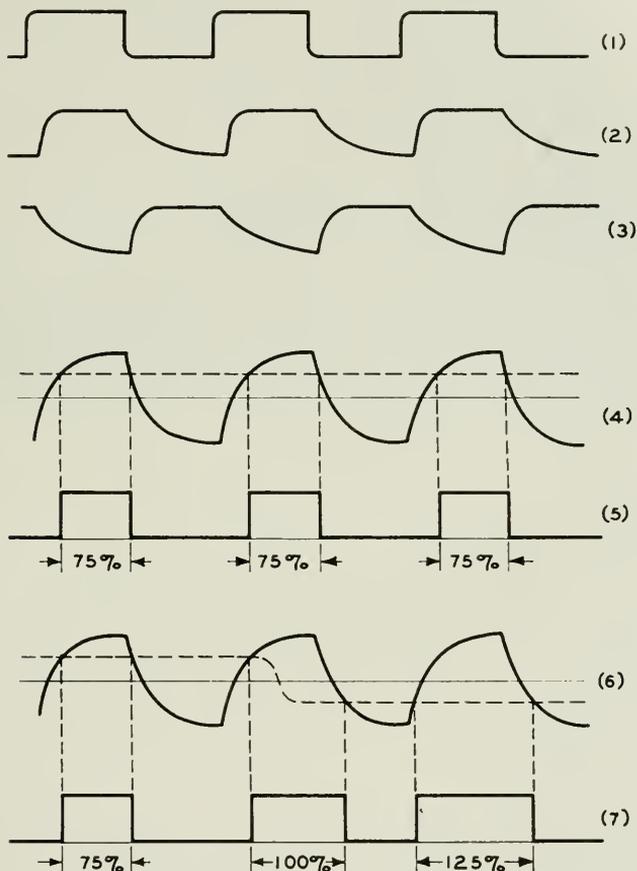


Fig. 19—Currents in bias-producing circuit.

bias-producing circuit. The current in the receiving relay input circuit is indicated by 1. These are undistorted substantially square-topped reversals. The currents flowing through the operating windings of the Bias relay in response to the marks and spaces of the reversals are shown by 2 and 3. In the case shown by 2, it is seen that at the beginnings of the marking pulses the wave fronts are steep and that the current gradually decays to zero at the beginnings of the spacing

pulses. In 3 this condition is reversed so that the net operating current for the Bias relay, being the difference between the currents of 2 and 3, has a rounded wave shape for both the beginnings of marks and the beginnings of spaces as is indicated at 4. This is a symmetrical wave about the zero axis and if there is no bias effect in the Bias relay the signals will be repeated unbiased. However, if current is passed through one of the auxiliary windings of the Bias relay the operating point of this relay will be shifted as indicated by the dotted line and the repeated signals will be biased, as shown by 5. In this case the bias is 25 per cent spacing so that the repeated unit marks are 75 per cent of their original length.

One of the auxiliary windings of the Bias relay is used for introducing marking bias and the other for spacing bias, and the sign of the bias may be changed by switching from one winding to the other. Then the bias effect reverses according to the dotted line of 6, and the signal is affected as is indicated by 7. By reversing the bias periodically under the control of a commutator arrangement an effect known as "switched bias" is produced. The reversing operation occurs 60 times per minute and is not synchronized with the signals and accordingly produces a fortuitous effect on some of the signals.

The signals obtained from this device are commonly used in testing the operating margins of subscriber teletypewriters. In this case perfect teletypewriter signals are applied to the input of the device and are distorted by a predetermined amount in passing through it. These signals are then applied to the circuit extending to the subscriber station. If the receiving teletypewriter at the station responds faithfully when set at its optimum orientation point it is considered to be satisfactory for service.

Distorted signals obtained from this device are also used to determine the extent of the distorting effects in line circuits. For this purpose the set may be connected at either the sending or the receiving end of the line and a distortion-measuring set at the receiving end to give an indication of the increase in distortion caused by predistorting the signals. With the set used at the sending end, the results of the test indicate how much distortion may be applied as from preceding telegraph sections. With distortion added at the receiving end, the test may be used to show the margin in the receiving apparatus before failure.

A front view of a panel-mounted telegraph signal-biasing set of the relay type is shown in Fig. 20.

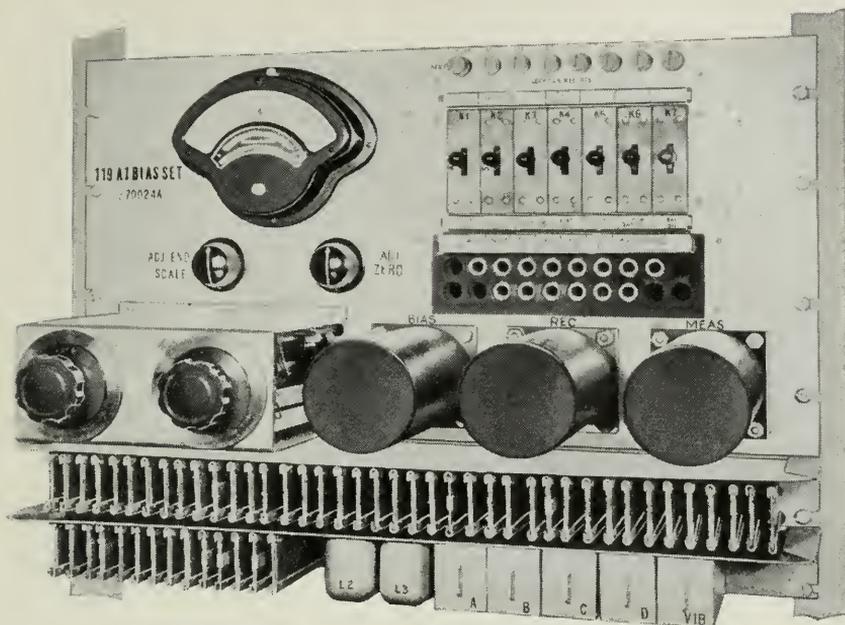


Fig. 20—Telegraph signal-biasing set.

F. Measuring with Teletypewriters

As is well known, the start-stop teletypewriter, when properly adjusted for this purpose, may be used as a transmission-measuring device. The usual procedure in field testing is to compare the range over which the orientation range finder may be shifted with substantially perfect signals to that obtained with signals from the line under test. The orientation range finder is shifted above and below the usual setting until perfect copy is no longer obtained, these limiting positions determining the margins.

At the time of the previous paper,¹ the orientation range test was the only test available to the field forces in the Bell System which utilized the start-stop principle. At that time the teletypewriter was not convenient to use and generally the results were not as accurate as desired. However, the machines have been much improved and better methods of use have been developed so that now orientation margin tests furnish a better indication of the grade of transmission. At the same time, the more convenient measuring sets described above have become available and this has, of course, reduced the field of use of the teletypewriter as a measuring device.

One of the improvements in the machines from the standpoint of

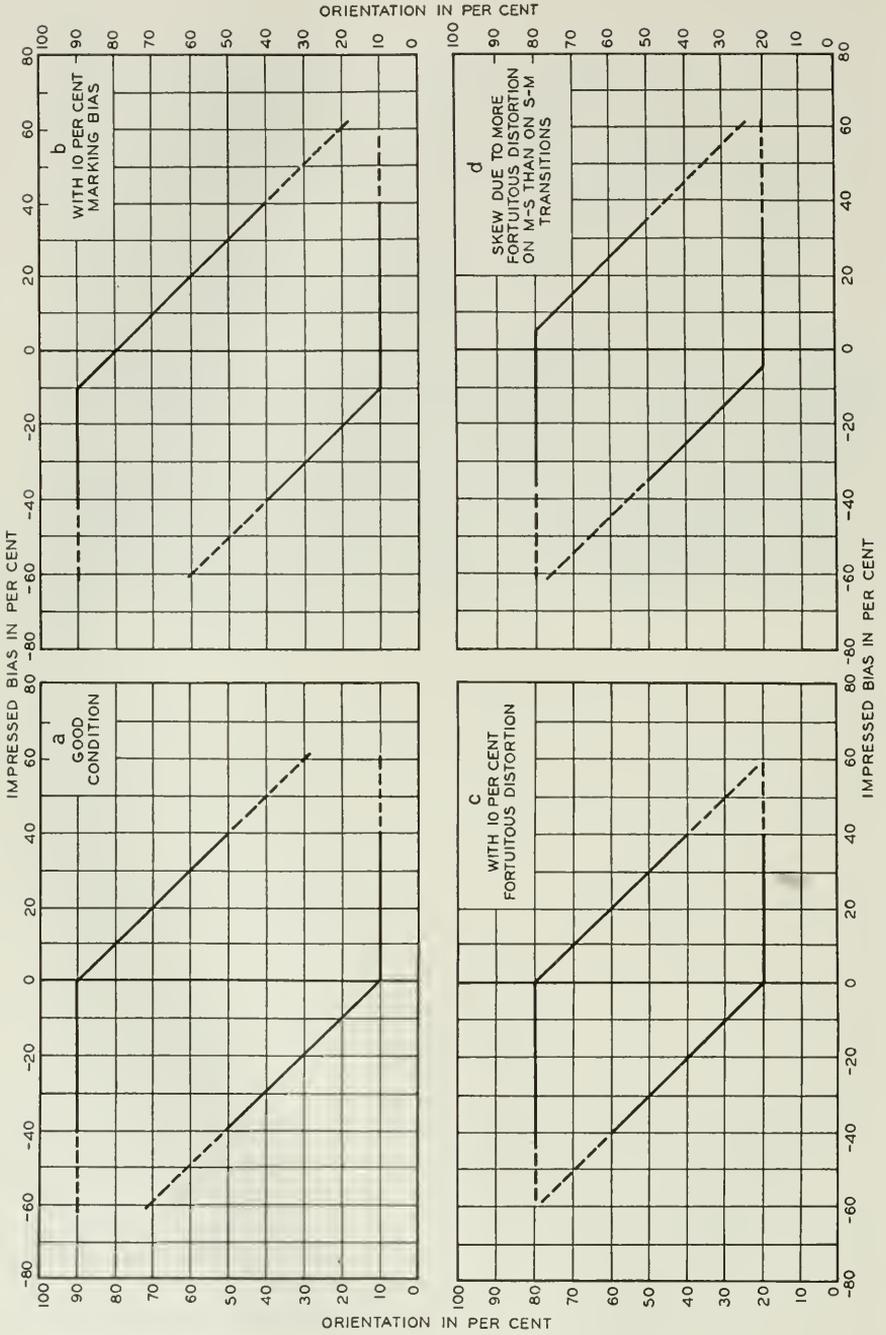


Fig. 21—Diagrams showing the effect of internal distortion on orientation limits of teletypewriters.

transmission testing consists in the addition of a small crank which extends through the cover and which is coupled to the range finder. The crank has a detent which assists in making settings to the nearest per cent, the scale being arranged to indicate directly the distortion in percentage of a unit selecting pulse. This crank and scale arrangement increases the convenience of measurement considerably.

It is of considerable importance to remove as far as practicable the internal distortion⁷ of teletypewriters used for measuring purposes. These internal distortions can be identified as bias, characteristic and fortuitous effects. These effects reduce the margin from the theoretical limit of ± 50 per cent to about ± 40 per cent for the usual machines operating at 60 words per minute.

The presence of internal distortion can be readily determined by using signals biased by various amounts and noting the effect on the orientation range. With a machine satisfactory for measuring purposes, the results obtained would be as indicated on Fig. 21*a*. Here the range is from 10 to 90 per cent for reception of perfect signals. Marking bias reduces the upper range in direct proportion and spacing bias likewise reduces the lower range in direct proportion; thus the machine would be satisfactory for measuring purposes.

If the machine had internal marking bias, the orientation parallelogram would be as shown in Fig. 21*b*. Here the range with perfect signals is from 10 to 80 and marking bias reduces the range in proportion to the bias but spacing bias first increases the range until the internal bias is compensated and then decreases the range as the impressed bias is further increased. With internal spacing bias the parallelogram would be shifted to the other side of the zero line by the amount of the bias. It is obvious therefore that biased machines do not give accurate results in measuring.

Characteristic and fortuitous effects may also be present in teletypewriters to such an extent as to make the machines unsatisfactory for testing purposes. It will be appreciated that internal characteristic distortion changes from signal to signal and when receiving miscellaneous teletypewriter characters, it has much the same effect as fortuitous distortion and the upper and the lower margin limits would be reduced about equally as shown in Fig. 21*c*. If the machine distortions do not have the same effect on mark-to-space and space-to-mark transitions, a skewing effect is produced in the orientation parallelograms. Fig. 21*d* shows skew due to the fortuitous effect of the mark-to-space transitions, being greater than the fortuitous effect of the space-to-mark transitions. Teletypewriters showing such skew effects do not give margin reductions proportional to the

impressed distortion and are, therefore, not suitable for measuring purposes.

It is apparent from the above discussion that it is necessary to adjust teletypewriters for minimum internal distortion before they can be used for measuring purposes. Where it is desired to use teletypewriters in testing, procedures have been established in the Bell System to insure that they will be in proper condition and fairly good results are obtained with them.

Although the effect of distortion on the orientation margins has been discussed previously^{4, 5, 7} it may be of value to state here how distortion affects the margins at the lower and upper orientation limits in connection with the use of teletypewriters for testing purposes. In general, the maximum reduction corresponds numerically to the total distortion as indicated by the start-stop type of measuring set described above. Distortions other than bias usually affect both orientation limits equally so that the amount of bias can be estimated by subtracting the smaller reduction of the two from the larger. In addition it is possible to obtain an idea of the characteristic distortion by the indications obtained during the orientation test. If the orientation limits are fairly definite, that is, if the copy changes from good to bad when the range finder is moved only a small distance, it is likely that the distortion is due to bias. If there is a definite range over which certain characters are found to be consistently in error this is due to characteristic distortion. If the limits are not definite, that is, if there is a range over which errors occur but not consistently on certain characters this is probably due to fortuitous distortion. Although a qualitative analysis of the distortion may be made in the manner discussed above, this indirect method is somewhat laborious and may give misleading results. Moreover, it is impossible to get a measurement of isolated distortions of high value.

G. Telegraph Service Monitoring Set

An automatic telegraph service monitoring set has been designed for the purpose of giving an alarm at repeater stations whenever the distortion on circuits becomes abnormally high or whenever an excessive number of large distortions or "hits" is experienced. This set is still under development; however, a description of it will be given because it is thought to be of general interest to those concerned with telegraph transmission measuring.

In the interest of economy and simplicity this set contains a so-called shortest-pulse type of measuring circuit rather than a start-stop type. Measurement on this basis will of course result in a loss in

accuracy for some signal combinations because of the fact, as mentioned earlier, the distortion of the stop pulse is not added to the distortion of other pulses. However for the purpose for which it is used, namely to detect trouble conditions on working circuits, the accuracy is believed to be adequate.

In this set two measuring circuits are provided, one for measuring marks and the other spaces. These are condenser-charging circuits in which the charge on the condenser is an indication of the duration of the pulse. These condenser voltages are compared to a reference voltage and only those less than the reference voltage are permitted to influence the distortion indicator. Therefore, only the shortest pulses are measured and this permits observation on working teletypewriter circuits without involving a start-stop arrangement. By adjustment of the time constant of the condenser-charging circuits, as for instance by means of continuously variable resistances, the percentage distortion for which an alarm is given may be varied at will.

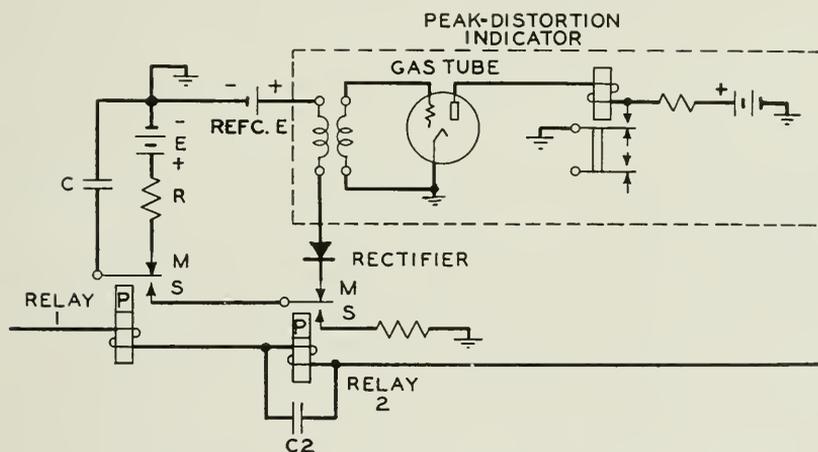


Fig. 22—Distortion-measuring circuit of telegraph service-monitoring set.

The distortion-measuring portion of the circuit used in the measurement of marks is indicated by Fig. 22. Condenser C is charged during marking intervals through high resistance R by voltage E ; thus a voltage is produced on the condenser which depends upon the duration of the marking interval as is indicated by Fig. 23. At the time of the transition from mark to space the condenser voltage is momentarily compared to that of a reference source by way of the armature and marking contact of relay 2 of Fig. 22. Immediately afterwards the condenser charge is dissipated by the armature of relay 2 moving to

its spacing contact. The small delay required in the operation of relay 2 with respect to relay 1 is obtained by connection of condenser C_2 around the windings of relay 2.

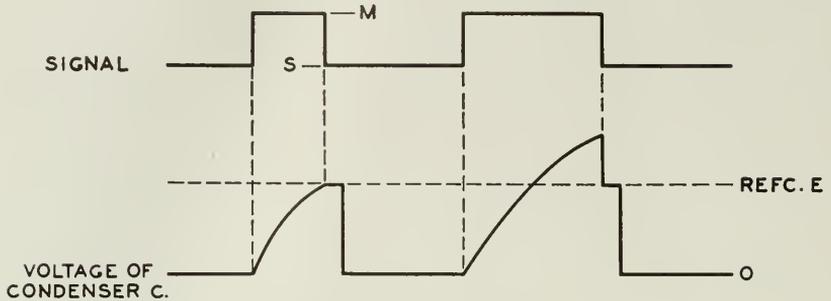


Fig. 23—Variation of charge on measuring condenser in telegraph service-monitoring set.

If at the time of the momentary comparison between the condenser and reference voltages these two voltages are equal, no current will flow in the peak distortion indicator. This condition can be obtained for marks of unit duration (for instance the duration of a selecting pulse of the teletypewriter code) by suitably adjusting the reference voltage REFC. E . For marks of longer duration however the condenser voltage will exceed the reference voltage but, by properly poling the rectifier in series with the peak-distortion indicator, current is effectively prevented from flowing at the time of the momentary comparison. With this poling of the rectifier the current flowing due to the marks being shorter than unit duration will affect the peak-distortion indicator and if suitably adjusted will cause a gas-filled tube to operate and thereby give an indication of distortion.

It will be apparent that the measurements of spaces may be accomplished by providing for that purpose another circuit of the same type as that of Fig. 22. For the measurement of both marks and spaces four relays are required, but only one peak-distortion indicator is necessary.

Associated with the measuring circuit is a counting circuit which counts the number of excessive distortions experienced in a given time. This circuit is indicated schematically by Fig. 24. In this circuit the charge of the condenser C is mixed with that of a larger condenser C_2 at each operation of the gas-filled tube of the distortion measuring circuit. Thus the charge on the larger condenser becomes an indication of the number of excessive distortions. By arranging a timer to discharge this condenser through an indicating circuit suitably

designed an alarm may be obtained whenever the number of excessive distortions exceeds any predetermined number up to about 7 within a given time. For this purpose another gas-filled tube is

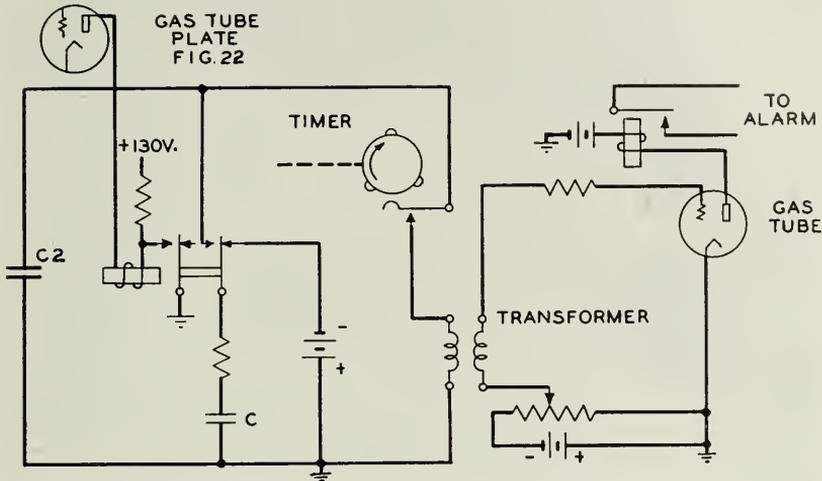


Fig. 24—Counting and alarm circuit of telegraph service-monitoring set.

employed having a potentiometer associated with it for the purpose of adjusting the grid-biasing potential so that the tube will fire only for voltages exceeding definite amounts corresponding to definite numbers of excessive distortions. It is apparent that this type of counting circuit could be replaced by a counting-relay circuit or by a selector-switch circuit such as is used in automatic telephony.

A front view of the monitoring set, arranged for mounting on a relay rack in a central office, is shown by Fig. 25.

In the present arrangement these sets have jacks at a number of places along the telegraph board in the central office, for the purpose of permitting attendants to use the set conveniently. Alarm lamps and signals are provided at the board to attract the attention of the attendant. As the use of these sets is developed, it may be found desirable to employ them in conjunction with a patrol arrangement by means of which a given set may be connected in turn to each of a number of circuits for a short interval. An arrangement of this sort was described by W. Schallerer.⁸

OTHER DEVELOPMENTS

Other types of measuring apparatus have been used experimentally in the Bell System and have been found of value in laboratory work

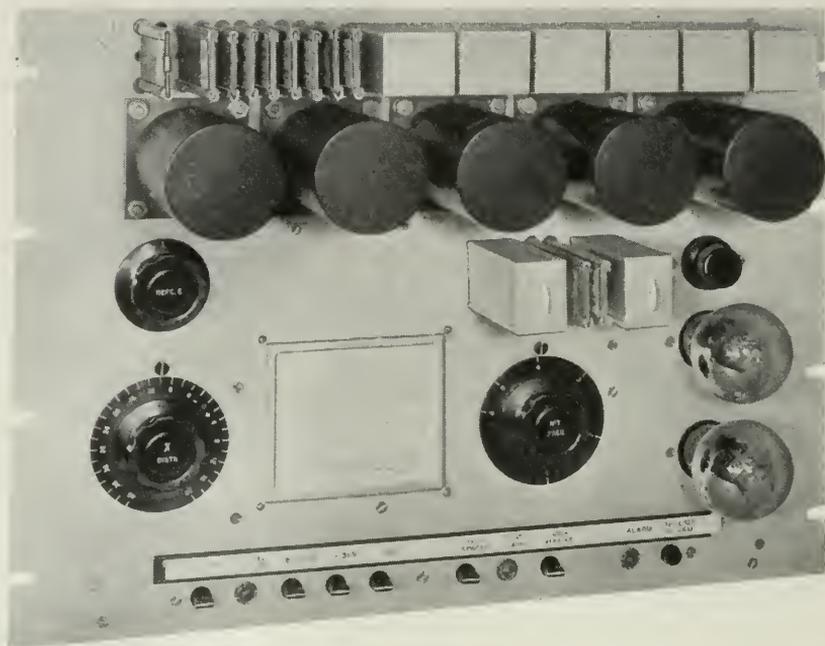


Fig. 25—Telegraph service-monitoring set, test design.

and special investigations. Among these devices are start-stop and synchronous distortion indicators employing flashes of light to indicate the position of the transitions of teletypewriter signals and a photographic recorder of teletypewriter signal transition points. A synchronous flashing indicator in combination with a distributor arranged to supply teletypewriter signals with adjustable bias is in use in shop tests of teletypewriters and in the laboratory.

ACKNOWLEDGMENT

In the development of the practical telegraph transmission measuring devices, which have been described, many members of the Bell System have contributed valuable ideas and effort. The authors wish to express their appreciation of the cooperation they have received.

APPENDIX

GENERAL

This appendix considers characteristic distortion of telegraph signals from the standpoint of the development of simple and convenient testing technique for application primarily to circuits transmitting start-stop teletypewriter signals.

A previous paper² developed methods of determining the correct transfer admittance for distortionless transmission under certain assumed conditions. One general and simplifying assumption was that the time interval between transitions in the telegraph signals would be an integral number of time units. If telegraph circuits were to be designed for the transmission of telegraph signals of this nature, distortionless transmission would be expected when the overall transfer admittance of the circuit was one of the many possible admittances discussed in the previous paper. Although a knowledge of the admittance requirements is a helpful guide in the design of circuits and permits the establishment of certain boundaries, the exact adjustment of transfer admittance to the proper value for satisfactory transmission on the basis of admittance measurements presents many practical difficulties and up to the present has not been generally used.

Another approach to the problem is the actual transmission of miscellaneous signals of the type required and the adjustment of the transfer admittance on a cut-and-try basis until the overall results are satisfactory. For relatively simple circuits satisfactory results can be obtained in this manner. However, for circuits which are electrically long and contain complex networks, a more orderly approach is desirable.

The problem may also be approached from the standpoint of adjusting the transfer admittance of the circuit so as to minimize the transient associated with each transition at the times at which succeeding transitions may occur. This may, of course, be done by means of oscillograph observations, but this procedure has serious practical disadvantages. An advantageous method, however, is to measure the characteristic distortion of simple signal combinations while making the adjustments. For this purpose, signals, each composed of two transitions, repeated at intervals long compared to the duration of the appreciable transient, are used. If the circuit is adjusted so as to transmit without distortion signals having respectively separations of one, two, three, etc. signal elements, between transitions, the requirements for distortionless transmission of miscellaneous signals of the nature under consideration are met, as will be shown below.

Repeated two-transition signals with varying integral time-unit intervals between transitions may be considered as telegraph reversals having a frequency determined by the period of repetition and bias determined by the interval between transitions. Therefore, telegraph reversals of varying bias may be used as a source of test signals, with a simple bias-measuring set at the receiving end and the input bias-output bias characteristic of a circuit determined for

checking the characteristic distortion. The possible use and meaning of such measurements are discussed in the following:

BIAS IN-BIAS OUT

For the purpose of discussion consider the particular code indicated by Fig. 26.

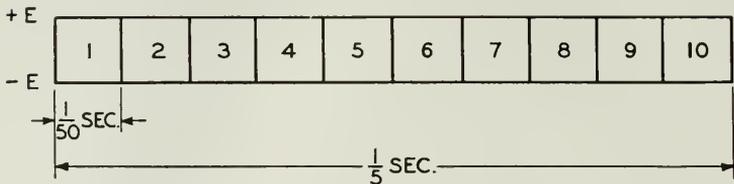


Fig. 26—Special telegraph signal. Each character consists of ten time units. Each time unit is one-fiftieth second and may be $+E$ or $-E$ depending on the character transmitted.

Each character to be transmitted is composed of ten time units, each unit being $1/50$ second, so that it takes $1/5$ second to transmit any of the 1024 possible characters. Although the nominal speed of signaling is 25 dots per second, the fundamental frequency of any character sent repeatedly is 5 cycles per second considered in the Fourier sense. Among the many possible signal combinations are the following which are equivalent to 5-cycle reversals with the amounts of bias indicated below.

Time Units		Time Bias $= 100 \frac{M-S}{M+S}$
Mark	Space	
10.....	0	+100%
9.....	1	+80%
8.....	2	+60%
7.....	3	+40%
6.....	4	+20%
5.....	5	0%
4.....	6	-20%
3.....	7	-40%
2.....	8	-60%
1.....	9	-80%
0.....	10	-100%

It is obvious for a symmetrical circuit that if signals with any amount of positive bias are transmitted accurately signals with the corresponding amount of negative bias will be transmitted since this corresponds to a reversal of the sending polarities.

The transfer admittance for distortionless transmission of these signals will now be derived for the case in which the band width is a minimum. It has been shown in a previous paper² that the minimum band width required is equal numerically to the speed of signaling.

Therefore a band width of 25 cycles will be necessary and sufficient for the signals listed above; this band will also meet the requirements for transmitting the remaining characters of the possible 1024.

The Fourier series for the biased reversal is

$$E(t) = E - \frac{bE}{2\pi} + \frac{E}{\pi} \sum_{n=1}^{\infty} \frac{1}{n} [\sin n(\omega t - b) - \sin n\omega t],$$

where E and b are defined in Fig. 27, which shows the impressed voltage

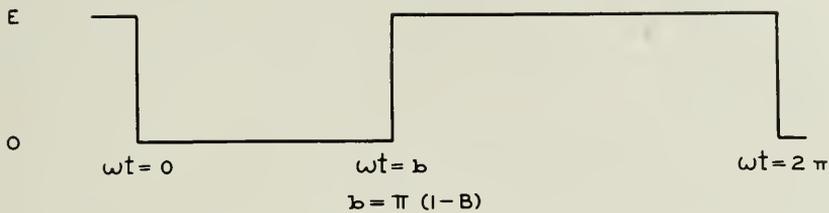


Fig. 27—Voltage wave for biased reversal.

wave at the transmitting end of the circuit, and B is percentage bias divided by 100.

Let the transfer admittance be Y_n at frequency corresponding to $n\omega$. Then, the requirement for perfect transmission of the reversal, assuming the receiving device to operate when the current passes through $EY_0/2$, is that

$$f(t) = EY_0/2 \text{ at } \omega t = 0 \text{ and at } \omega t = b,$$

where $f(t)$ denotes received current.

$$\text{At } \omega t = 0 \text{ or } b, f(t) = \left(E - \frac{bE}{2\pi} \right) Y_0 + \frac{E}{\pi} \sum_{n=1}^{\infty} \frac{Y_n}{n} \sin(-nb).$$

Hence, assuming for simplicity that $Y_0 = 1$

$$\frac{E}{2} = E - \frac{bE}{2\pi} + \frac{E}{\pi} \sum_{n=1}^{\infty} \frac{Y_n}{n} \sin(-nb).$$

Substituting the value of b and transforming

$$\frac{B\pi}{2} = \sum_{n=1}^{\infty} \frac{Y_n}{n} \sin nb.$$

In accordance with the assumption that band width is a minimum,

$Y_n = 0$ for $n \geq 5$. Therefore, the values of Y_n may be determined from the following equations wherein the values of nb are in degrees.

$$\begin{aligned} \text{For } B = .2 \quad .1\pi &= Y_1 \sin 144 + \frac{1}{2}Y_2 \sin 288 + \frac{1}{3}Y_3 \sin 432 + \frac{1}{4}Y_4 \sin 576, \\ B = .4 \quad .2\pi &= Y_1 \sin 108 + \frac{1}{2}Y_2 \sin 216 + \frac{1}{3}Y_3 \sin 324 + \frac{1}{4}Y_4 \sin 432, \\ B = .6 \quad .3\pi &= Y_1 \sin 72 + \frac{1}{2}Y_2 \sin 144 + \frac{1}{3}Y_3 \sin 216 + \frac{1}{4}Y_4 \sin 288, \\ B = .8 \quad .4\pi &= Y_1 \sin 36 + \frac{1}{2}Y_2 \sin 72 + \frac{1}{3}Y_3 \sin 108 + \frac{1}{4}Y_4 \sin 144. \end{aligned}$$

Solving these simultaneous equations it is found that $Y_1 = .967$, $Y_2 = .865$, $Y_3 = .685$, $Y_4 = .408$.

The computed admittance is the same as that which may be computed from the information given in Appendix III of a previous paper² which is:

$$Y = \frac{\pi f}{2 f_s} \cot \frac{\pi f}{2 f_s},$$

Y = transfer admittance at frequency f ,
 f_s = dotting speed.

Since the bias of the signals at the circuit output must equal the bias of the signals at the circuit input for the magnitudes of bias entering

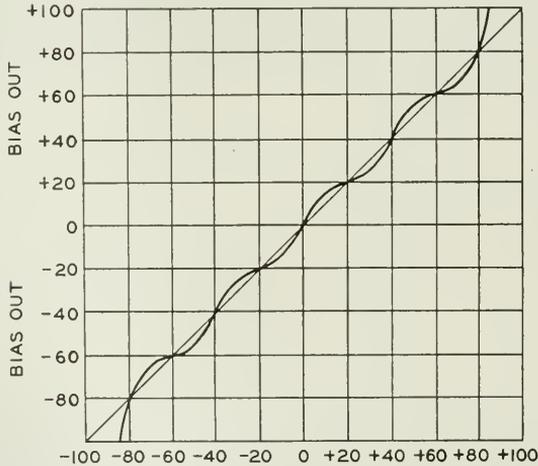


Fig. 28—Bias in-bias out characteristic for transmission of telegraph signals formed in accordance with Fig. 26 over a system having a band width equal to the dotting speed.

into the determination of the admittance, and since the transmission of additional harmonics of the fundamental frequency would be required to make the input bias equal to the output bias at additional magni-

tudes of bias, the Bias In-Bias Out characteristic for a circuit having an admittance as computed above would have the general characteristics indicated in Fig. 28.

Considered in terms of the transient behavior, if the circuit had the transfer admittance defined by the above equation, or any of the infinite number of other prescribed admittances using higher frequencies, the transient resulting from a single transition would be such as to have zero value at each of all possible future transition points. Also if the circuit were adjusted so that the four simple characters were transmitted with negligible distortion, the transients would fulfill the conditions for the satisfactory transmission of the other 1020 possible characters.

From Fig. 28, it is obvious that a circuit which had a perfect transfer admittance and a frequency band width large compared to the character repetition frequency, would have a Bias In-Bias Out characteristic which crossed the 45-degree line at many points and approached it as a limit. It is interesting to note the relation between the deviations of the Bias In-Bias Out characteristic from the 45-degree line and the frequency band width. In the example under discussion there are five waves in the characteristic which correspond to the frequency band width divided by the number of characters per second. The band width required is numerically equal to the product of the number of waves in the Bias In-Bias Out characteristic and the number of characters transmitted per second.

START-STOP TELETYPEWRITER SIGNAL

Start-stop teletypewriter systems may employ varying speeds and signal arrangements. The 60-word-per-minute (60-speed) system is the one most generally used in the Bell System and is taken as an example in this appendix. Similar methods of analyses and tests could be applied to other systems.

The 60-speed teletypewriter signal consists of a starting unit which is always spacing, five selecting units, and a stop signal which is 1.42 units in length and is always marking. The duration of a unit signal pulse is 22 milliseconds and the total length of each character is, therefore, $1 + 5 + 1.42 = 7.42$ times units or 163 milliseconds. With no pause between succeeding characters there are 368 operations per minute or

$$\frac{368}{60} = \frac{1}{.163} = 6.13 \text{ operations per second.}$$

The problem of transmitting these signals without distortion is

similar to the problem just discussed except instead of uniform spacing between possible transitions of $1/50$ second, the transitions may be spaced at intervals of either .022 second, .031 second, or combinations of these intervals, and the frequency of repetition is 6.13 instead of 5 per second.

As indicated in Fig. 1, there are six teletypewriter characters (Blank, *T*, *O*, *M*, *V*, and Letters) which correspond to 6.13 cycle reversals biased by certain amounts. It will now be shown that if these six characters can be transmitted without distortion, the other 26 characters will also be transmitted without distortion. The method is the same as that used in the preceding problem.

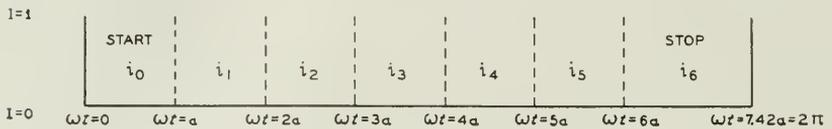


Fig. 29—7.42-Unit code start-stop teletypewriter signal.

Figure 29 indicates any teletypewriter signal where $i_0 = 0$, and $i_6 = 1$. i_1, i_2, i_3, i_4, i_5 , may have values of 0 or 1 depending on the particular character. The Fourier series for the received current over a circuit with a transfer admittance Y having unit value at zero frequency, is

$$\begin{aligned}
 f(t) = & \frac{i_1 + i_2 + i_3 + i_4 + i_5 + 1.42}{7.42} \\
 & + \frac{1}{\pi} \sum_{n=1}^{n=\infty} \frac{Y_n}{n} \{ i_1 [\sin n(\omega t - a) - \sin n(\omega t - 2a)] \\
 & + i_2 [\sin n(\omega t - 2a) - \sin n(\omega t - 3a)] \\
 & + i_3 [\sin n(\omega t - 3a) - \sin n(\omega t - 4a)] \\
 & + i_4 [\sin n(\omega t - 4a) - \sin n(\omega t - 5a)] \\
 & + i_5 [\sin n(\omega t - 5a) - \sin n(\omega t - 6a)] \\
 & + 1 [\sin n(\omega t - 6a) - \sin n \omega t] \}.
 \end{aligned}$$

Suppose that the transmitted frequency band is limited to 6 times the fundamental, i.e. $n = 1$ to 6, and Y_n adjusted so that the characters "Letters" *V*, *M*, *O*, *T* and "Blank" are transmitted perfectly.

The transfer admittance will now be determined for this case. The expression for $f(t)$ may be written for the characters just mentioned, and would have the value $1/2$ at $\omega t = a$, for "letters," at $\omega t = 2a$ for *V*, at $\omega t = 3a$ for *M*, etc. Accordingly there result six equations which may be simplified as follows:

$$\begin{aligned}
 \text{“Letters”} & \quad \sum_1^6 \frac{Y_n}{n} \sin na = \frac{2.71\pi}{7.42}, \\
 V & \quad \sum_1^6 \frac{Y_n}{n} \sin 2na = \frac{1.71\pi}{7.42}, \\
 M & \quad \sum_1^6 \frac{Y_n}{n} \sin 3na = \frac{.71\pi}{7.42}, \\
 O & \quad \sum_1^6 \frac{Y_n}{n} \sin 4na = \frac{-.29\pi}{7.42}, \\
 T & \quad \sum_1^6 \frac{Y_n}{n} \sin 5na = \frac{-1.29\pi}{7.42}, \\
 \text{“Blank”} & \quad \sum_1^6 \frac{Y_n}{n} \sin 6na = \frac{-2.29\pi}{7.42}.
 \end{aligned}$$

The values of Y_n computed by solving these equations were plotted on Fig. 30 and a curve was drawn through them. It will be understood, that, on the scale of abscissae, F is the fundamental frequency of the character and the coefficients of F are values of n .

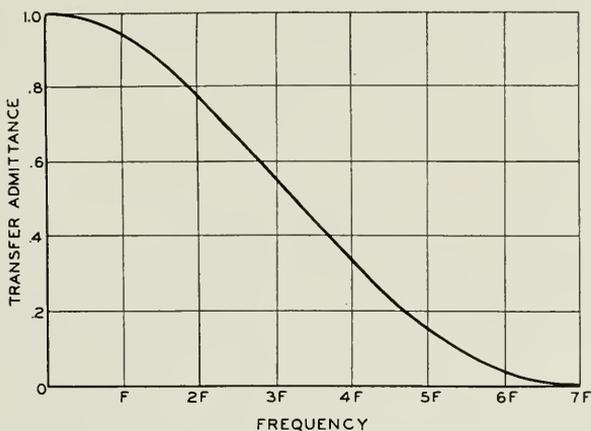


Fig. 30—Transfer admittance characteristic for undistorted transmission using frequencies up to seven times the fundamental frequency of the character.

By the methods employed in the previous paper² it can be shown that, for signals employing the same permissible intervals between transitions, the solutions of equations similar to the above approach the smooth curve as a limit as the period of repetition is lengthened.

These equations may be used to prove that when the six simple characters are perfectly transmitted, all other teletypewriter signals may be transmitted on a repeated basis. For distortionless transmission $f(t)$ should equal $1/2$ at each transition point, assuming that the

relay operates at this value of current. If the expression for $f(t)$ is written for each of the conditions which may occur, and is found to equal $1/2$ at the transition points, the proof is complete. For this purpose the summation signs may be eliminated by substituting from any of the six simultaneous equations numerical values in place of summations of terms, in the equation for $f(t)$, with the following results:

CONDITIONS

$$\begin{aligned} \omega t = 0 \quad i_0 = 0 \quad i_6 = 1 \quad f(t) &= \frac{1}{7.42} [i_1 + i_2 + i_3 + i_4 + i_5 + 1.42 + i_1(-1) \\ &\quad + i_2(-1) + i_3(-1) + i_4(-1) \\ &\quad + i_5(-1) + 1(2.29)] = \frac{1}{2}, \\ \omega t = a \quad i_0 = 0 \quad i_1 = 1 \quad f(t) &= \frac{1}{7.42} [1 + i_2 + i_3 + i_4 + i_5 + 1.42 + 1(2.71) \\ &\quad + i_2(-1) + i_3(-1) + i_4(-1) \\ &\quad + i_5(-1) + 1(1.29 - 2.71)] = \frac{1}{2}, \\ \omega t = 2a \quad i_1 = 0 \quad i_2 = 1 \quad f(t) &= \frac{1}{7.42} [0 + 1 + i_3 + i_4 + i_5 + 1.42 + 0(2.71) \\ &\quad \text{or} \\ &\quad i_1 = 1 \quad i_2 = 0 \\ &\quad + 1(2.71) + i_3(-1) + i_4(-1) \\ &\quad + i_5(-1) + 1(.29 - 1.71)] = \frac{1}{2}, \\ \omega t = 3a \quad i_2 = 0 \quad i_3 = 1 \quad f(t) &= \frac{1}{7.42} [i_1 + 0 + 1 + i_4 + i_5 + 1.42 + i_1(-1) \\ &\quad \text{or} \\ &\quad i_2 = 1 \quad i_3 = 0 \\ &\quad + 0(2.71) + 1(2.71) + i_4(-1) \\ &\quad + i_5(-1) + 1(-.71 - .71)] = \frac{1}{2}, \\ \omega t = 4a \quad i_3 = 0 \quad i_4 = 1 \quad f(t) &= \frac{1}{7.42} (i_1 + i_2 + 0 + 1 + i_5 + 1.42 + i_1(-1) \\ &\quad \text{or} \\ &\quad i_3 = 1 \quad i_4 = 0 \\ &\quad + i_2(-1) + 0(2.71) + 1(2.71) \\ &\quad + i_5(-1) + 1(-1.71 + .29)] = \frac{1}{2}, \\ \omega t = 5a \quad i_4 = 0 \quad i_5 = 1 \quad f(t) &= \frac{1}{7.42} [i_1 + i_2 + i_3 + 0 + 1 + 1.42 + i_1(-1) \\ &\quad \text{or} \\ &\quad i_4 = 1 \quad i_5 = 0 \\ &\quad + i_2(-1) + i_3(-1) + 0(2.71) \\ &\quad + 1(2.71) + 1(-2.71 + 1.29)] = \frac{1}{2}, \\ \omega t = 6a \quad i_5 = 0 \quad i_6 = 1 \quad f(t) &= \frac{1}{7.42} [i_1 + i_2 + i_3 + i_4 + 0 + 1.42 + i_1(-1) \\ &\quad + i_2(-1) + i_3(-1) + i_4(-1) \\ &\quad + 0(2.71) + 1(2.29)] = \frac{1}{2}. \end{aligned}$$

Hence it is concluded that if an admittance can be found such that "Blank" T , O , M , V and "Letters" are transmitted without distortion any other teletypewriter signal may be transmitted on a repeated basis, since the correct current value ($1/2$) will be obtained at any transition regardless of what other signal combination is used for the remainder of the character.

Computations have also been made for the case of a repeated signal combination consisting of any two teletypewriter characters. The results indicate that this may also be transmitted without distortion if the requirements have been met for the six simple characters sent repeatedly.

When distortion is present, there is deviation of the received current from the desired value at the time it should be $1/2$ and the deviation to be expected at any transition for any character may be computed from equations similar to those given above. This, of course, differs from the amount of distortion which would be measured in per cent. It is of interest that the characteristic distortion on the more complicated characters may be materially greater than that measured on the simple two-transition characters.

The transient behavior of a circuit having the transfer admittance specified by the smooth curve of Fig. 30 may be determined by methods utilizing the Fourier integral. However, certain characteristics of this transient, namely, the points at which it must be zero, are known in advance from the conditions entering into the determination of the admittances. Although these conditions do not directly prescribe the magnitude of the transient at other points, additional information may be obtained from an inspection of Fig. 31.

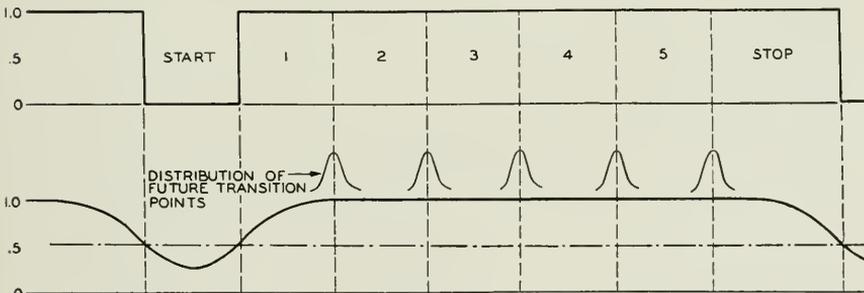


Fig. 31—Received current for "letters" character assuming the transfer admittance of Fig. 30.

This figure shows the computed received current over a telegraph circuit having the admittance shown in Fig. 30 when a repeated teletypewriter "Letters" signal is transmitted. It may be noted that the transient decays to inappreciable amplitudes at times greater than the shortest time unit of .022 second. This is significant, since it means that a particular arrangement of transitions is not of practical importance as long as transitions do not come at intervals closer than .022 second. With this admittance, therefore, no difficulty would be

expected from keyboard sending, the signals from which differ from the signals previously discussed, in that the lengths of the stop signals occur on a random basis, and are never shorter than .031 second, depending on the typist.

However, if signals containing bias or other distortion such as indicated on Fig. 31 as distortion of future transition points were transmitted over the circuit, thus decreasing the minimum interval between transitions, the transient from one transition, for example, might affect following transitions. A circuit having a shorter build-up time could be made to introduce less distortion on transitions spaced at very short intervals. The rate of build-up is a function of the area under the transfer-admittance curve and from a design standpoint it is necessary to provide a suitable frequency band-width to make the slope of the received signals sufficient so that characteristic distortion will not be excessive when closely spaced transitions are transmitted, and also so that certain types of interference will not cause excessive fortuitous effects.

Where the available band width is limited the area under the admittance curve could be increased by transmitting the permissible harmonic frequencies at a greater amplitude. The transient of such a circuit would continue at appreciable magnitudes for a greater length of time and the minimum distortion would be increased but the general circuit stability might be improved.

The foregoing discussion of telegraph signal transmission has shown that when a circuit is adjusted to transmit without distortion certain selected repeated characters, the circuit can transmit on a repeated basis any of the characters possible with signals of the nature represented by the selected characters. This is true not only for the signals in which transitions are spaced at integral units of time but for signals in which transitions are spaced at predetermined non-integral units of time, such as start-stop teletypewriter signals. Incidental to the development of the proof of the later statement, the prescribed admittance for transmission without distortion was evaluated. The prescribed admittance for signals employing integral units between transitions was also similarly determined and found to be the same as that which had been determined from a somewhat different approach in a previous paper.²

The admittances considered have been idealized somewhat, inasmuch as no physical circuit will cut off completely at the higher frequencies and, in addition, the effective transfer admittance is not only a complex, and frequently nonlinear quantity, but is determined in part by the characteristics of transmitting and receiving relays and

other terminal equipment. In the present state of the art, it is difficult to make practical use of transfer admittances in predicting the performance of a telegraph circuit.

The significant point is that the satisfactory transmission of the selected characters is an indication of the ability to transmit the desired telegraph signals satisfactorily. Also, the measurement of the distortion on the selected characters is particularly useful when it is desired to equalize individual circuits of varying length and makeup to secure a minimum of distortion.

The testing procedures suggested by the considerations of the foregoing have been incorporated into the testing instrumentalities discussed in the main paper. These methods have been used for several years in the adjustment and maintenance of telegraph circuits and found to be of considerable utility.

REFERENCES

A. References Cited:

1. "Measurement of Telegraph Transmission," Nyquist, Shanck and Cory, *Trans. A.I.E.E.*, 1927, Vol. 46, p. 367.
2. "Certain Topics in Telegraph Transmission Theory," H. Nyquist, *Trans. A.I.E.E.*, 1928, Vol. 47, p. 617.
3. a. "Telephone Typewriters and Auxiliary Arrangements," R. D. Parker, *Bell Telephone Quarterly*, July, 1929, p. 181.
- b. "Modern Practices in Private Wire Telegraph Service," R. E. Pierce, *Trans. A.I.E.E.*, Jan. 1931, Vol. 50, p. 45.
4. "Fundamentals of Teletypewriters Used in the Bell System," E. F. Watson, *Bell System Technical Journal*, Oct., 1938, p. 620.
5. "A Transmission System for Teletypewriter Exchange Service," Pierce and Bemis, *Bell Sys. Tech. Journal*, Oct., 1936, p. 529; *Elec. Engg.*, Sept., 1936, p. 961.
6. "Metallic Polar-duplex Telegraph System for Cables," Bell, Shanck and Branson, *Trans. A.I.E.E.*, 1925, Vol. 44, p. 316.
7. a. "Der Spielraum des Siemens-Springschreibers," M. J. deVries, *Telegraphen und Fernsprech Technik*, Jan., 1934, p. 7.
- b. "Der Spielraum des Springschreiber," M. J. deVries, *T. F. T.*, Sept., 1937, p. 213.

B. Additional References:

8. "Ein Neues Mess- und Überwachungsgerät für Springschreiberverbindungen," W. Schallerer, *T. F. T.*, Feb., 1935, p. 40.
9. "Versuche über eine günstige Verteilung der Trägerwellen in der Wechselstromtelegraphie," H. Stahl, *T. F. T.*, Nov., 1930, p. 340.
10. "Verzerrungsmesser für Telegraphie," A. Jipp and O. Römer, *T. F. T.*, May, 1932, p. 121.
11. "Telegraph Transmission Testing Machine," F. B. Bramhall, *Trans. A.I.E.E.*, June, 1931, p. 404.
12. "A Telegraph Distortion Measuring Set," V. J. Terry and C. H. W. Brookes-Smith, *Elec. Comm.*, July, 1933, p. 15.
13. "The Measurement of Telegraph Distortion," V. J. Terry, *Elec. Comm.*, April 1933, p. 197.
14. "Determining the Transmission Efficiency of Telegraph Circuits," E. H. Jolley, *P.O.E.E. Jl.*, April, 1933, p. 1.

Contemporary Advances in Physics, XXXII Particles of the Cosmic Rays

By KARL K. DARROW

Even after fifteen years of intensive research following on two decades of more desultory study, the cosmic rays are still a store of new and remarkable data. The question of their ultimate origin, though by no means extinct, has been set aside by many physicists in favor of a fuller inquiry into their qualities. The distinctive mark of the cosmic-ray particles is the immensity of their energies; for, great by all previous standards as are the energy-values which physicists now can impart in their laboratories, those manifest in the cosmic rays are greater by factors not of thousands merely, but often of millions. To this remote and exalted energy-range belong the penetrating particles capable of cleaving through a metre of lead, and the wonderful and beautiful phenomenon of cosmic-ray showers. It is not to be wondered at that with energies so high, particles so familiar as electrons and photons should be invested with unfamiliar powers. So evidently they are; but some of the charged corpuscles of the cosmic rays have properties such that their strangeness cannot be ascribed to high energy alone, but apparently must be based upon some fundamental difference (perhaps a difference of mass) from all the particles thus far identified.

WHEN a new member is admitted to a small and jealously-restricted club supposedly already filled for all time, the event has a dramatic aspect. When a concept is formed in a nebulous way and rapidly gains precision with the passage of the years, the story is of philosophic interest. When physicists extend their knowledge into ranges of energy heretofore unsuspected, and find them inhabited by particles classifiable as electrons but in possession of powers ordinarily unknown, and also by particles which must be put in a class by themselves—when such things are available for telling, the tale has scientific value. When evidence comes in the form of pictures so striking as those which can here be shown, the science of lifeless matter has an aesthetic splendor such as rarely embellishes it. All of these features appear in the recent advances of the study of cosmic rays.

The small and exclusive club consists of the subatomic particles, long supposed to comprise only the negative electron and the proton and other positive atom-nuclei. Into it the positive electron had been forced in 1932, and the neutron in 1933; a vacant chair was

being reserved for the negative proton, which as yet has not turned up to claim it; few if any expected the actual applicant. The concept now hardening into the definite form of this applicant is that of the "mesotron." This is a particle presumed to be equal in charge to the electron, but in mass a couple of hundreds of times as great. In so naming it I follow (C. D.) Anderson's recent proposal, though other titles such as "barytron" and "heavy electron" are already more or less firmly rooted in the literature. The quality which marks it out, when it appears with enormous energy among the cosmic rays, is an extreme and almost incredible power of penetration. This means that the so-called mesotrons are able to traverse decimetres, nay even metres of lead (or of dense matter generally). Like electrons, mesotrons may be of either sign of charge. As for the cosmic-ray particles still classified as electrons, *they* are marked out by their power of producing one of the most magnificent phenomena of Nature, the "shower of cosmic rays," or "shower" for short. Shower-production by the supposed electrons, penetration by the supposed mesotrons, ionization along the course of either corpuscle through air: these are the three phenomena which will furnish most of the illustrations, much of the text of this article. The story of their incorporation into the structure of physical theory will furnish the remainder.

(But negative electrons and protons, not to speak of other atom-nuclei, have been identified through having their charge-to-mass ratios measured with the aid of electric and magnetic deflecting fields in elementary classical ways. Why then do I not cut this introduction short by giving the results of such a measurement upon the mesotron? The reason is, that no such measurement has yet been made. Probably one will be made ere long. Should it give something near to the result expected, the delay will not have been regrettable; for the end of the delay will mark the beginning of the time, when the story to be related in these pages will be regarded as being "of historical interest" only—which is to say, that it will then be liable to be forgotten.)

So that the reader may see at once the three phenomena which are to bulk so largely in this story, I draw his attention at once to some of the pictures which decorate this article.¹ Nearly all of them were made (of course) with the aid of the cloud-chamber or expansion-chamber of C. T. R. Wilson, that device so precious in physics and precious in so many ways.

¹They decorate it with particular clarity, thanks to the kindness of Messrs. Anderson, Auger, Brode, Corson, Fowler, Fussell, Neddermeyer, Stevenson and Street in supplying me with prints of their splendid photographs.



Fig. 1—Track of a cosmic-ray particle (probably an electron) in the expansion-chamber, time having been allowed for the ions to drift apart before the expansion. (Corson and Brode, University of California)

At the beginning I place, as Fig. 1, a picture of the track of a cosmic-ray particle believed to be an electron. Anyone who has ever studied the pictures of cloud-chamber tracks will at once be impressed by seeing how distinctly the droplets stand apart. This separation was achieved by letting half a second elapse from the instant when the electron shot through, to the instant when by expansion the gas of the chamber grew suddenly cool and the water-vapor suspended in the gas condensed itself as dewdrops on the ions. These ions, formed by the passage of the electron, had been diffusing through the gas during the half-second intervening, and the diffusion-process had served in the main to carry them apart (though there must also have been cases of ions approaching and possibly even combining with each other). The counting of these droplets is germane to the question as to whether the traversing particle was or was not an electron. This question, however, we leave till later, and turn to photographs in which the droplets of the tracks lie close together and are uncountable, because the expansion took place before there had been time for much diffusion. Tracks so formed have the advantage of sharpness over what they lose in detail.

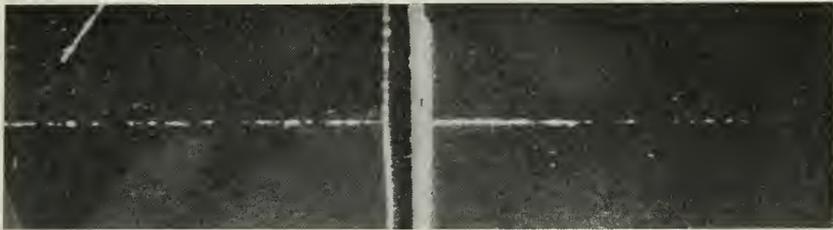


Fig. 2—Track of a particle, presumably a mesotron, traversing a metal plate without sensible deflection. (Auger; Université de Paris)

Figure 2 presents the track of a particle which traversed a plate of lead as it shot across the chamber. In passing through the lead, it underwent no sensible deflection; no other particle sprang from the lead; and there is nothing in the aspect of the track which differs on the two sides of the metal. It would be more impressive yet to present a similar picture for a particle traversing ten or fifty centimetres of lead, but here the practical limitations on the size of a Wilson chamber defeat the physicist, or at any rate no one has overcome them yet. Ehrenfest has lately circumvented them by the laborious scheme of setting up *two* Wilson chambers, one above the other, with as much as 9 cm. of lead or gold between them. However, the passage of single

charged particles through thicknesses as great or even much greater is amply attested by the scheme of apparatus sketched in Fig. 3, even without the cloud-chamber there indicated by "Ch."

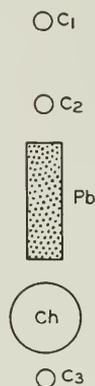


Fig. 3—Scheme of apparatus for observing very penetrative particles with counters and cloud-chamber.

In this sketch of Fig. 3, the objects C_1 and C_2 and C_3 are Geiger-Müller counters: that is to say, gas-filled discharge-tubes of a very special design, the two electrodes of each being an axial wire and a coaxial cylinder, and the electrode-size, voltage, and gas-content being very carefully adjusted. These long large cylinders, usually called simply "counters" without the prefixed names, are familiar sights in almost every laboratory where cosmic rays are studied. If a charged flying corpuscle penetrates such a tube, a momentary discharge takes place in the gas. If such discharges spring up simultaneously in all the three tubes of such a system as Fig. 3 exhibits, the event is recorded by a mechanism. ("Simultaneously" is of course a word which requires detailed exegesis; it meant at first that in all tubes discharges began within 0.01 second of each other, but this interval has been pushed down to .0001 second and lower.)

These events, the "threefold coincidences," do actually occur. Of course, since in each of the tubes a discharge occurs now and then by itself, some of the coincidences must be the result of chance; but the probable number of these meaningless ones can easily be estimated from the frequency and the duration of the individual discharges, and in the best experiments they are a small minority. For the great majority, the simplest of explanations is to attribute each of them to a single vertically-flying particle cutting through all of the counters in succession. Yet there are other thinkable causes, and confirmation

of this simplest idea is needed. It was supplied when the cloud-chamber, "Ch" in the figure, was inserted. The chamber was compelled by mechanism to expand, always when and only when a three-fold coincidence happened; and at the great majority of its expansions it showed a vertical track. Figure 3 exhibits the arrangement of Street, Woodward and Stevenson at Harvard, who found the track of the traversing particle at 202 expansions out of 219. Auger and Ehrenfest at Paris had already set up *four* counters and a cloud-chamber and a block of lead in a vertical line, and found the track of the single traversing particle at fifty-five expansions out of sixty-nine. Another test is made by displacing one of the counters out of line with the others, whereupon it is found that the coincidences fall off in number sharply. And now to come to the point which most concerns us: there were 45 cm of lead between the counters in the experiment of Fig. 3, and 50 cm in the experiment by Auger and Ehrenfest, and no fewer than 101 cm in an early experiment of Rossi's with counters though without the chamber! Such is the power of penetration of some of the charged corpuscles of the cosmic rays.

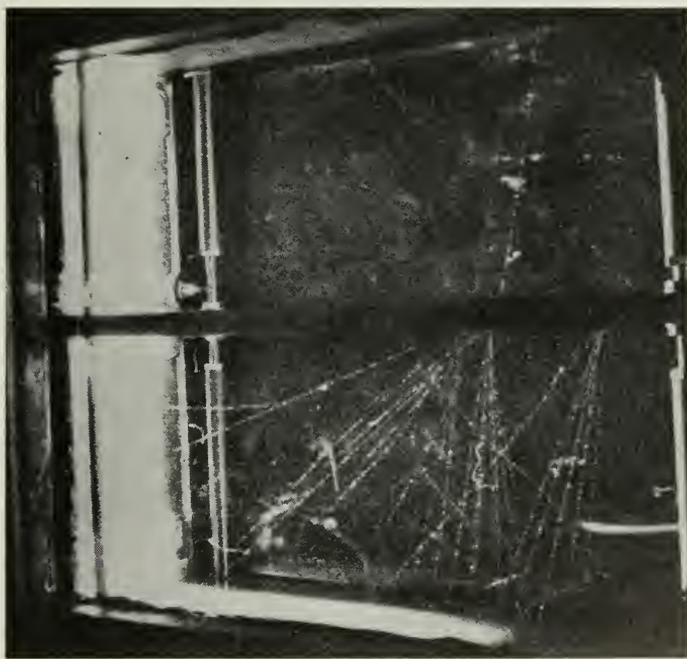


Fig. 4—Three showers, two evoked by charged particles and one presumably by a photon. (Street and Stevenson, Harvard University)

The reader has now been introduced to charged particles which bore through quantities of lead, apparently without doing or suffering anything. Next he is to be introduced to particles which begin to do something startling, when they have scarcely more than entered into a thin metal plate. This is vividly shown to him in Fig. 4, in which—after he can detach his eyes from the pretty sight beneath the transverse leaden plate—he will see that two of the “showers” beneath spring from the places where the metal was entered by two charged particles coming from above. These are accordingly called “shower-producing particles.”

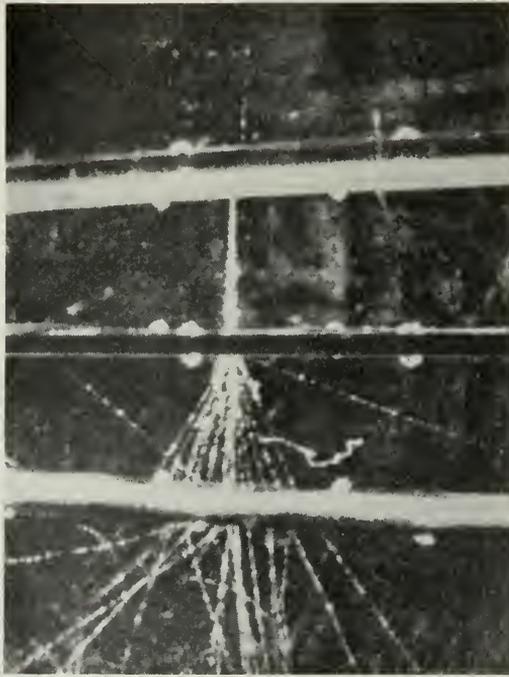


Fig. 5—Shower begun by a charged particle impinging on a 6.3-mm lead plate, and multiplied as it passes through a second such plate; in the third plate, 0.7 mm thick, only deflections occur. (Fussell, Harvard University)

Figures 5 and 6 and 7 show examples of showers even more gorgeous—regular cloudbursts, to continue with the metaphor (and indeed the term “burst” is often used as a synonym for “very large shower”). Of these, the special value of Fig. 6 is that the tracks that start in the gas itself bear witness to corpuscles of light—photons—included in the shower; for these are the tracks of electrons ejected by photons

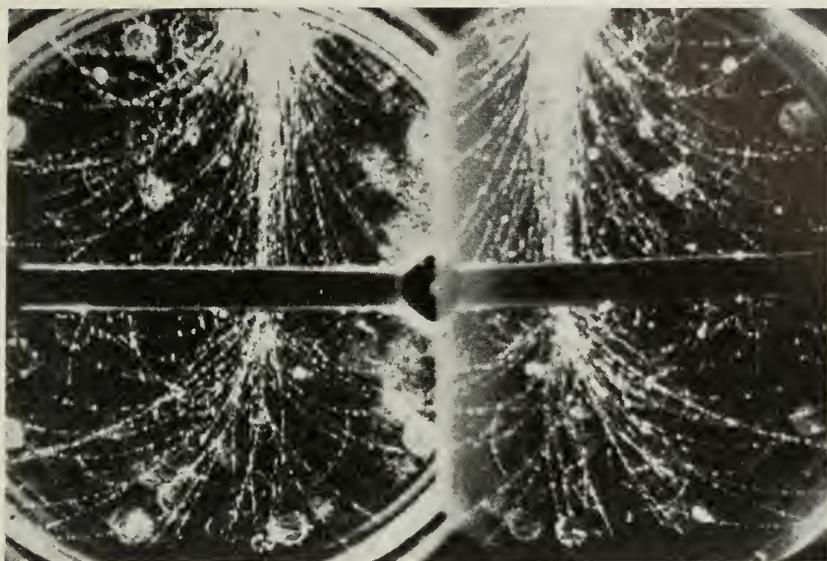


Fig. 6—Shower comprising photons attested by the (curled) tracks of slow electrons released in the gas. (Anderson and Neddermeyer, California Institute of Technology)

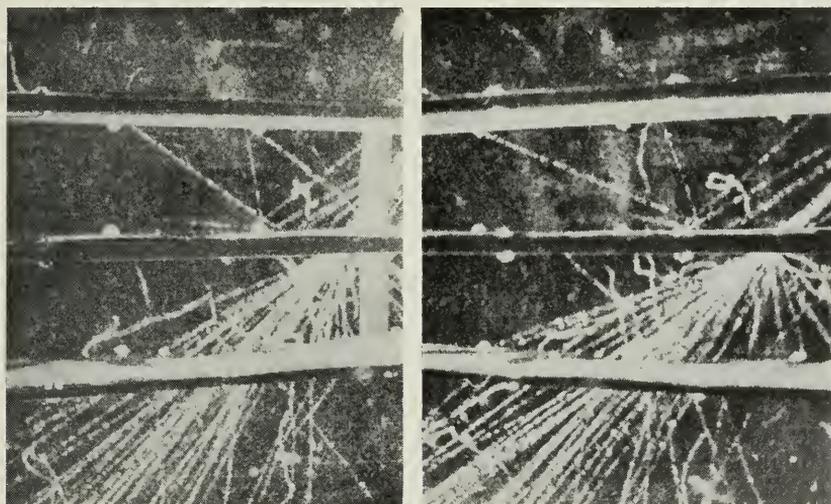


Fig. 7—Another example of a shower undergoing multiplication as it passes through metal plates. (Fussell)

from atoms of the gas. (The agent which bends them into curlicues is, of course, a magnetic field applied to the whole of the Wilson chamber.) Showers, then, comprise photons as well as charged particles. The

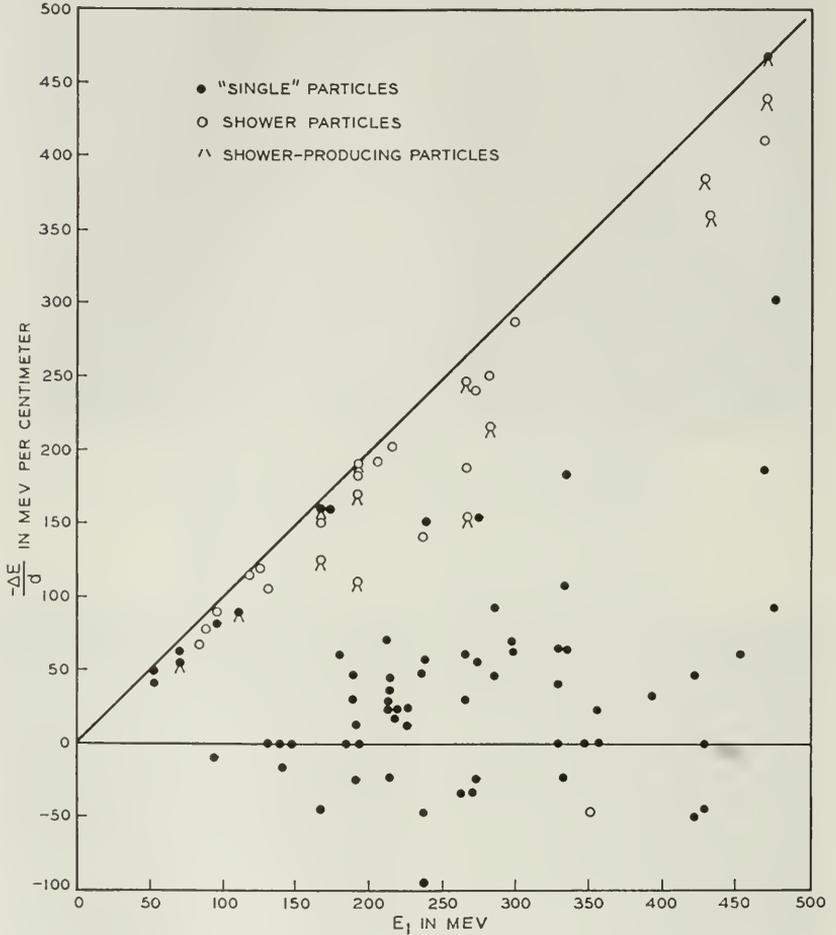


Fig. 8—Energy-losses per unit length of path (in Mev/cm) suffered by 94 cosmic-ray particles in traveling through platinum. (Anderson and Neddermeyer)

special value of Figs. 5 and 7 is, that they show the progressive aggrandizement of showers as these pass onward through dense matter. This is called "the multiplication of showers." *Shower particles are themselves capable of being shower-producing particles.* One could not tell from these figures whether the multiplication is due to the charged particles or the photons, to either singly or to both. Here again the

reader may consult Fig. 4, in order to notice that one of the three showers there depicted sprang from a place in the plate to which no charged particle came. This suggests that a photon may cause a shower, and that the multiplication of a shower already begun is due to the action of its charged particles and of its photons both.

Two classes of charged particles begin to take shape: the penetrating ones on the one hand, the shower particles and the shower-producing particles classified together on the other. To bring out another aspect of the distinction, I now turn to the data underlying Fig. 8.

These data are derived from cloud-chamber photographs such as Fig. 9 exemplifies. If the track of a charged particle is sensibly curved in such a magnetic field as it is possible to apply to a Wilson chamber, it may be possible to infer the momentum and the energy of the particle.² I digress to give the formulae, so as to make it clear just what can be deduced from what amount of knowledge. The elementary procedure consists in pointing out that the charged body describes a circle in the plane perpendicular to the magnetic field, and that consequently the force exerted on it by the field is to be equated to the product of its mass by its centrifugal acceleration. Putting ne for the charge (in electrostatic units) of the corpuscle, m for its mass, v for its speed and p for the magnitude of its momentum in the plane normal to the field, ρ for the radius of the circle and H for the field-strength, and writing down the two members of the equation, one finds:

$$Hnev/c = mv^2/\rho, \quad (1)$$

$$p = (ne/c)H\rho. \quad (2)$$

These equations remain valid when (as usually is the case with cosmic-ray electrons) the speed is so great that relativistic mechanics must be used instead of ordinary. At such high speeds equation (2) retains its aspect. Equation (1) may also be left unaltered, but one must be sure to remember that m is a certain function of v :

$$m = m_0\sqrt{1 - v^2/c^2}, \quad (3)$$

m_0 being known as the "rest-mass" of the body.

² Curvatures of tracks being so very important in this field of research, it is necessary to examine with the greatest of care into all of the causes (apart from magnetic field) which may produce or affect them. Notable among these are currents in the gas, which are especially obnoxious if there is a metal plate in the chamber. Indeed it seems strange that the currents should not be more hampering than they are, considering the expansions which occur. Sometimes people observe that in the absence of magnetic field, there is a slight curvature of the tracks; then in the presence of magnetic field, they deduct this amount from the curvatures observed. The papers of Anderson and Blackett abound in information on these delicate questions.

Equation (2) does not involve the mass at all. In the usual loose phrasing, $H\rho$ gives the momentum of the particle provided that its charge is known. The like cannot be said for the energy, which is given by $H\rho$ only if both the charge and the rest-mass are known. For particles of the cosmic rays it is best to disregard the ordinary expression for kinetic energy ($\frac{1}{2}mv^2$) and adopt for good the relativistic expression mc^2 , to wit, $m_0c^2/\sqrt{1-v^2/c^2}$. Of this the portion m_0c^2 is not kinetic energy: it is the "rest-energy" associated with the "rest-mass" m_0 , inseparable from the particle so long as this exists; it amounts to about half-a-million electron-volts or 0.5 Mev for the electron, to about 1000 Mev for the proton. The remainder may be called kinetic energy. For nearly all of the electrons and most of the other cosmic-ray particles, this remainder is by far the greater part. The dependence of the kinetic energy upon $H\rho$ is exhibited, for electrons and for protons, by Fig. 13 (page 213). One sees that for different

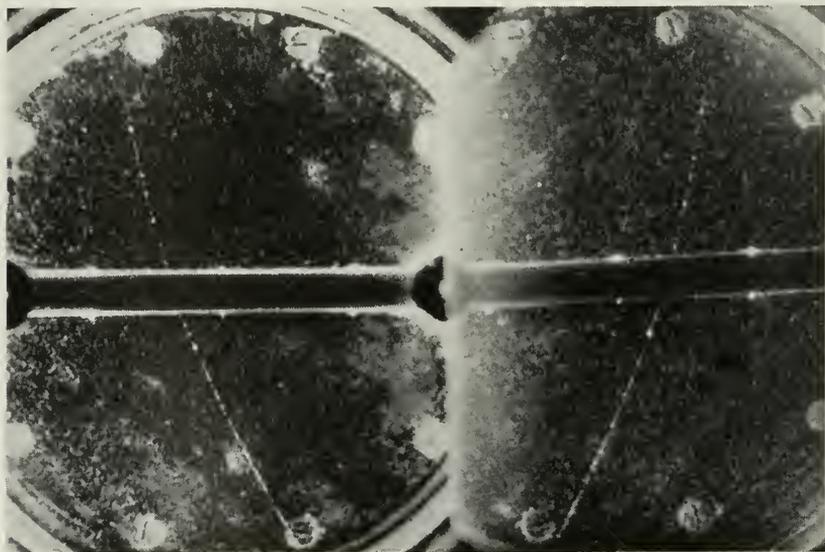


Fig. 9—Track exhibiting measurable and unequal curvatures on the two sides of a metal plate, thus indicating changes of energy and momentum suffered in the traversal. (Anderson)

masses a given $H\rho$ -value leads to different energy-values, but also that the error due to an incorrect estimate of rest-mass becomes proportionately smaller as the $H\rho$ -value increases. Yet the possibility of the error is always there, if the mass of the particle is not certainly known; and it affects many published "energy-values"

based on the presumption—often admitted in the context to be more than doubtful—that the particles to which they refer are electrons. The danger might be mitigated by describing these as “quasi-energy-values” expressed in “quasi-Mev.”—For actual electrons with momenta as great as those figuring in the cosmic rays, the energy-value in electron-volts is practically equal to 300 times the $H\rho$ -value expressed in gauss-centimetres.

Many a cosmic-ray particle suffers no deflection that can be detected in its entire course across a Wilson chamber (diameter, 15 cm. or even more) in a magnetic field as strong as can be applied over so great a volume (field-strength, 20,000 gauss or thereabouts). One might well be tempted to think such a particle chargeless, for if this were the case, the field would have no grasp at all upon it; but if it were chargeless it could not ionize the molecules of the gas and therefore could not form the chain of ions on which the droplets are founded. In some of the finest of the experiments (those in Pasadena and those in Paris) a detectable curvature of the track would be shown if this were made by an electron of energy so enormous as $2 \cdot 10^{10}$ electron-volts (20,000 Mev!). The uncurved tracks accordingly speak of electrons of energies greater than 20,000 Mev, if these particles are electrons; and the inference is not much less drastic, if they are more massive than an electron.

We, however, are more interested, for the present, in the tracks which are sensibly curved; and most of all, in the tracks which are intersected by a metal plate and which show a curvature on one side of the plate and a larger curvature on the other (Figure 9). From the two ρ -values one can deduce the momentum-loss Δp and the energy-loss ΔE suffered by the particle in passing through the plate. (Yet I emphasize again that Δp is computable only if the charge is correctly guessed, and ΔE only if the rest-mass is correctly guessed in addition to the charge.) With this ambition Anderson inserted such plates for the first time into a Wilson chamber, in 1931. The idea had a wonderful and unforeseen result, some years ago recounted in these pages. Notice that above I spoke of the momentum-loss and the energy-loss suffered by a particle in going through a plate. In so doing I was making the assumption that it is a loss and not a gain which happens. If this highly plausible assumption is correct, then the sense in which the particle is traveling its path is knowable; it is from the side of the plate on which the curvature is less, to the side on which the curvature is greater. If the sense of the motion is knowable, so also the sign of the charge of the particle is knowable, being positive or negative according as the track is bent with its

concavity toward the left or toward the right of an observer looking into the chamber from the north-seeking pole of his magnet. Without the plate, neither sense nor sign would be knowable except in the rarest of cases.³ Anderson in August 1932 found on one of his photographs the track of a particle which by this criterion was positive, and which by the density of droplets along its track (we take up this topic later) he identified as an electron. He thus became the discoverer of the positive electron.

Concentrating on the measuring of ΔE after the excitement of the positive electron had subsided, Anderson presently found that its values are very fluctuating. Thus in 1934 he published the details of nine traversals, made by particles assumed to be electrons, through thicknesses of lead from 7 to 15 mm. (Even with a single metal plate the effective thickness varies, since corpuscles traverse the plate with varying degrees of obliqueness.) These were by no means identical in initial energy, this ranging from 38 to 240 Mev; nevertheless one might have expected the energy-loss per unit length of path in lead to be about the same for all, and yet the nine values thereof were scattered all the way from 18 to 120 Mev/cm! Such fluctuations suggest that the energy is lost in great amounts at a few events, and not in dribbles at many. They did not deter Anderson and Neddermeyer from making such measurements on hundreds of later particles, classifying the particles into groups according to their energy-values, and averaging the energy-losses within each group. What then was found has a bearing upon the problem; but we pass over it for the time being, and consider in Fig. 8 the record of ninety-four particles which, during a later experiment, passed through a plate of platinum one centimetre thick.¹

Plotted horizontally are the energy-values of the particles while above the plate, vertically the energy-changes divided by the lengths of path in the platinum. The axis of abscissæ is the locus of energy-losses imperceptibly small; the line slanting at 45° is the locus of energy-losses which are total, the particles shown on this line having been stopped by the plate. The fact that some of the representative points lie below the horizontal axis means only that for every particle the observers subtracted its energy below the plate from its energy above, irrespective of its direction of motion. Suppose that these

³ One might be misled by the adjective "cosmic" into believing that all cosmic-ray particles come from above, their sense of motion making an angle of less than 90° with the downward-pointing vertical. Many, however, including Anderson's first positive electron, have been found by this criterion to be moving upward (i.e. at more than 90° to the downward-pointing vertical). The showers of Figs. 6 and 7 show that this is not a forced interpretation.

¹ I am indebted to Dr. Anderson for a plate exhibiting data thus far unpublished.

subjacent points correspond to upward-going corpuscles, and transfer them across the horizontal axis. Then, the sprinkling of points extends all the way from axis to slanting line; and this is the sign of fluctuations such as Anderson from the start had observed. Notice however that the representative points are of four aspects: solid dots and hollow circles, with or without downward-pointing barbs. The dots refer to tracks which were seen in the chamber singly; the circles, to particles which "entered the chamber accompanied by other particles." The lonely particles are prevailingly able to pass through matter without suffering energy-losses nearly so great as those which the others incur! Thus by itself and without any theory, Fig. 8 establishes a distinction between the singly-appearing corpuscles on the one hand, and those which appear in company on the other. Moreover the barbs are often attached to the hollow circles, bearing out the inference from Figs. 5 and 7 that shower particles are likely to be shower-producing particles; but rarely are they attached to solid dots, never to those which lie far off from the slanting line.

(This seems the best place for mention of the similar work now being done in England by Blackett and (J. G.) Wilson, in France by Ehrenfest. The Englishmen have set plates of gold, lead, copper and aluminium, of various thicknesses from 3.3 mm to 2 cm, into the middle of an expansion-chamber in Anderson's fashion; Ehrenfest, using a pair of cloud-chambers one over the other, was able to put between them a block of gold no less than 9 cm thick! Their way of reducing their data for plotting is not the same as that employed at Pasadena, and their diagrams therefore look very different¹ from Fig. 8. Their energy-range runs much further upward, as far as 5000 Mev, and the great majority of the particles which they plot lie beyond the limit of Fig. 8. Many of Ehrenfest's particles got through the great thickness of gold without losing anywhere nearly the whole of their energy, and are therefore to be classed as much more penetrating than electrons should be. So did nearly all of the particles of energy greater than 250 Mev observed in England, but there were a few of these which lost most of their energy in 0.33 cm of lead, and of these few about half seemed to belong to showers.

¹For the benefit of those who may consult the original papers, I give the difference. Let E_1 and E_2 stand for the (quasi) energy-values of a particle before and after passing through a thickness d of metal; ΔE for $(E_1 - E_2)$; x for $\frac{1}{2}(E_1 + E_2)$. What is plotted by Anderson and Neddermeyer (Figure 8) is $\Delta E/d$ as ordinate and E_1 as abscissa. Blackett (in all his papers but the earliest), Wilson and Ehrenfest begin by subtracting from ΔE a quantity sd which is supposed to be the amount of energy spent by the particle in detaching electrons from atoms while traversing the metal (Blackett assigns the value 15 Mev/cm to s in lead, Ehrenfest takes 28 for gold); they then plot $(\Delta E - sd)/xd$ as ordinate and x as abscissa. Their ordinate (denoted by them as R) is then more nearly ready for comparison with theory.

At energy-values below 200 Mev Blackett finds almost no penetrating particles, a singular contrast with the Pasadena observations; he suspects that the penetrating particles become ordinary electrons when they are slowed down into this energy-range. I mention also the measurements made on some twenty penetrating corpuscles by Leprince-Ringuet and Crussard, leading to the exceptional conclusion that positives suffer smaller energy-losses than negatives.)

But granting that there are two sorts of particle with a right to different names: has either a right to the name "electron"? To settle this question, and for several other reasons, it is time to call upon theory.

It is now some thirty years since there entered into physics a German word, *Bremsstrahlung*, which can be translated literally into English as "braking radiation," and would no doubt be so translated if "braking" did not sound like another English word of entirely different meaning. This is chiefly observed emerging from X-ray tubes, being emitted from their metallic targets when these are struck by the stream of bombarding electrons. It consists of photons or corpuscles of light, each containing at least a part of the kinetic energy of one of the incident electrons. The distribution-in-energy of the photons makes it clear that the electrons frequently lose large fractions of their initial energy *en bloc*, throwing it off in individual parcels which are these photons (indeed it sometimes happens that the entire kinetic energy of an incident electron is shed in the form of a single corpuscle of light). This radiation forms the so-called "continuous X-ray spectrum" or "X-ray continuum" emerging from targets of X-ray tubes. With the spectrum-lines which are sometimes seen superposed on this continuum we have nothing here to do.

By the classical theory of thirty years ago this continuous spectrum is attributed to the slowing-down of the electrons as they penetrate into the metal, whence the name *Bremsstrahlung*. By the quantal theory of today it is still ascribed to the slowing-down, which must now be conceived as taking place in instantaneous jerks, occurring probably in the close vicinity of atom nuclei. At each of the jerks, the electron-speed is suddenly reduced and the kinetic energy goes forth in the form of light. The later theory in its quantitative form gives a competent account of the continuous X-ray spectrum as it springs from the tubes of the laboratory, with their bombarding electron-streams energized by voltages of a few tens or hundreds of thousands. For a long time nobody seemingly troubled to extend it to voltages of the order of thousands of millions; a futile extension indeed this would have been, so far as X-ray tubes are concerned.

When finally the extension was made by people interested in the cosmic rays, it turned out that according to the quantal theory the liability of electrons to these "radiative energy-losses" goes up so greatly with increasing speed, that electrons of even the cosmic-ray energies should not be able to bore their way through as much as five centimetres of lead!

After the meaning of this inference sank in, there ensued a period lasting for months (in 1935 and 1936) in which several eminent theorists were willing to concede that Nature must have set a limit to the scope of quantal theory. It was beginning to be believed that somewhere between the energy-range attainable in the laboratory and the energy-range manifest in the cosmic rays, there is a critical energy-value beyond which the electron escapes from the sway of the quantal laws, and is exempted from losing its energy by the process of *Bremsstrahlung*. This belief was an artifice for permitting the penetrative particles of the cosmic rays to be called by the name of electron. It might have remained a credible artifice, if the penetrative particles had been the only ones—if, that is to say, there had never been any evidence for the existence of particles among the cosmic rays having the properties required of electrons by the quantal theory. Such a situation may have seemed to exist at the time when the belief was dominant. It exists no longer, as the description of Fig. 8 has just suggested; but before considering further the data, I must introduce something more of what the theory has to say.

Since 1934 it has been known that a photon of energy greater than about one million electron-volts is capable, when in the vicinity of an atom-nucleus, of converting itself into a pair of electrons of opposite sign. About one million electron-volts—1.02 Mev, to be somewhat more precise—becomes "rest-energy" of the twin electrons, being incorporated with their rest-masses; the remainder ($h\nu - 1.02$, if by $h\nu$ we denote the photon-energy in Mev) becomes kinetic energy of the electrons. The process may be produced at command and exhibited to the eye, by projecting the photons known as gamma-rays against metal targets contained in expansion-chambers. The gamma-rays originally used for this purpose proceeded from natural radioactive substances; mostly they were those emitted by a certain substance (thorium C'') with a photon-energy of 2.62 Mev. Nowadays gamma-rays of energy several times as great can be produced by effecting certain transmutations, in the course of which (or afterward) they emerge from the new-born nuclei. Figure 10 shows an admirable example of an electron-pair formed out of such a photon. Moreover, the converse process is well-known: positive electrons falling against

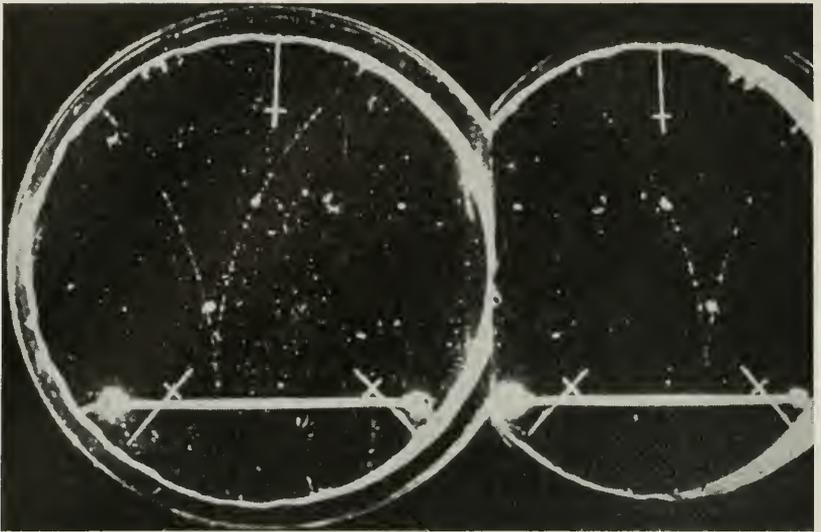


Fig. 10—An electron-pair born from a photon. (W. A. Fowler, California Institute of Technology)

a plate of dense matter bring about the emission of photons of energy 0.51 Mev, and these are just what are to be expected if the positive electrons (after being slowed down) unite with some of the innumerable negative electrons already in the plate and produce, at every such union, a pair of equal photons.⁴ Much too abundant to be here described is the evidence for the ability of electron-pairs to pass into light and light to pass into electron-pairs, making it permissible to imagine a continual alternation of energy between these two so sharply contrasted forms.

Formation of the photons of *Bremsstrahlung* by electrons of enormous energy, and formation of electron-pairs out of such photons: these reciprocal processes engaged the attention of several theorists (Bethe, Heitler, Sauter, Weiszaecker, Oppenheimer) in the years 1933 and 1934. The problem was, to evaluate by quantal theory the chance that electron or photon would spend its energy in producing photon or electron-pair, while traversing given thickness of given element.

⁴Evidently this is not quite the converse of the process previously described, which if reversed would consist in the merger of a positive and a negative electron with the formation of a single photon bearing away all of their energy. Some evidence exists for the occurrence of this process. There is no sign of the fourth conceivable process (the meeting and merger of two photons to form two electrons) which must obviously be very rare in practice owing to the feeble concentration of photons in actual beams of gamma-rays. Nevertheless this last is the process first predicted by the theorist Dirac.

Approximations had to be made in the calculation, as nearly always in quantal problems; but they are supposed not to affect the rightness of the main result. To quote Oppenheimer's description of this result: "a beam of high-energy electrons should have a good part of its energy converted into photons in a centimetre of lead; in an equal distance these photons will be largely reconverted into pairs."

Such was the result from which, in 1935, it was inferred that quantal theory must be wrong because it was predicting something which could not be found in Nature; and from which, in 1936 and thereafter, it was concluded that quantal theory not only was correct but had made a splendid triumph, in explaining the phenomena of showers! It is not altogether clear why the later conclusion was not drawn at the start; perhaps the reason is, that as lately as the summer of 1936 fine photographs of showers were still rather rare, while such pictures as Figs. 5 and 7 with their examples of self-augmenting showers had not as yet been made. On the other hand it would be premature to say and misleading to imply that the process which the theory describes is in exact and quantitative accord with the observations on showers. There are at any rate good grounds for hoping that as the mathematics of the theory is more fully worked out and the art of the experiments refined, the agreement will grow better and better. The most that seems safe to say is, that now we have a general scheme for the interpretation of showers of a certain type, and a very hopeful prospect that this general scheme will be converted into a detailed and quantitative explanation as the mathematics of the theory on the one hand, the aptness and precision of the observations on the other hand are gradually improved.

By inserting the words "of a certain type" in the foregoing sentence, I leave open the possibility that showers may be classified into more than one type, and all of these but one be ascribed to other processes. This is no mere possibility but already almost a certainty. Certain showers which include "heavy tracks" due to protons or still more massive particles are ascribed to nuclear explosions provoked by cosmic rays. If a shower fails to undergo the "multiplication" illustrated in Figs. 5 and 7, it is taken as belonging to this other type. Exception made for such cases, it is strongly plausible to say that shower particles and shower-producing particles are electrons; that accordingly high-energy electrons exist among the cosmic rays, behaving as the quantal theory says that they should; and that consequently the other particles, setting themselves apart from electrons by their penetrative power and their failure to make showers, are of another sort.

Ability to penetrate matter, inability⁵ to make showers: these are the complementary aspects of the property which distinguishes this other type of particle, the mesotron. If one wishes to contrive a particle having this property and differing otherwise as little as possible from the electron, how must it be done? The electron has the qualities of charge and mass; also those of spin and magnetic moment, but these are considered (perhaps wrongly) to be little or not at all concerned with shower-production. If we imagine the mass to be increased while the charge remains the same, the liability to *Bremsstrahlung* will diminish; for *Bremsstrahlung* occurs when sudden sharp deflections or decelerations occur, and these are less sharp and sudden the more massive the particle is. Now *Bremsstrahlung* is the prelude to the entire manifold process of the forming of a shower, and hence a mere increase in the mass of the hypothetical particle leads in the desired direction. The theory indicates that a particle with the electronic charge and a few dozen times the electronic mass will be penetrating enough. We do not need, however, to be contented with such vague intimations, for there is yet another phenomenon in respect of which the mesotron differs from the electron, and from this the mass can be deduced more sharply.

So far, we have been considering the passages of particles through solids. There, the paths are concealed, the adventures of the particles can only be inferred—from the difference between energy before and energy after traversal, or from the photons and the secondary electrons which are driven out of the solid. Now we are to consider the passages of charged particles through the gas of the Wilson chamber, which, unlike the scriptural way of the eagle through the air, are preserved for our inspection by the droplets. Figure 1 has shown to us a track in which the number of droplets in unit length of path can rather readily be counted. What does this number signify? And is it truly an indication of the mass of the traveling particle, as I hinted on an early page?

The latter question might perhaps be sufficiently answered without reference to the former; but for completeness, and for the sake of its own interest, the former ought to be treated more fully than it was in that brief earlier mention. In the voyage recorded in Fig. 1, nothing so drastic happened to the traversing particle as would have been the losing of a large part of its energy in the form of a photon of *Bremsstrahlung*. It lost its energy in dribbles, spent in detaching electrons from molecules and giving them a small extra bonus of kinetic energy

⁵ It is better to say "relative inability" since occasional showers are attributed to mesotrons, which perhaps operate by making a violent impact on an electron and so giving it the energy needful for starting the process.

with which to go wandering around in the gas. They had not speed enough to wander far, even in the half-a-second afforded them before the condensation. Probably they had already adhered to molecules before the condensing water immobilized them. One speaks of the droplets as being condensed partly on negative, partly on positive ions; the last-named are the molecules from which the electrons were ref. (If, during the half-a-second, an electric field of suitable strength is applied, the ions of the two signs drift in opposite ways, and when the water-vapor comes down there are seen two parallel trails of droplets with an empty space between.)

The simplest idea is that the traversing particle tears off one electron from each of many molecules through or near which it passes, and that half of the droplets are formed on these electrons and the other half upon the molecules bereft. This is too simple to be true. It is likely that sometimes the particle removes two electrons or more from a single molecule, so that there will be more negative ions than positive. Much more serious is the certain fact that often when an electron is thus released by the direct action of the traversing particle, it shoots away with speed and energy enough to enable it to release one or several more from neighboring molecules. Now and then one comes on a cloud-chamber photograph in which there appears a track with branches (Fig. 11); each of these is the trail of an electron which



Fig. 11—Tracks of a charged particle bristling with short branching tracks, made by electrons ejected from atoms with energy sufficient to ionize. (Auger)

has received a truly abnormal and extraordinary amount of energy. Much commoner, in fact universal, is the "beaded" appearance of such trails as appear in most of the pictures of this article: it is presumed that each of the beads is an unresolved cluster of droplets formed on a cluster of ions, all but one pair of them made in the indicated way. Occasionally one sees a picture in which the interval allowed for diffusion has been so happily chosen that the droplets in the clusters are far enough apart for counting, and yet consecutive clusters do not overlap. In making Fig. 1 the interval allowed was a little too long, and yet perhaps it is possible to think that the ions are denser in some parts of the trail than in others, as though they had been formed in clusters which have broadened almost but not quite to the point of losing their identity.

It is therefore necessary to distinguish, in mind if not in fact, between the "primary ionization" consisting of the electrons and the molecules torn apart from each other by the direct immediate action of the traversing particle,⁶ and the "entire ionization" (sometimes called "probable ionization") consisting of these together with all the ions formed by the directly-ejected electrons. Under ideal conditions it is presumed that the measure of the former would be the total number of droplet-clusters,⁷ the measure of the latter would be the total number of droplets, in unit length of path. Not many physicists have tried to evaluate both of these numbers. Of those who have, the data have been scanty, but the consensus of opinion is that the latter is about or not quite twice as great as the former. It is, however likely that the value of the ratio of the two is not important when one wants only to distinguish between electron and mesotron, as we shall presently see.

The problem of the primary ionization is one of the major tasks of theoretical physics. Classical and quantal theorists alike have spent great labor on the question: given a charged particle of specified charge and mass and speed traversing air (or any other gas), how many electrons will it set free from the molecules in unit length of path? At this point I will give only one of the results—or rather, something which is not a result at all, but a part of the assumptions. It is assumed that as the traversing charged particle flies along through or close to a molecule, it operates upon the electrons thereof by virtue of the ordinary electric forces between its charge and the charges of the electrons. It follows, then, that *whatever expression finally may be derived for the primary ionization must depend only upon the charge and the speed of the traversing particle, and not upon its mass.* (Mass and momentum of the particle must indeed be great enough to hold it on a sensibly straight course as it plows onward through the gas, despite its losses of energy as it detaches electrons; but this condition is always realized, with the corpuscles of the cosmic rays.)

I seem to have said that the primary ionization gives no power of distinguishing between an electron on the one hand, a particle of equal charge and different mass on the other. However, it *does* confer on us this power, for the reason that the curvature of a particle-track in a known magnetic field is a measure not of particle-speed but of

⁶ Unluckily called "secondary ionization" by some of the German theorists.

⁷ Best to observe the droplet clusters as individual entities, one would wish the expansion to occur before the ions have any time at all to diffuse. To attain this, Williams and Pickup caused the chamber to expand at moments taken at random, and trusted to luck for the appearance of cosmic-ray tracks formed at just the right instants. Luck served them with no fewer than four tracks betokening particles of a distinctive mass.

particle-momentum (equation 2). If by luck an experimenter should happen upon two tracks having the same curvature but made by particles having masses⁸ standing to one another in the ratio (say) 100 : 1, the speeds would stand to one another in the ratio 1 : 100, and this might well entail a perceptible difference in the primary ionization. It would come to the same thing, if someone should take the data for a large number of tracks, and plot primary ionization as function of curvature: if there are really two kinds of particle differing in mass, there should be two sets of points lying along two curves, and from the ordinates of these curves at any abscissa the ratio of the masses would be derivable.

Perhaps the last sentence suggests that someone already has made this correlation, and has found that the points for all of the single or penetrating particles lie upon one curve, and all the points for shower-particles and shower-producing particles lie on another. This has not been done. The reason is, that many of the penetrating particles exhibit no perceptible curvature of track at all, and most of the others a very small curvature. The former are moving so fast that their momentum cannot even be estimated, except as being beyond a certain critical value. As for the latter, the speeds of even these are so great as to approach the speed of light; for a given momentum-value the speed varies only a little with the mass, and the primary ionization varies too little to serve as an index of mass. To make a profitable correlation, one must use only the particles of which the tracks are notably curved. Nearly all of these are shower-particles, which already are presumed to be electrons. To find a penetrating particle with a highly-curved track, one must find it when it is near to the end of its course and its energy wellnigh gone. Such is the principle which directed some of the recent successful searches for particles proclaiming themselves by their ionization to be more massive than electrons.

Before looking at the track of one of these particles, we ought to notice a couple of questions concerning ionization. One of them is: is the distinction between primary and entire ionization—or rather, our lack of perfect ability to make it in practice—likely to lead to trouble? Many observers are far from clear in reporting whether what they observe is more like the one or more like the other; but it seems probable that the second like the first is dependent only upon the speed and the charge of the traversing particle, not on the mass thereof; and this diminishes the dangers from confusing the two. The question is implicated with the second: to what extent do experi-

⁸ Allowance being made for the relativistic dependence of mass on speed.

ment and theory aid us in identifying the shower-particles with the electrons? As to experiment, there exist the records of a few studies made by the Wilson chamber upon particles acknowledged to be electrons, of energy-values ranging from about 2 Mev downward to some 25000 electron-volts. In respect of the trend with energy, they agree fairly well with the assertions of the quantal theory; but when one inquires whether the absolute value for the number of clusters of ions in unit length agrees with the absolute value of the quantal expression for the primary ionization at any particular energy, one is confronted with the fact that the quantal expression contains a multiplying factor which depends on intimate details of the structure of the molecule, and is not exactly known. The quantal theory, however, predicts a minimum in the curve of primary ionization vs. energy, at an energy of about 2 Mev. Such a minimum (Fig. 12) was

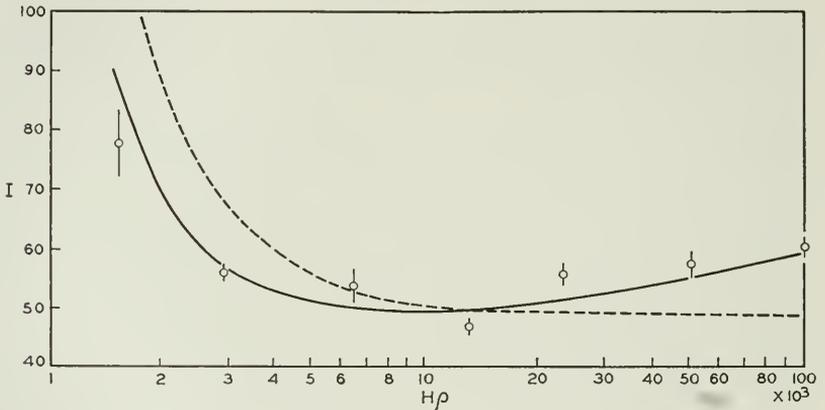


Fig. 12—Ionization-density (entire) along the tracks of cosmic-ray particles, plotted as function of $H\rho$. The continuous curve is that of a theoretical function containing a multiplying factor which has been adjusted to get the best fit to the data. (Corson and Brode)

actually found by Corson and Brode in their study of some fifty particles of the cosmic rays, and probably is to be ranked as evidence for the electronic nature of these particles quite as forcible, as would be an absolute agreement between the observed ionization and the predictions of a reliable theory.

Street and Stevenson, with a row of counters and an interposed cloud-chamber such as appeared in Fig. 3, adjusted their counters in such a way that the chamber expanded only when the counters above the chamber had simultaneous discharges and the counter below did *not*. A thousand photographs yielded to them the track

of one particle having a notable curvature and displaying an ionization six times as great as that attributable to an electron; they inferred a "mass 130," *i.e.* a rest-mass one hundred and thirty times as great as that of an electron. Neddermeyer and Anderson transposed the bottommost counter into the very centre of the cloud-chamber itself,

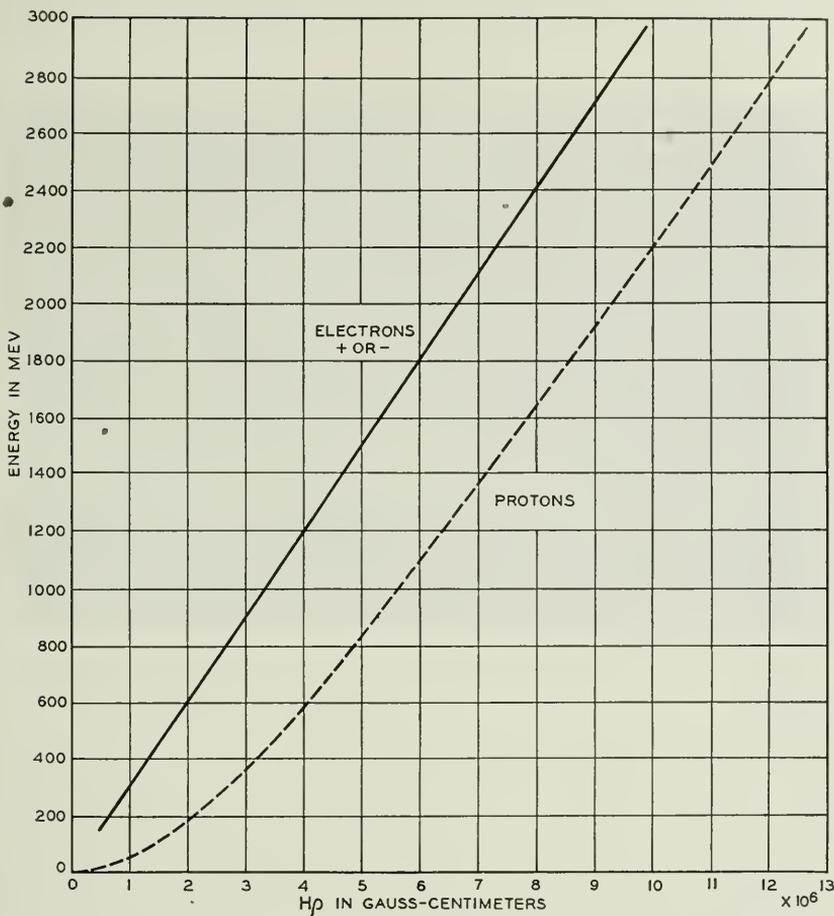


Fig. 13—Relation between energy and $H\rho$ -value for electrons (of either sign) and protons. (Anderson)

and there it appears in Fig. 14, neatly intersected by the course of a particle which above it made a track lightly curved and thinly studded with droplets, and beneath it made a track sharply curved and densely congested. Comparing ionization with curvature along the track above and the track below, they found 240 to be a satisfactory ratio

of the mass of the traversing particle to the electron-mass. Williams and Pickup, to whose technique I have already alluded (footnote 7 on page 210), observed four tracks of which three were compatible with a rest-mass of about 200, the remaining one requiring a mass-value between 430 and 800. A few more such tracks have appeared in the literature, but instead of describing them I turn for the climax to another and an exacter way in which Fig. 14 furnishes the desired value of mass.

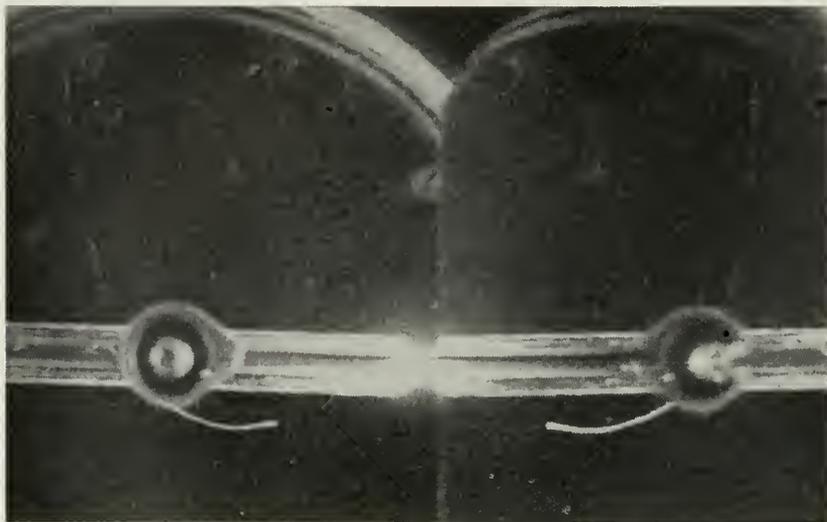


Fig. 14—Track of a mesotron slowed down by an obstacle in a Wilson chamber and finally brought to a stop in the gas of the chamber itself. (Neddermeyer and Anderson)

In Fig. 14, the track beneath the counter comes to a sudden end. One could take a sheet of coordinate-paper, and plot along the horizontal axis the curvature of the path as it emerges from the counter, and along the vertical axis the length of the path from that point of emergence onward to its end. This would give a single point of what is known as a "range-vs.-curvature relation" or a "range-vs.-momentum" relation. A second point can be found by measuring the thickness of the glass counter-wall twice traversed by the particle, converting it into an equivalent thickness of gas, adding this to the length of the path beneath the counter, and correlating the sum with the curvature of the path at the point where the particle enters the counter. Now, range-vs.-curvature relations are among the best-studied of the features of the charged particles already known—

electrons, protons, alpha-particles. These two points pertaining to the particle of Fig. 14 lie far from the curves appropriate to any of the three. An electron departing from the counter in a path of such a curvature as there is shown would have traveled 2000 times as far before reaching the end of its course! a proton, on the other hand, only one seventy-fifth as far! This at the moment is deemed the sharpest and most clear-cut evidence for the existence of a particle intermediate in mass between proton and electron, to which Anderson now assigns a mass of $220 (\pm 35)$ times the electron-mass.⁹

It is fitting to end this article by mention of several other kinds of evidence which have bearing on the question of the mesotron; mainly they are relatively indirect, and would require much space to describe and assess. Inferences have been drawn from the number of electrons ejected with high energy from metal plates by penetrating particles traversing these: J. G. Wilson derives a mass-value greater than 100. A curious inference has been drawn from the deflections suffered by these particles in traversing metals: the magnitude of these should by theory be independent of the mass of the particle—since it *does* appear to be the same for penetrating particles as for electrons, it is deduced that the mesotron and the electron can differ only in mass. Inferences have been drawn from the trend of cosmic-ray intensity with elevation in the atmosphere, and from the trend of cosmic-ray intensity beneath metal screens as function of the material and thickness of these last (it was thus that Auger as early as 1934 was led to suspect the existence of two kinds of charged particle among the rays).

Inferences have also been drawn from nuclear theory. To enter adequately into this difficult field is impossible here: it must suffice to say that Yukawa conceived, as a constituent of nuclear structure, of a particle possessing the charge of an electron and a mass of about the magnitude which the mesotron appears to have, and possessing in addition the quantity of *instability*. The "Yukawa particle," that is to say, has the qualities demanded of the mesotron, and in addition is liable to emit an electron; what is left behind is then a neutral particle which could elude observation. The emission is expected to follow the law familiar in radioactivity, the durations of individual Yukawa particles being distributed according to the law of chance about a mean value. Is there evidence that the mesotron behaves in this way?

⁹ Values diverging from this by more than the estimated uncertainties have been published by other observers of other particles, and may betoken an underestimate of the uncertainty or the existence of particles of several masses. A "nomograph" for facilitating the evaluation of mass from curvature of path combined with ionization-density or range is given by Corson and Brode.

For this there is some evidence, of the following kinds. First let us compare (in imagination) the number (per unit time per unit area) of penetrating particles flying vertically downward and the number flying obliquely downward. The comparison can be readily made with such an apparatus as that sketched in Fig. 3, the cloud-chamber being superfluous and the lead absorber reduced to the least thickness sufficient to stop electrons; the axis is oriented first at 90° and then at various lesser angles θ to the horizontal plane. Even the whole of the atmosphere is insufficient to stop such mesotrons as the cloud-chamber discloses; and yet the observations show a marked decline of the number thereof as θ decreases. But the particles which travel obliquely traverse a greater distance from the top of the atmosphere than those which come vertically down, and take a longer time in doing so; the decline of number with decrease of θ may therefore be ascribed to the perishing of the mesotrons *en route* to the apparatus as the route grows longer and longer. Second: Let us compare the effect of the obliquely-traversed atmosphere with that of a sheet of lead in cutting down the number of particles arriving at the apparatus. One must make a guess as to the thickness of lead which would be required to produce a falling-off of the number of particles equivalent to that observed in the atmosphere, if the falling off were due to actual stopping of mesotrons in air and lead respectively, and the impermanence of the mesotron did not enter in at all. It is commonly conjectured that the equivalent thicknesses of lead and air would stand to one another inversely as the densities of these materials. When, however, the effects of such "equivalent" thicknesses are compared, it is found that the falling-off beyond the lead is decidedly less than that beyond the air. Now the mesotrons take very much less time for traversing the sheet of lead than the wide expanses of the atmosphere; and the "anomaly," as it has been called, is tentatively explained by assuming that few of them perish in the lead, many in the long journey through the atmosphere.

Estimates of the mean life of the mesotron thus made yield values of the order of a millionth of a second. It is supposed by many that the mesotrons are born in the upper layers of the atmosphere. Such conjectures, however, lead beyond the scope of this article, which must be confined to these few recent fruits of the seemingly exhaustless cornucopia of the cosmic rays.

SELECTIONS FROM THE LITERATURE

ENERGY-LOSSES OF PARTICLES TRAVERSING METALS: Anderson and Neddermeyer, *London Conference on Nuclear Physics* (1934); *Phys. Rev.* **50**, 263 (1936); Neddermeyer and Anderson, *Phys. Rev.* **51**, 884 (1937); Blackett and Wilson, *Proc. Roy. Soc.*

160, 304 (1937); Blackett, *ibid.* **165**, 11 (1938); Wilson, *ibid.* **166**, 482 (1938); Leprince-Ringuet and Crussard, *Comptes Rendus* **204**, 112, 240 (1937); Ehrenfest, *Comptes Rendus* **207**, 573 (1938).

PENETRATING PARTICLES DETECTED WITH COUNTERS, WITH OR WITHOUT CLOUD-CHAMBERS: Street, Woodward and Stevenson, *Phys. Rev.* **47**, 891 (1935); Street and Stevenson, *Phys. Rev.* **51**, 1005 (1937); Auger and Ehrenfest, *Comptes Rendus* **199**, 1609 (1934).

TRACKS OF PARTICLES WITH CHARACTERISTIC IONIZATION DENSITIES: Electrons: Corson and Brode, *Phys. Rev.* **53**, 773 (1938). Mesotrons: Street and Stevenson, *Phys. Rev.* **52**, 1003 (1937); Ehrenfest, *Comptes Rendus*, **206**, 428 (1938); (E. J.) Williams and Pickup, *Nature* **141**, 684 (1938); Maier-Leibnitz, *Naturwiss.* **26**, 677 (1938); Neddermeyer and Anderson, *Phys. Rev.* **54**, 88 (1938), and literature there cited.

THEORY OF SHOWERS: Oppenheimer, *Phys. Rev.* **50**, 389 (1936); Carlson and Oppenheimer, *ibid.* **51**, 220 (1937); Bhabha and Heitler, *Nature* **138**, 401 (1936), *Proc. Roy. Soc.* **159**, 432 (1937); Bhabha, *Proc. Roy. Soc.* **164**, 257 (1938); Montgomery and Montgomery, *Phys. Rev.* **53**, 955 (1938).

DEFLECTIONS OF PARTICLES: Blackett and Wilson, *Proc. Roy. Soc.* **165**, 209 (1938).

ELECTRONS RECOILING FROM IMPACTS OF MESOTRONS: Wilson, *Nature* **142**, 73 (1938).

INSTABILITY OF MESOTRON: Blackett, *Nature* **142**, 992 (1938); Rossi, *ibid.* 993; (T. H.) Johnson and Pomerantz, *Phys. Rev.* **55**, 104 (1939).

GENERAL REVIEWS: Euler and Heisenberg, *Ergebnisse d. exakten Naturwiss.* **17**, 1 (1938); Froman and Stearns, *Rev. Mod. Phys.* **10**, 133 (1938).

Hurricane and Flood—September 1938

By W. H. HARRISON

Editor's note: The following was presented by Mr. Harrison as the closing address of a symposium on the effects of the hurricane and floods of September 21, 1938 on transportation, power and communication utilities. The symposium was held in New York at the Winter Convention of the American Institute of Electrical Engineers, Thursday, January 26, 1939. After the close of the meeting a motion picture on the hurricane prepared by the Bell System for the information of its own employees was shown.

THE experiences of the telephone companies are naturally much the same as those already described. The aftermath tally showed that more than one-half million telephones were put out of service—in the New England States about thirty per cent of the telephones in that area. Through the destruction of toll lines, the storm temporarily cut off telephone communication with the outside from over two hundred towns. The total damage to telephone plant was in the neighborhood of ten million dollars.

The story of restoration—the immediate provision of emergency services—the handling of emergency supplies in unprecedented quantities—the augmenting of forces locally to supplement the normal forces—and the mobilization of forces from other areas—all are replete with engineering interest and are very intriguing, but it would not be appropriate to take the time to tell the story here. A few facts will give you a sketchy idea of the situation.

As to materials:

3,500,000 feet of lead covered cable
54,000,000 feet of paired wire
7,000,000 feet of steel strand for guys
and supporting cables

As to mobilization of forces:

Local construction forces were expanded from 3,000 to 5,000. In addition, 2400 highly skilled linemen, cable splicers and installers and over 600 fully equipped construction trucks and other special motor

vehicles were brought from fourteen other telephone companies as far south as Virginia and as far west as Nebraska and Arkansas.

Of striking significance in the prompt restoration of service was the traditional Bell System background of standardization of materials and methods. This standardization greatly facilitated the collection of large quantities of suitable supplies and made possible maximum effectiveness of the men who came from many parts of the country. The striking effectiveness of these measures is a great tribute to the engineers who long ago by their recognition of the value of standardization laid the broad foundation for this effective work.

In every disaster much is learned with regard to formulating plans and caring for specific situations. Of interest in this specific situation, there had been serious floods in much of this territory in 1936. The experience at that time pointed to certain precautionary measures and we know of no case where these did not prove effective in the present situation. For example, while the water rose five feet above the ground floor level of the main telephone building in Hartford, it was successfully kept out of the building by bulk-heads about the doors and windows, provided since the 1936 flood. Also, at various places where lines had been carried away due to the failure of bridges or other forms of river crossings the restored lines did not fail.

Over and above all of these more or less specific points, which I might say are somewhat routine, lies a broad engineering fundamental vividly illustrated by this whole experience.

Engineers by their work have made a pattern of life which has come to make individuals and communities dependent to a large extent in their day-to-day activities and mode of living, on the proper functioning of the services of power, transportation and communication.

Having done this, they have seen their works fall before the fury of nature—have seen the utter disruption of the organized scheme of life, with all the anguish that goes with such disruption.

It is in the light of this experience that an engineering fundamental of first magnitude presents itself, and one which offers a long range problem that is going to call for nicely balanced judgment, both on the part of the engineer and the management. This fundamental stands out clearly—dependability of service, and specifically the degree to which dependability can soundly and wisely be built into the physical plant.

It is trite to say that dependability is fundamental to good service, that it is of prime consideration in the design, construction and operation of all communication, power and transportation facilities. On the other

hand, it would be foolhardy to assume that any man-made structure could completely withstand the fury of the elements, as typified by this storm.

Consider the circumstances. For four days rain was progressively heavier. It totaled between five and ten inches at many New England points. At some places more than six inches of rain fell in one day. As a result large rivers were brought to flood stage and small brooks and streams became raging destructive torrents. And then came the hurricane—then the seas. Wind velocities as high as from 120 to 180 miles per hour have been reported. Raging flood and tidal waters inundated important sections of many communities. Our services extended over the entire band of the storm and we can definitely trace the relationship of high wind velocities and resultant damage.

Another important circumstance, and bearing particularly on engineering consideration, is that nothing like this had happened in this area since the year 1815.

Obviously, to build plant to be unyielding to the sea and to be hurricane tight against such occurrences at century intervals would be as unsound as to ignore them altogether. Thus a challenge is presented to the engineer, taxing his best judgment. On the one hand, not failing to take every reasonable precaution in the future design of the plant, such as the avoidance of known exposures, the provision of alternate routes, the use of emergency restoration facilities of every conceivable character, adequate emergency operating routines; and on the other hand, not to be led by the tragedy of the storm to recommend extreme construction and operating procedures such as wholesale substitution of underground for aerial plant, which would obviously not be in the public interest.

This, it seems to me, is the broad lesson that we draw from this experience and the challenge presented to the engineer.

It was my good fortune to have been in the midst of the restoration work. It was comforting and inspiring to see how the men and women of all service agencies responded to the call, each presented with a trying problem of his own but ever ready to lend helpful and effective cooperation to those in other utilities, and all motivated with the common objective of maximum service to the community in this period of great distress. I know we in the telephone end could not have done our job had we not had the help of others, including the highway and other public agencies.

The final measure of any man's work is, has it been for mankind? A grateful public has put the mark of approval on the work of the men

and women of the utilities and transportation groups in the stricken area. My admiration for them knows no bounds. Frequently when we fail of expression we turn to the pens of immortals.

Two lines in one of Kipling's poems—"Sons of Martha"—beautifully express the work of these men and women:

"Not as a ladder from earth to Heaven, not as a
witness to any creed,
But simple service simply given to his own kind
in their common need."

A Terrain Clearance Indicator*

By LLOYD ESPENSCHIED and R. C. NEWHOUSE

There is described a radio altimeter that gives continuously on the plane a measurement of the separation between the plane and the earth's surface or projections therefrom. There is projected from the plane and reflected from the earth back to it a very short radio wave, the frequency of which is continuously swung back and forth. The returned wave is thereby made to differ from the outgoing wave in frequency by an amount that is proportional to the echo path; and the difference or "beat" frequency is indicated on a frequency meter calibrated in feet of separation. The paper outlines some of the early efforts in this field, some of the technical problems involved, the theory of the system and the practical experimental results that have been obtained.

INTRODUCTION

THE problem of an altimeter for aviation has engaged the attention of many inventors and experimenters for twenty years or more. As a result, about every conceivable fundamental method of attacking the problem, by the utilization of acoustic or electric phenomena, is disclosed in the art, including the many U. S. patents on the subject.

The familiar aneroid altimeter has reached a high degree of perfection and enables the pilot to maintain level flight at any desired altitude but it gives no clue as to the variation of the elevation of the terrain beneath. The pilot has to know his position at all times and perform a mental calculation, in order to know his height above the ground at any given moment. A number of airplanes have drifted off their normal courses and have crashed on higher ground.

An altimeter based upon the use of a sound echo is subject to two fundamental limitations. The first of these limitations is the extremely high noise level produced by the airplane's motors and propellers, which tends to submerge the relatively weak echo at heights of more than a few hundred feet. The second is that the speed of sound is not enough greater than the speed of airplanes. At a height of one thousand feet approximately two seconds are required for a sound to travel to the ground and return. In this time interval a modern airplane would travel six hundred feet and the clearance may have changed materially.

* Read before the Institute of Aeronautical Sciences at the Chicago meeting, November 19, 1938, and to be printed in the Journal of the Institute.

There is in radio the corresponding phenomenon of an echo, an electric-wave reflection. The velocity of a radio signal is so great that an echo from the earth's surface is almost instantaneous; in fact, the time interval is so small as to give rise to a problem in measuring it. For instance, for heights less than a thousand feet the time to be measured is less than two millionths of one second.

The method used in the present instrument is extremely simple in theory. A radio transmitter is provided on the airplane which sends toward the earth a signal, the *frequency* of which changes at a definite rate with respect to time. The signal is reflected by the earth and returns as an echo after a time delay equal to twice the height, divided by the velocity of propagation. During this interval the frequency of the transmitter has changed and now differs from that of the echo by an amount equal to the product of the rate of change of frequency and the time of transit. The reflected wave is combined in the plane receiver with some of the outgoing wave energy and the difference or "beat" frequency is measured by a frequency meter. Since the reading of the meter is that of the "beat" frequency, it is proportional to the time delay of the echo and, hence, to height and thus can be calibrated directly in feet.

EARLY EFFORTS

The evolution of this method is interesting because it illustrates how one art is built upon another, and also the familiar story of separate inventors arriving at the same answer almost simultaneously, actually somewhat in advance of the existence of instrumentalities having the characteristics required to make the invention practically serviceable.

Many systems employing electromagnetic waves for the purpose of indicating altitudes of an aircraft have been proposed.¹ Among early workers in this field who independently of each other were concerned with methods involving frequency modulated waves were J. O. Bentley² of the General Electric Company; Professor W. L. Everitt³ of Ohio State University and certain students in his department of Electrical Engineering including the junior author⁴ and M. W. Hively; and the senior author.

Under the direction of Professor Everitt, some experimental work on the frequency modulation method, using wire lines, was undertaken in the school year 1928-29. On the basis of this work a grant was made by the Guggenheim Fund for the promotion of aeronautics and an investigation was continued with experimental tests, during the following school year under the auspices of the Ohio State Engineering Experiment Station. The experiments were reported upon in the

bulletins of the Station, and in a graduate thesis ⁵ of the junior author and J. D. Corley.

As early as 1920, the senior author proposed the use of electric wave reflection in railway safety systems ⁶ and entertained the idea of frequency-modulated transmission with beat-tone detection for measuring distance along a track. Radio wave reflection for aircraft altitude determination was considered at times from 1926 to 1930 when a patent application was filed ⁷ for an arrangement similar to that which has been worked out, including the use of a frequency meter to give continuously a visual indication of the altitude.

At that time, however, a really practical terrain clearance indicator could not be built due in large part to the lack of suitable radio instrumentalities. Vacuum tubes capable of operating on frequencies approximately fifty times higher than those generally available were indicated as necessary before a satisfactory system could be built.

A long-range program, however, of vacuum tube development for high frequencies was under way in Bell Telephone Laboratories. This resulted in the production of suitable tubes, and they were described by A. L. Samuel ⁸ to the Institute of Radio Engineers in October, 1937. One of these was capable of providing a stable output of between five and ten watts at a frequency of approximately 500 megacycles, so it became feasible to undertake the development of a practical terrain clearance meter.

The Japanese have been experimenting recently with apparatus operating upon the same basic theory and a paper ⁹ was published in Japanese in 1936. A later paper ¹⁰ was published in English in 1938 by the same author, which describes the apparatus and the results of tests made on the ground over short distances with the equipment at rest.

TECHNICAL PROBLEMS

At the time this development was undertaken a number of questions presented themselves as to what the earth's surface would do to the incident wave in reflecting it. It seemed possible that the signal might be so scattered and broken in reflection by small irregularities that the echo would be more like static than a useful signal.

Even if the reflected signal proved satisfactory over the smoother surfaces, it was hard to predict what would happen when flying over timber land or over very irregular mountainous terrain. There was also the question of what would happen when the surface happened to be that of a city where an airplane flying at 250 to 300 feet per second passes over several buildings and streets with abrupt altitude changes of possibly hundreds of feet several times in the course of one second.

Even with the most directive systems that can be devised, the beam radiated from the airplane is so spread that echoes can be expected to arrive simultaneously from several surfaces, for instance from both the leaves on the trees and the ground between the trees, or from the top of a building and from the adjacent street.

Several problems were anticipated in the apparatus itself. The theory is based upon a frequency-modulated signal free from any amplitude modulation, and it was questioned whether a transmitter could be built to operate on ultra-high frequencies which would be sufficiently free from amplitude modulation, when subjected to the vibration of the airplane, to be satisfactory. Since the receiver utilizes both the direct and reflected signals in making the altitude measurement, it is necessary that some signal be picked up directly from the transmitting antenna but not enough to overload the receiver and thus prevent reception of the echo. It was expected that difficulty would be encountered in sufficiently reducing the direct signal.

After considering all these problems, it was decided that the cheapest and easiest way of determining the answers was to build the apparatus and try it out to see if correct operation could be obtained, first, under the more or less ideal conditions of flying over smooth water and, then, over less favorable surfaces.

Most of the measuring equipment available for radio frequency test work is useless at ultra-high frequencies. Hence, it was necessary to get the system functioning as a whole before any means were available for determining the best adjustment of the radio-frequency parts of the system. Because of the difficulty of providing, while on the ground, an adequate reflector at distances of from a few feet to thousands of feet from the apparatus, it was necessary to install the equipment in an airplane very early in the development and make most of the tests during flights. Nearly a hundred airplane flights were made in one of the Bell Telephone Laboratories' airplanes during the development period of seven months which preceded the public demonstrations made in the United Air Lines Flight Research Airplane.

OPERATION AND THEORY

The fundamental parts of the altimeter in relation to their application are shown in Fig. 1. An ultra-high frequency oscillator is provided, whose frequency is varied up and down by a modulator which consists of a small rotating variable condenser driven by a motor. The oscillator is connected through a coaxial transmission line to a transmitting antenna which is located on one of the lower surfaces of the airplane. The signal is radiated downward by this antenna. A

radio receiver is connected through a similar coaxial line to a second antenna similarly located but arranged in such a way that a minimum of direct signal is received from the transmitting antenna and as much echo as possible from the ground. The direct and reflected signals are

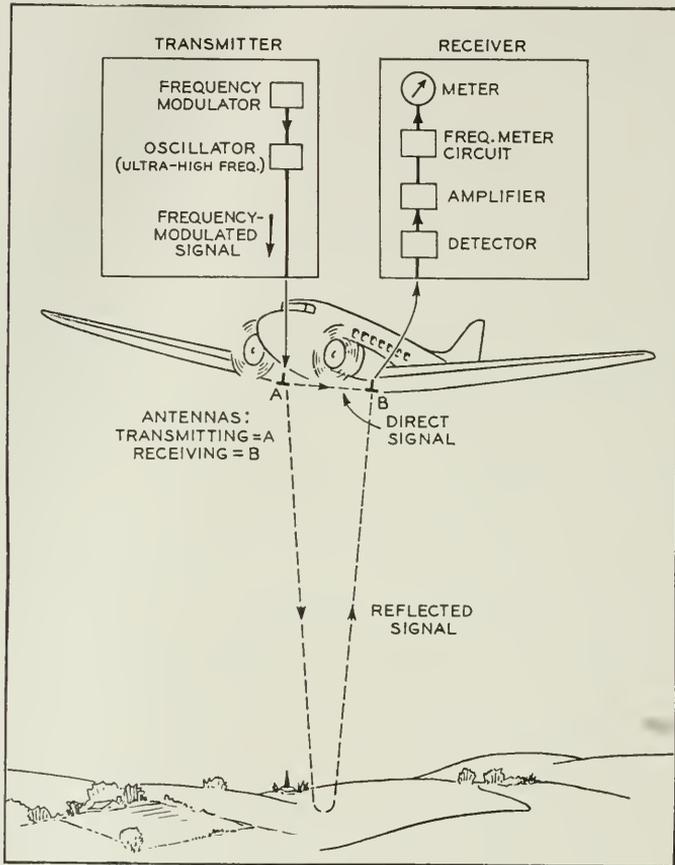


Fig. 1—Overall system.

applied to a detector circuit in the receiver. The output of this detector is a signal of a frequency equal to the instantaneous difference existing between the direct and the reflected signals and is proportional to the height of the plane above the terrain. This signal is amplified by the receiver and applied to a frequency meter or counter circuit which is so designed that a current proportional to the frequency and, hence, to the height flows through a meter calibrated in feet and located on the airplane's instrument panel. A number of types of

indicating frequency meter circuits¹¹ of the condenser charge and discharge variety have been described in the technical literature.

The operation of the system can be understood more easily by reference to Fig. 2. The variation of the transmitter frequency with

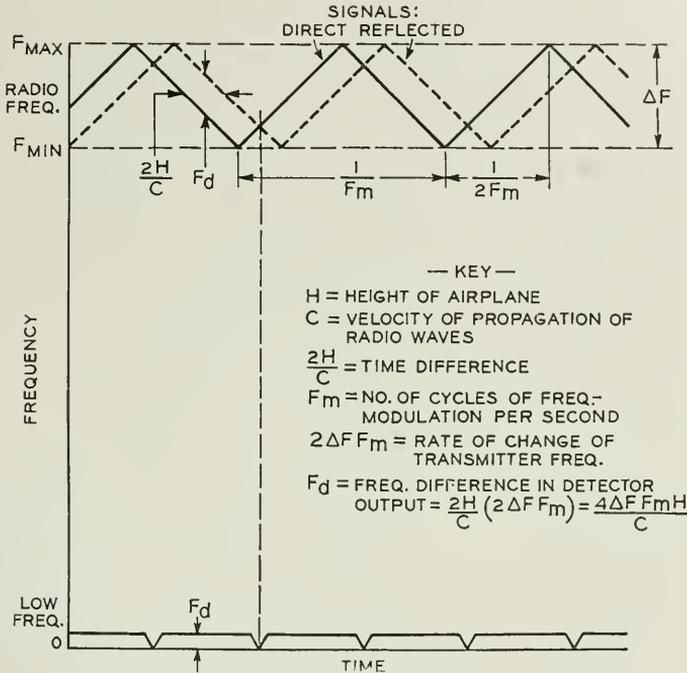


Fig. 2—Operating theory.

time is indicated by the solid sawtooth line.* The value of the ordinate of this curve at any point is the transmitter frequency for the corresponding time. The frequency is varied from $F_{MIN.}$ up to $F_{MAX.}$ and back F_m times per second, so the rate of change of frequency is $2\Delta F F_m$ when ΔF is substituted for $F_{MAX.} - F_{MIN.}$ The linear frequency variation shown, while ideal, is not essential for the successful functioning of the apparatus. The dashed sawtooth line represents the variation with time of the frequency of the echo signal from the earth's surface. This curve is displaced to the right by a time equal to twice the height divided by the velocity of propagation, or, in other words, the time it took the radio signal to go down to the earth and

* A simple harmonic wave that changes in frequency from instant to instant is no longer a single frequency but a series of discrete frequency components. In the present instance, the number of cycles of frequency modulation per second is small compared to the transmitter frequency swing, so the spectrum occupied by the signal is substantially that of the swing itself.

return. This results in a frequency difference between the direct and reflected signals which is equal to the product of the time delay $2H/C$ and the rate of change of frequency, and is given by the equation,

$$F_d = 4\Delta F F_m H/C \text{ cycles per second.}$$

The difference is plotted again at the bottom of the diagram and appears as a series of trapezoids of height F_d . The time delay, $2H/C$, has been greatly exaggerated in comparison with $1/F_m$, the time interval corresponding to one cycle of frequency modulation, in order to make the difference, F_d , large enough to show on the diagram. F_d is actually only a few cycles in hundreds of millions. It will be noted that F_d drops momentarily to zero twice for each complete sawtooth variation of the transmitter frequency. This is due to the necessity of varying the transmitter frequency first up and then down, instead of forever in one direction. Hence the theory must be considered from the standpoint that one altitude measurement is made for each upward and another for each downward sweep, ΔF , of transmitter frequency so that a total of $2F_m$ measurements are made per second. The number of cycles of frequency F_d , occurring during one frequency sweep, is

$$F_s = F_d \times \frac{1}{2F_m} = 2\Delta F H/C,$$

since $\frac{1}{2F_m}$ is the time of one sweep, ΔF . F_s is directly proportional to both the height and to the amount of transmitter frequency change, ΔF .

The fact that $2F_m$ separate measurements are made per second is important only when considering small altitudes. The height which gives a value of unity for F_s corresponding to a frequency meter signal of $2F_m$ cycles per second is the minimum height which can be indicated since lower altitudes give the same reading. Lower altitudes cause only a fraction of a cycle of frequency, F_d , to be generated per sweep, but since this fraction is repeated $2F_m$ times per second, it constitutes a signal of the same frequency $2F_m$ and is so counted by the frequency meter. In order to make this minimum altitude small, it is necessary that ΔF be large, since they are inversely proportional to each other. A frequency sweep of approximately 25 megacycles is required to provide measurements down to the present minimum of about twenty feet. If a high antenna efficiency is to be obtained over a band 25 megacycles wide, it is necessary that the percentage variation from the average frequency during the modulation cycle be small. This

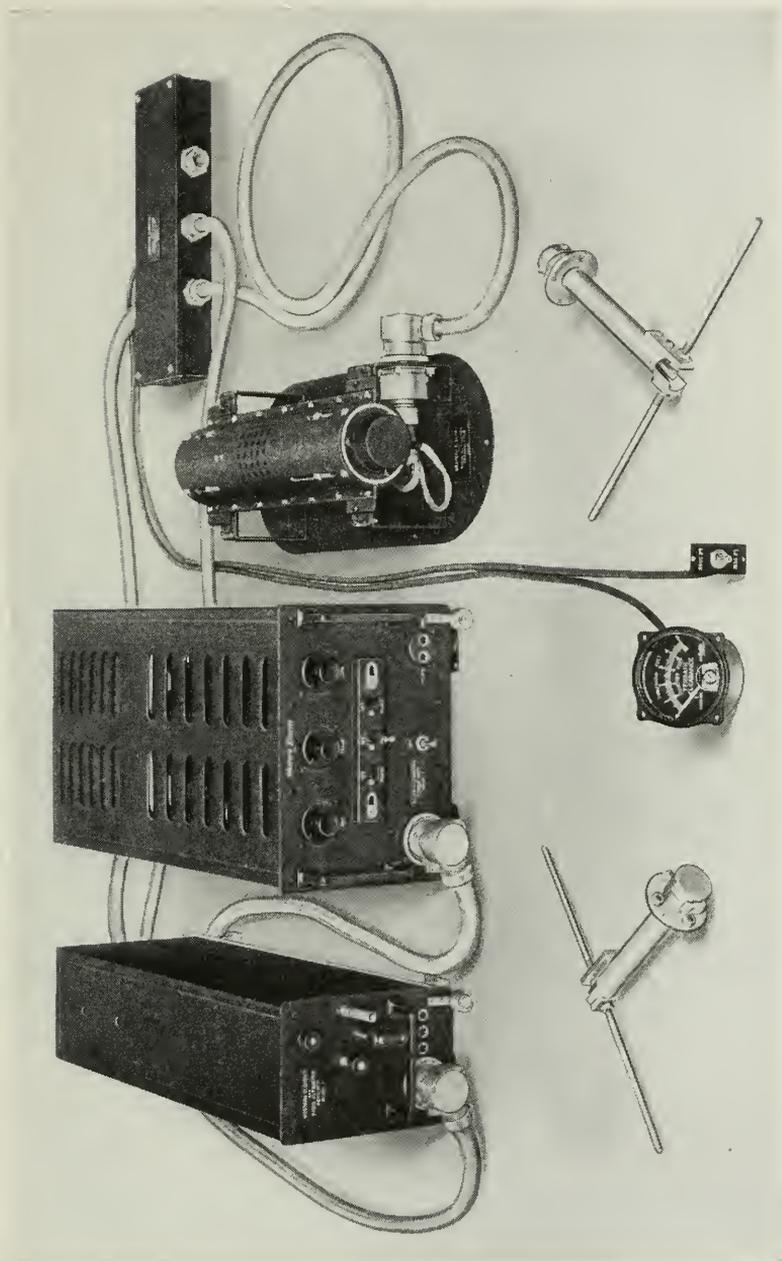


Fig. 3—Terrain clearance indicator units.

percentage variation is made small by the use of an average frequency of approximately 450 megacycles. The use of this ultra-high frequency has the additional advantage that the antennas can be both small and efficient and cause little drag upon the airplane.

APPARATUS

Figure 3 is a photograph of all the units of the altimeter, with the exception of the transmission lines used to connect the antennas to their respective units. The units are as follows: left to right, receiver, power unit, and transmitter, with a junction box in the upper right. In the foreground are the two dipole antennas and the indicating meter with its range-shift switch. The meter and one of these antennas are shown in larger scale in Fig. 4. The meter has two scales, the upper

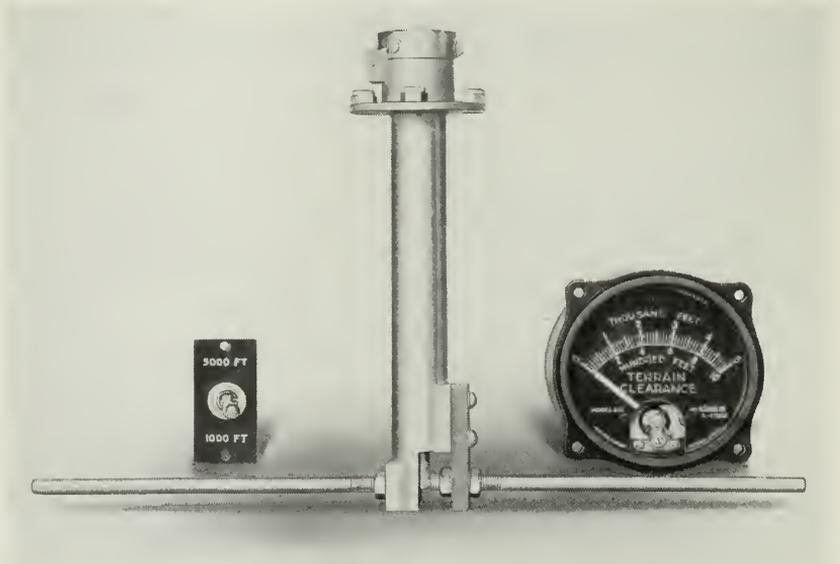


Fig. 4—Antenna, meter, and range switch.

extending from 0 to 5000 feet and the lower 0 to 1000 feet. The position of the range switch determines the scale to be used in reading the meter.

Figure 5 shows an assembly of the various units located approximately as they would be installed in an air transport. The transmitter, power unit, receiver and a junction box are installed in the baggage compartment just aft of the cockpit with cable connections

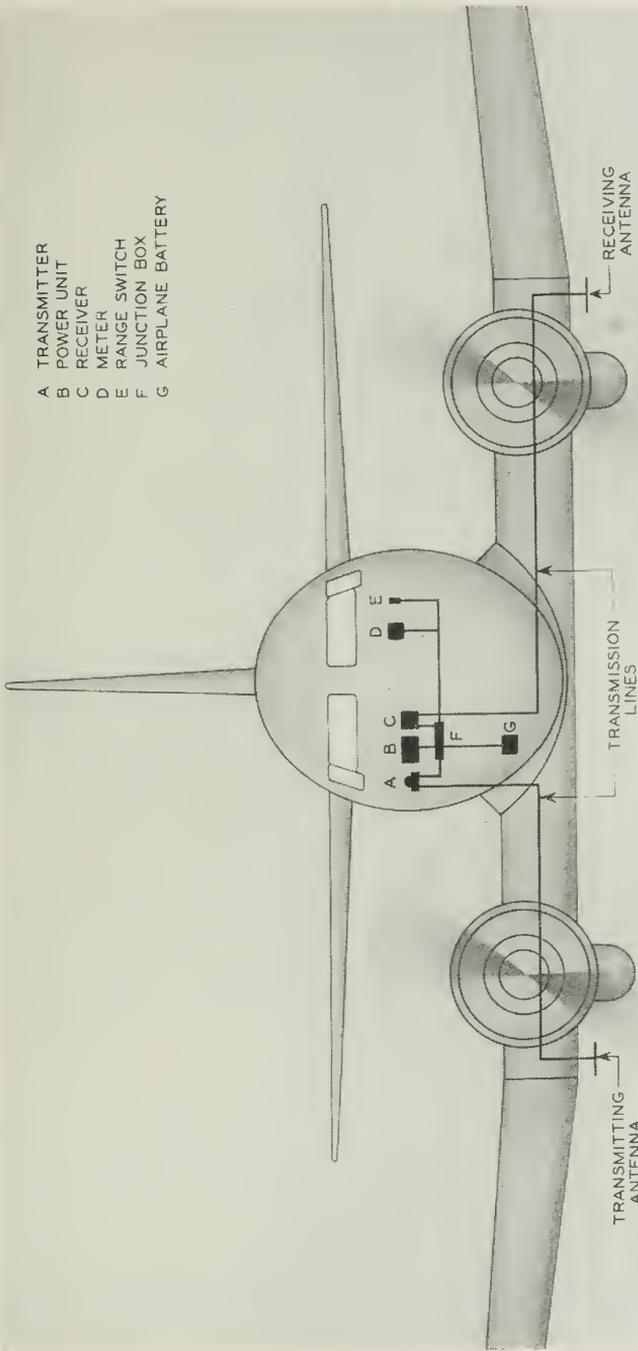


Fig. 5—Airplane installation.

to the airplane battery and to the meter and range switch on the instrument panel. The transmitting antenna is shown below the wing to the left of the engine nacelle and the receiving antenna to the right of the other engine nacelle. Coaxial transmission lines connecting the antennas to the transmitter and receiver, respectively, are indicated by the lines extending through the wings from the antennas. It was necessary to exaggerate the size of some of the units in order to make them large enough to see in the diagram.

The installation with apparatus as pictured in Fig. 3 weighs complete with all cables and connections about seventy pounds. Since the equipment shown in the pictures represents a working model built with the idea of attaining performance rather than minimum weight, undoubtedly some reduction in weight will be obtained in future models.

The antenna installation shown utilizing half-wave dipole type antennas approximately a quarter wave-length below the reflecting surface of the wing is not particularly directional. The signal is radiated over approximately the whole hemisphere below the wing centered on the transmitting antenna. The strength of the signal is greatest in the downward direction but does not fall off rapidly in other directions. The advantage of this antenna arrangement is that the distance to the nearest reflecting surface is measured regardless of whether it is directly beneath, or to the front or side. As a result very little change in reading occurs when the airplane banks steeply. Some advance indication also is given when the airplane in level flight approaches higher terrain.

PERFORMANCE

The terrain clearance indicator in its present experimental form indicates altitudes between approximately twenty and five thousand feet. When over smooth water or land, it is subject to errors as indicated by a consideration of the fundamental equation upon which the altimeter is based,

$$F_d = 4\Delta F F_m H/C.$$

Since F_d is directly proportional to both ΔF and F_m , any variation of a given percentage in either will result in a corresponding percentage error in the reading of the meter. It is believed from the data available that the errors due to variation of either ΔF or F_m do not exceed ± 1 per cent.

Additional errors can also occur in the frequency meter circuit. These errors are believed to be less than ± 7 per cent, so that a total error of ± 9 per cent might occur if all the errors were simultaneously

in the same direction. Fortunately, all these are of a percentage nature, so that the error in feet becomes smaller as the ground is approached. An absolute error in the indication is still possible because of the limitations of the milliammeter used on the instrument panel. The Weston aircraft meter used is guaranteed to be correct to within one per cent of its full scale reading at any point on its scale, which permits maximum errors of ten feet on the 1000-foot scale and fifty feet on the 5000-foot scale.

When flying over rough water, wooded terrain or cities, reflected signal is received from surfaces at different distances simultaneously, resulting in addition and subtraction interference effects, thus sometimes momentarily reducing the echo signal below the minimum required for accurate indication. In such a case, the meter hand may swing down momentarily as much as 10 per cent. For the present limited transmitter power and receiver sensitivity, at altitudes above 2500 feet, these momentary signal reduction effects become progressively more serious when flying over irregular surfaces so that for a substantial part of the total time the echo signal may be below the minimum required for correct meter reading. This is indicated by a reading fluctuating between 3000 and 5000 feet when flying at 5000 feet over a surface dotted by buildings, timber, etc. The meter swings up to the correct reading every time the airplane passes over a smooth field or body of water of any size. Up to 2500 feet the echo signal has proved to be sufficient for steady operation over all kinds of terrain.

Tests have been made over New York, Raritan, Newark and San Francisco Bays, Great Salt Lake, Lakes Erie and Michigan, the timbered mountains of Washington and Oregon, the deserts and mountains of the southwest and the cultivated areas of the midwest during the period of the recent demonstration flights made with the equipment installed in the United Air Lines Flight Research Airplane.

An indication of the character of the surface over which the airplane is flying is given by the variations in the meter reading. A city usually causes rapid fluctuations of the order of fifty feet, depending, of course, upon the height and the spacing of the buildings. Cultivated farmland causes fluctuations of lower frequency and amplitude. An isolated high object such as a skyscraper or a chimney is indicated only by a slight meter kick as the airplane passes over it, which may not be noticed by the observer. If the airplane passes over only a few feet above the object and the top is large enough to contribute momentarily most of the echo signal received by the airplane, the indication is unmistakable and the correct distance to the object is indicated by the meter. For instance, the gas storage tank

near the Chicago airport is an excellent object upon which to demonstrate the altimeter performance. The instrument is useful as a position indicator when approaching an airport on a course which crosses an obstruction of appreciable height and size since the moment of passage over the obstruction is clearly indicated. In fact, use as a position indicator may be one of the altimeter's most valuable applications.

A study of the circumstances in connection with a number of crashes in the west during recent years has revealed that in most of the cases the airplanes crashed after having been within a few feet of the ground without the pilot knowing it for several minutes before they struck. In such a situation the terrain clearance indicator should be capable of warning the pilot in ample time to avert a crash.

The writers wish to express their appreciation of the contributions of a number of other members of the technical staff of the Bell Telephone Laboratories to the success of this project.

REFERENCES

1. H. Loewy, *U. S. Patent* 1,492,300.
H. Loewy, *U. S. Patent* 1,585,591.
C. F. Jenkins, *U. S. Patent* 1,756,462.
E. F. W. Alexanderson, *U. S. Patent* 1,969,537.
F. H. Drake, *U. S. Patent* 1,987,587.
2. *U. S. Patent* 2,011,392 issued August, 1935 to J. O. Bentley.
3. Page 29 of "Solving the Problem of Fog Flying," a publication of the *Daniel Guggenheim Fund for the Promotion of Aeronautics*, 1929.
4. "Altitude Measurements by Reflected Electro-Magnetic Waves" by Murray Hively and R. C. Newhouse, Ohio State University Library, 1929.
5. "An Electro-Magnetic Altimeter" by R. C. Newhouse and J. D. Corley, Ohio State University Library, 1930.
6. *U. S. Patent* 1,517,549 issued December, 1924 to Lloyd Espenschied.
7. *U. S. Patents* 2,045,071 and 2,045,072 issued June, 1936 to Lloyd Espenschied.
8. "A Negative Grid Triode Oscillator and Amplifier for Ultra-High Frequencies," A. L. Samuel, *Proceedings of Institute of Radio Engineers*, 25, Oct. 1937 (1243).
9. "A Research of Direct Reading Altimeter for Aeronautical Use by Radio Reflection Method," Sadahiro Matsuo, *Journal I. E. E. Japan*, No. 571, February, 1936.
10. "A Direct-Reading Radio-Wave-Reflection Type Absolute Altimeter for Aeronautics," Sadahiro Matsuo, *Proceedings of Institute of Radio Engineers*, Vol. 26, 1938 (848).
11. *Trans. A. I. E. E.* 49, 1930 (1331).
Proc. I. R. E. 19, 1931 (659).
Review of Scientific Instruments 6, 2, January, 1935.
Journal of Scientific Instruments 14, 1937 (136).

Transcontinental Telephone Lines *

By J. J. PILLIOD

Late in 1937 a large construction project was completed which added 16 telephone circuits to the transcontinental layout, and the work was so planned that 48 additional circuits can be obtained by the addition of equipment but without stringing additional wire. A brief description of some features of this project and the general development of the transcontinental telephone routes since the first one was opened for service in 1915 is given in this article. Although most of the discussion relates to transcontinental lines, the methods described are generally applicable to other similar situations.

LESS than twenty-five years ago, it was impossible to talk by telephone from coast to coast across the United States. Furthermore, it was impossible to talk between points separated by any such distance anywhere in the world. By 1915, technological advancement had reached a point such that telephone service could be established across the country, and three telephone circuits had been built which connected San Francisco and the Pacific Coast with points in the East. Four telegraph circuits were also provided by the new wires. An improved loading system and especially the successful development of the vacuum-tube telephone repeater were outstanding factors which made telephone connections of such length possible for the first time in history.

Open-wire lines played the major role in the early transcontinental telephone circuits. The transmission losses caused by cable were so great that it was avoided wherever possible. The steady improvement of telephone repeaters, types of loading for use on cable circuits, and carrier telephone systems for use on open-wire lines made it possible to provide rapidly and economically more telephone circuits across the continent as use of the service grew. In the cross section

* This paper has been prepared from an address given before the Communications Group of the A. I. E. E., New York Section, March 22, 1938, and published in *Electrical Engineering* for October, 1938. Since the paper was written, three type J 12-channel carrier systems have been placed in service on the new line. Two of these systems operate between Oklahoma City and Whitewater, 1200 miles, and the third between Oklahoma City and Albuquerque, N. M. Twelve additional intermediate repeater stations have been constructed. Three of these are located at such remote distances from primary power that experiments are being made in generating by means of wind-mill power plants part of the power required. One such station is shown in Fig. 8. *Editor.*

just west of Denver there are today one hundred and forty through telephone circuits and about the same number of telegraph circuits carried by four open-wire routes, the last of which was completed during 1937. While open wire was used almost exclusively as a matter of necessity in the first transcontinental telephone lines, cable is now used for about half of the circuit mileage. This is a striking illustration of the large-scale changes which have taken place in the interest of more reliable toll telephone service.

CONTINUED IMPORTANCE OF OPEN-WIRE LINES

The open-wire line seems destined to continue to play an important part in long-distance telephone communication, particularly where distances are great and circuit requirements on any one route are relatively small. Improvements in the usage to which the wires may be put have made this increasingly so. The three circuits on the first transcontinental line were operated at voice frequencies and were obtained from two pairs of line conductors, the third circuit being derived by means of phantom circuit arrangement of these two pairs. The development of carrier telephone systems made it possible to obtain three additional circuits on some pairs of wires, using frequencies above those required for existing voice-frequency circuits. Carrier telephone systems were first installed on a transcontinental route in 1926 and were quickly followed by others, so that today ninety-six of the one hundred and forty circuits mentioned earlier are obtained by means of these three-channel carrier telephone systems. Development work, however, has been continued, and it is now expected that it will be possible, by means of carrier telephone systems using still higher frequencies, to obtain as many as twelve more telephone circuits on some pairs of wires. It has been with a view toward using such systems and obtaining a total of sixteen telephone circuits on a pair of wires that the latest of the four transcontinental routes has been designed.

CONSTRUCTION OF NEW TRANSCONTINENTAL LINE

Early in 1937, it became clear from a study of loads carried on existing transcontinental routes that additional circuits would be required in the near future. Circuits in cable were available as far west as Omaha, Kansas City, Oklahoma City, and Dallas. After consideration of all the factors, it was decided to construct the new facilities west from Oklahoma City to Los Angeles on the route shown in Fig. 1. It was also decided to carry out the work in such a way that the route could be utilized for the future addition of a relatively

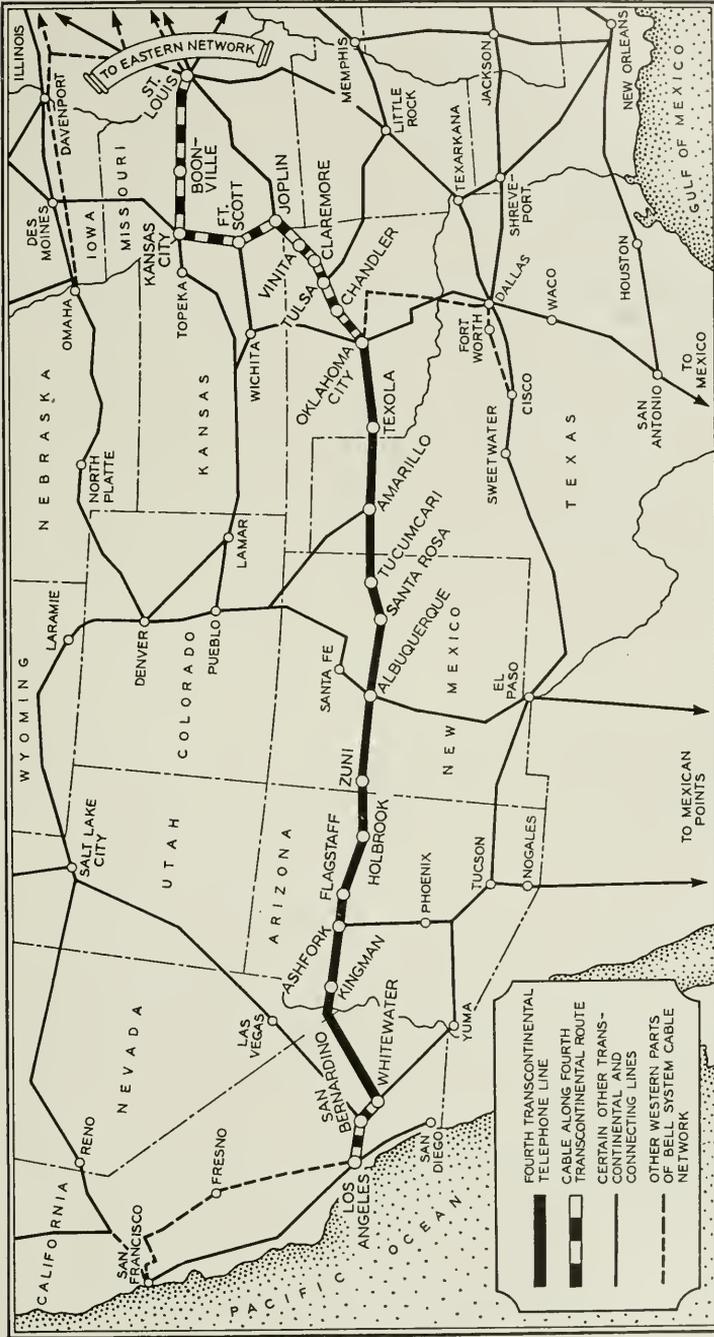


Fig. 1—Route of new transcontinental line across western states.

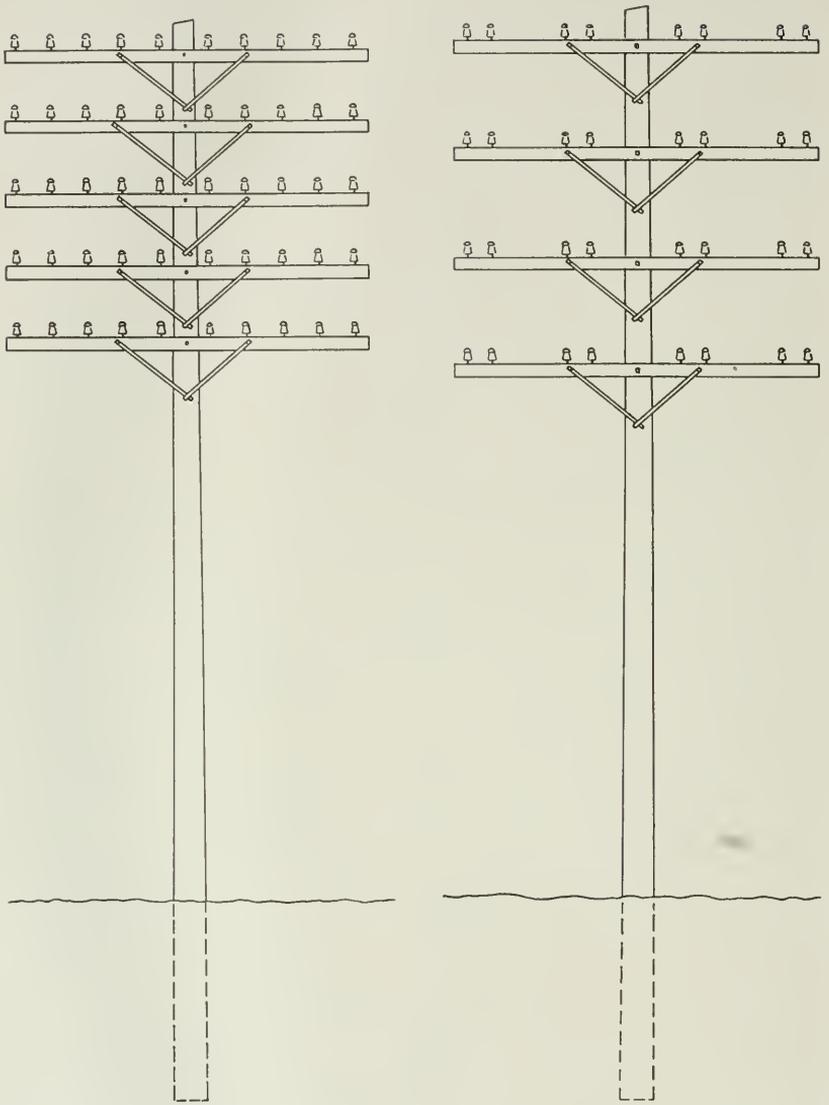


Fig. 2—Cross-arm arrangements. Left—50-wire phantom line, capacity 77 telephone circuits. Right—32-wire non-phantom line, capacity 256 telephone circuits.

large number of circuits through the application of the twelve-channel carrier-current telephone system then under development. Among the conditions favoring this particular route is freedom from winter storm hazards throughout most of the distance, which, looking ahead, is particularly important to the future application of twelve-channel carrier telephone systems. The work done in 1937 consisted of building a length of nearly three hundred miles of new pole line and stringing four pairs of wires throughout most of the section from Oklahoma City to Whitewater, California, a distance of 1,200 miles. Initially the voice channel and three-channel carrier telephone systems have been developed on these four pairs, providing a total of sixteen telephone circuits.

WIRE SPACING AND TRANSPOSITIONS

Open-wire telephone lines designed to carry frequencies up to 140 kilocycles per second, as used in the operation of the twelve-channel carrier telephone systems, have structural requirements substantially more stringent than those designed to carry only three-channel systems, which use frequencies up to 28 kilocycles. The usual type of open-wire toll telephone line has ten wires on each crossarm, spaced at about one-foot intervals, five on each side of the pole and with the crossarm spaced twenty-four inches apart. In the case of the line designed to conduct high carrier telephone frequencies, this configuration has been changed and is illustrated by Fig. 2. Eight wires are strung on each arm, grouped as four pairs, two on each side of the pole. The wires of the pair are spaced eight inches apart, and the nearest wires of the two pairs on each side of the pole are spaced twenty-six inches, while the spacing at the pole is thirty inches. Cross-arms are spaced thirty-six inches apart.

These new wire spacings reduce the coupling between pairs on the same line or between pairs on this line and pairs on other lines which may parallel it. New transposition systems are used further to reduce this coupling. Transpositions are closer together and a transposition bracket of the type shown in Fig. 3 is used to turn the wires completely over at as nearly a given point as possible. Transpositions in one or more pairs are installed on every pole with an occasional exception, and certain pairs are transposed at every other pole. The wires of a pair must be adjusted to the same sag within close limits. These sag variations are held to a fraction of an inch, and a check of the completed work indicated that fifty per cent of the spans had been adjusted to within one-quarter inch. Telescopes are used to help obtain these close sag adjustments, and a final check is made by oscillating the wires in a span and observing the periods at which they oscillate.

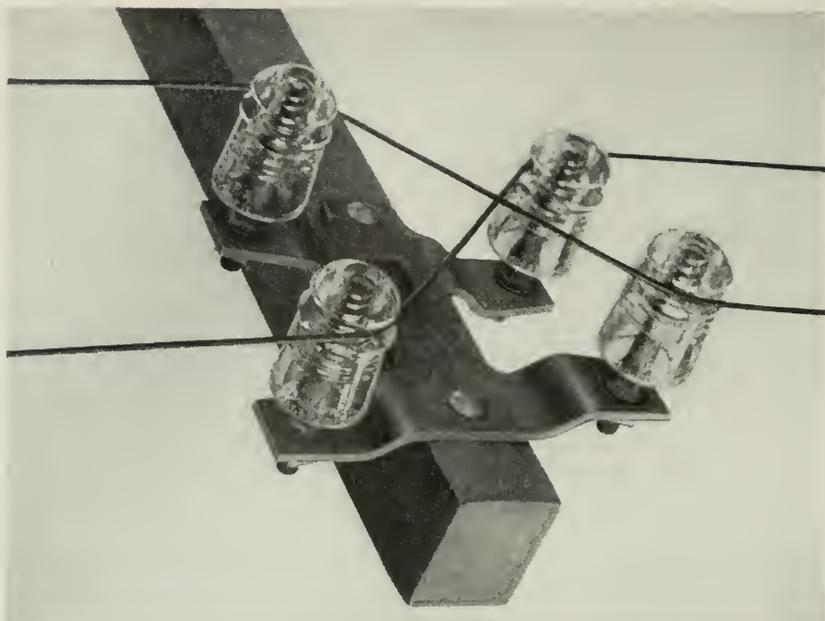


Fig. 3—Transposition bracket shown as installed. Tie wires are not used with this type bracket.



Fig. 4—Flat tie wire shown as installed.

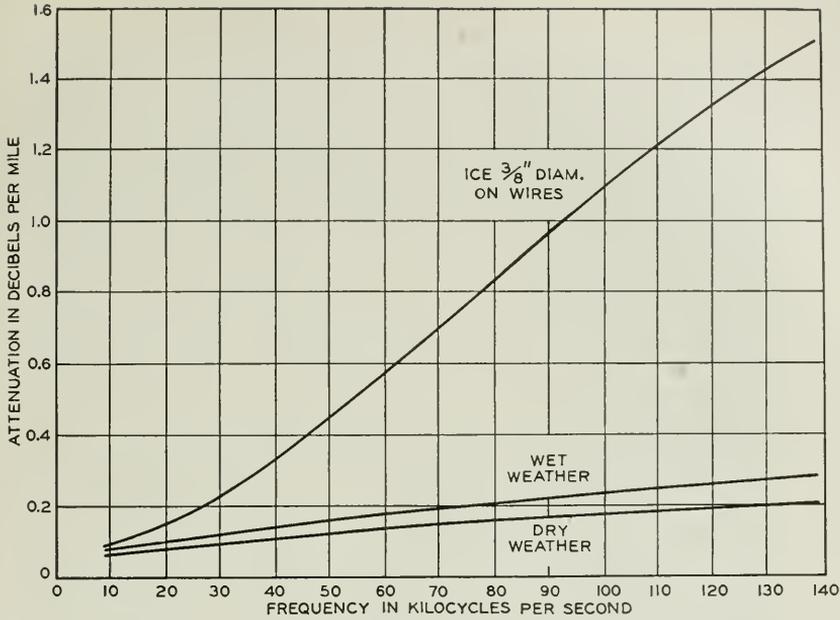


Fig. 5—Attenuation of open-wire pairs, 165-mil copper, 8-inch spacing, CS insulators.

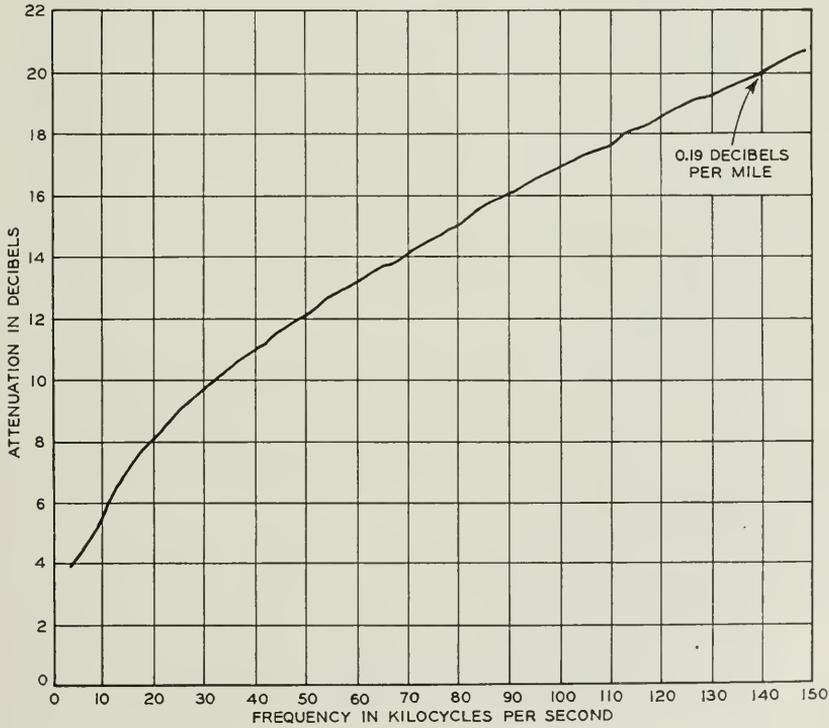


Fig. 6—Attenuation of 165-mil copper, 8-inch spaced pair of wires measured on a 105-mile section of the Amarillo-Albuquerque line in clear weather.

POLE SPACING AND INSULATORS

Poles must be spaced uniformly in order that the transpositions may be most effective, and an occasional deviation of only thirty-five feet is the maximum permitted. Where it is impossible to locate poles within this limit, such as is the case at long-span crossings, special fixtures are suspended from steel cables at the proper points to permit making the transpositions.

New types of insulators on steel pins, each pair of which is electrically bonded, are used to improve the stability of the transmission characteristics.

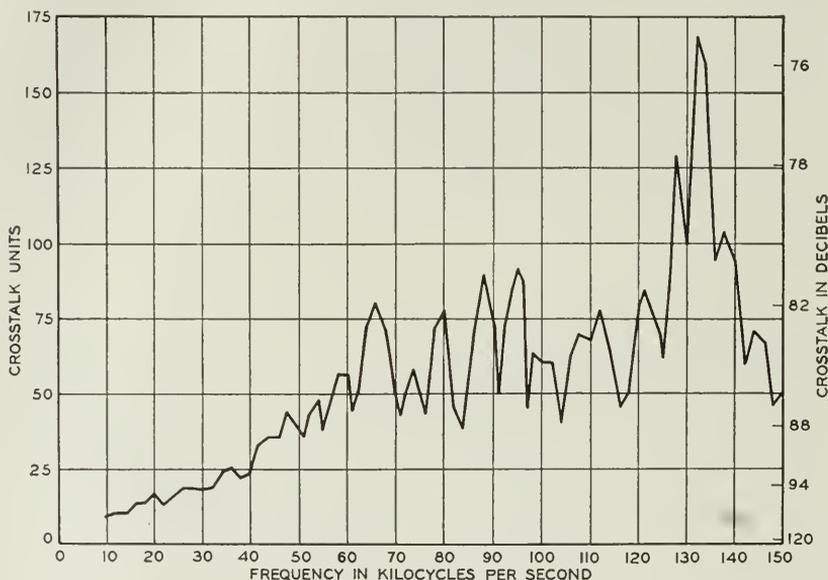


Fig. 7—Far-end crosstalk between wires 7-8 and 9-10 of Amarillo-Albuquerque line, measured from pole 1 to pole 4236, a distance of 105 miles.

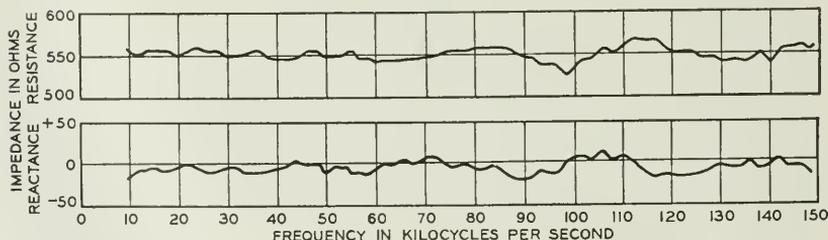


Fig. 8—Impedance of 165-mil copper, 8-inch spaced, CS insulated pair of wires on Amarillo-Albuquerque line, measured from pole 1 to pole 4236.



Fig. 9—Type J carrier repeater station at San Fidel, New Mexico (about sixty miles west of Albuquerque), showing wind mill power generating plant. This station is unattended.



Fig. 10—Long span crossing at Tucumcari Creek, New Mexico. This span is 1080 feet between supporting towers. Normally, the creek carries very little water and the long span is to avoid damage at times of flood.

Wire 165 mils in diameter was used on this construction west of Oklahoma City because of its strength and resultant relative freedom from interruptions. Transmission losses were also a factor in this case. Rolled sleeve joints were used in splicing the wire because of their strength and good electrical characteristics. Flat tie wires, shown in Fig. 4, were used to reduce chafing of the line wire when the wires vibrate. Tie wires are not used at transposition points, as may be seen in Fig. 3.

REPEATER STATIONS

The computed losses on open-wire pairs of this type at high-carrier frequencies and under different weather conditions are shown by Fig. 5. Field tests confirm these data. To offset the line losses it will be necessary to locate repeater stations at intervals of from fifty to one hundred miles, depending upon the weather conditions which may be expected. Between Oklahoma City and Whitewater, California, in order to operate the twelve-channel carrier telephone systems, it will be necessary to equip sixteen intermediate repeater stations. Most of these will be unattended and maintained from other offices.

It is not practicable, of course, to bring the open-wire pairs directly into all repeater stations and in some cases entrance cables several miles long must be used. Although ordinary non-loaded cable pairs may be used for this purpose, their usage involves transmission difficulties, and except where other factors dictate the use of this type of facility, it is planned to use low-loss cable conductors of a new design. These cable pairs have more favorable impedance characteristics as well as lower losses.

With the building and the further equipping of this latest open-wire line across the western states, open-wire facilities have played one more important part in the development of long-distance telephony. Although cable is being found more and more useful, there still remain many important links in the nation-wide telephone communication network where, for the present at least, the open-wire line can serve best and the development of it toward maximum usefulness is still being carried on.

Abstracts of Technical Articles from Bell System Sources

*Paper as a Medium for Analytical Reactions.*¹ B. L. CLARKE and H. W. HERMANCÉ. Absorbent paper has long been used in chemical laboratories for filtering suspensions. Paper has also found a special use as a container or holder for certain testing reagents; litmus paper is a common example. In this article and a preceding one (*Indus. & Engg. Chem., Anal. Ed.*, June 15, 1937), are reported exploratory investigations on the extension of the use of absorbent papers in chemical analysis.

Rapid identification "spot tests" have been described by Feigl in which, by successively placing drops of unknown and reagent solutions on filter paper, characteristic color changes are produced. Methods and apparatus are described in the present articles whereby some of the variables in such tests are controlled. Chief among these innovations is the use of semi-soluble instead of soluble reagents, and the precipitation of these compounds directly on the paper fibres to form a more or less permanent test paper. By these changes in technique the sensitivity—the smallest amount of a given metal detectable—is decreased from ten to one-hundredfold. For example, 0.002 microgram of copper can be detected by the new method, as compared with 0.2 by the old.

In another application a very dilute solution of some metal ion is slowly siphoned through a small circular piece of reagent paper suitably mounted. The metal is entrapped on the paper in an insoluble form strongly adsorbed by the paper. Theoretical analysis indicates that copper, for example, may be removed from a solution in this way so completely that only 8×10^{-12} microgram will be left in a liter.

*Neutral Particles in Physics.*² KARL K. DARROW. During the early days of science, the elementary particles which scientists and philosophers alike saw fit to postulate were always imagined as chargeless. With the remarkable growth of the understanding of electricity during the nineteenth century, and with the invention of instruments for detecting small charged particles during the twentieth, it became the custom to suppose that the fundamental particles of

¹ *Indus. & Engg. Chem., Anal. Ed.*, October 15, 1938.

² *Amer. Phil. Soc. Proc.*, September 30, 1938.

matter all bear charges and that the forces exhibited in Nature are all electrical (exception being made for gravitation). A noted and serious objection to this view was temporarily met by the adoption of quantum mechanics. Since 1930 a reversal of trend has set in, heralded by the discovery of the neutron as a subatomic chargeless particle capable of independent existence; and at present there is a strong tendency to develop the view that neutral as well as charged particles of subatomic size, and non-electrical as well as electrical forces, exist together in Nature.

*Electrical Networks for Sound Recording.*³ F. L. HOPPER. Electrical networks are employed in sound recording for modifying and limiting the frequency-response characteristic. The necessity for their use, application, and design is described. Particular emphasis is placed upon the constant-resistance type of structure.

*Sound Pictures in Auditory Perspective.*⁴ FRANKLIN L. HUNT. Soon after sound reproduction in auditory perspective was demonstrated over telephone circuits between Philadelphia and Washington in 1933, experimental sound pictures in auditory perspective were made at the Bell Telephone Laboratories' sound picture laboratory. Listening tests showed that they distinctly enhanced the illusion that the sound originated at its apparent source on the screen and they strikingly improved the feeling of spaciousness and reality. The auditory perspective effect is not primarily dependent upon perfect synchronism of the two sound-tracks required, nor on frequencies above the present commercial range. Existing equipment can be converted to project sound pictures in auditory perspective without great difficulty.

*Composition and Colloidal Properties of Balata Latex.*⁵ A. R. KEMP. This paper reports the composition and colloidal properties of two types of balata latex from Dutch Guiana. The white variety is shown to be superior to the red, owing to its higher content of hydrocarbon.

It is shown that balata latex is very stable owing to the presence of a highly protective water-soluble substance in its serum. It cannot be coagulated by acids or salts, but is readily coagulated by alcohol or acetone.

The balata latex particles are spherical and vary in diameter from about 0.1 to 2.5 microns with an average diameter of about 0.5 micron.

The balata latex particles are shown to enclose the resins, which

³ *Jour. S. M. P. E.*, November 1938.

⁴ *Jour. S. M. P. E.*, October 1938.

⁵ *India Rubber World*, December 1, 1938.

appear to be present in a dispersed state in the hydrocarbon. The particles are shown to contain about 18% of water, determined as water of retention in pressed coagulum.

The "resins" have been separated from both types of balata latex as water-white viscous liquids which deposit crystals of β -amyryn acetate on standing. The red balata latex resin is shown to be more viscous than the white and to differ from it as regards its iodine value, refractive index, and solubility in cold 95% ethyl alcohol.

The serum constituents have been separated into four main fractions: protein, carbohydrate, gummy substance, and ash. Minor constituents such as tannin and amino-acid have also been noted.

A complete analysis of balata ash has been made and compared with the analysis of ash from *Hevea* latex by Bruce. Balata ash was found to contain higher contents of CaO, Na₂O, and MgO and lower contents of K₂O and P₂O₅ than *Hevea* latex ash.

New data are presented on the density, refractive index, dielectric constant, and heat of combustion of balata hydrocarbon which are believed to be more reliable than similar data previously available in the literature. Data on the effect of temperature on the refractive index of balata and gutta percha hydrocarbon are presented, showing the crystallization of the gutta hydrocarbon on cooling, which starts at about 37° C. resulting in an abrupt increase in refractive index occurring between 37° and 35° C.

*A Short-Wave Single-Sideband Radio Telephone System.*⁶ A. A. OSWALD. There is described briefly a short-wave single-sideband system which has been developed for transoceanic radio telephone service. The system involves the transmission of a reduced carrier or pilot frequency and is designed to include the testing of twin-channel operation wherein a second channel is obtained by utilizing the other sideband.

The paper indicates the reasons which led to the selection of this particular system and discusses at some length those matters which require agreement between the transmitting and receiving stations when single-sideband transmission is employed.

*The Oxide-coated Filament. The Relation between Thermionic Emission and the Content of Free Alkaline-earth Metal.*⁷ C. H. PRESCOTT, JR. and JAMES MORRISON. The oxide-coated filament had its beginning in the sealing-wax era of vacuum technique. The obscure accident of its origin is not recorded, but all of our older physicists knew

⁶ *Proc. I. R. E.*, December 1938.

⁷ *Jour. Amer. Chem. Soc.*, December 1938.

that an enhanced emission of electrons could be obtained by smearing sealing-wax on a platinum ribbon and burning it off in air. The first authentic study is recorded by Wehnelt, who investigated the voltage-drop in a gas discharge tube with cathodes coated with various metallic oxides. Its further evolution and development to the status of a cathode in Western Electric vacuum tubes has been described by H. D. Arnold. A comprehensive treatment of its history, the various modifications in current use, and divergent theories of its preparation and behavior has been given by Saul Dushman in a treatise on "Thermionic Emission." A later review is given by J. H. deBoer.

The present work is devoted to a quantitative determination of the relation between thermionic emission and the content of free alkaline earth metal. To this end we have employed a filament which is a platinum rhodium core coated with barium, strontium, and nickel carbonates. On heating in a reducing atmosphere this coating becomes a grossly homogeneous colloidal mixture of barium oxide, strontium oxide, and free nickel. After a thorough preliminary clean-up of the experimental tube, the requisite amounts of free alkaline-earth metal are generated by reaction with methane. The electrical measurements are summarized by the use of the Richardson equation for thermionic emission. Free alkaline earth metal has been determined by oxidation with carbon dioxide and analysis of the gaseous reaction products.

Using a filament coated with a colloidal mixture of barium oxide, strontium oxide, finely divided nickel, and free alkaline earth metal, we have investigated the quantitative relation between thermionic emission and the content of active metal. A high level of activity was found from 15 $\mu\text{g./sq.cm.}$ to 60 $\mu\text{g./sq.cm.}$ of equivalent Ba, with a slight apparent maximum at 30 $\mu\text{g./sq.cm.}$ where the thermionic current at 1050° K. is 600 m. a./sq.cm. The electron work function is 1.37 v.

The radiant emissive power at 0.66 μ is approximately 64%, independent of the content of active metal.

The free alkaline earth metal was determined by oxidation with carbon dioxide and analysis of the gaseous reaction products.

*A Single-Sideband Receiver for Short-Wave Telephone Service.*⁸
A. A. ROETKEN. A new radio telephone receiver has been developed for the reception of reduced-carrier single-sideband signals in the frequency range from 4 to 22 megacycles. This receiver employs triple detection in which the first beating oscillator is continuously

⁸ *Proc. I. R. E.*, December 1938.

variable and the second is fixed in frequency. The first oscillator is a very stable tuned-circuit type, the proper adjustment of which is maintained through the use of an improved type of synchronizing automatic-tuning-control system. The second oscillator is crystal controlled. Separation of the carrier and sideband is accomplished in the receiver by means of band-pass crystal filters which provide extremely high selectivity. Unusually high stability and selectivity characterize the performance of the receiver.

*Dielectric Constant and Dielectric Loss of Plastics as Related to their Composition.*⁹ W. A. YAGER. Data are presented for the frequency variation of the dielectric constant and dielectric loss factor of various plastics over a broad frequency band extending from 1 kc. to 35 mc. The extremely low loss of polystyrene compared to that of polar plastics confirms the theory that a hydrocarbon is inherently more satisfactory from a dielectric point of view. Of the several possible mechanisms of dielectric loss which might account for the high-frequency dielectric absorption observed in polar plastics, the rotation of polar units in the chain and of polar side groups appears most probable. The fact that the loss factor maxima of phenol fibers, phenol fabrics, and phenol or urea formaldehyde molding compounds containing cellulosic fillers occur at essentially the same frequency is viewed as evidence that this dielectric absorption is an intrinsic property of cellulose. Substitution of mineral fillers for cellulose reduces the high-frequency loss to that residing in the polar resin binder. Furthermore, the dielectric loss of mineral-filled molding compounds is less moisture-sensitive. The large increase in dielectric loss at low frequency always found in materials of relatively high free-ion conductivity manifests itself in Duprene, and the humidified phenolic plastics containing cellulose fillers or laminations.

⁹ *Electrochemical Society Preprint* No. 74-24, October 12-15, 1938.

Contributors to this Issue

H. A. AFFEL, S.B. in Electrical Engineering, Massachusetts Institute of Technology, 1914; Research Assistant in Electrical Engineering, 1914-16. American Telephone and Telegraph Company, Engineering Department and the Department of Development and Research, 1916-34; Bell Telephone Laboratories, 1934-. As Assistant Director of Transmission Development, Mr. Affel is concerned with toll transmission problems, including the development of carrier telephone systems.

J. W. BEYER, B.S., State College of Washington, 1915; Westinghouse Electric and Manufacturing Company, 1915-16. Instructor in Electrical Engineering, State College of Washington, 1916-19. Western Electric Company, Engineering Department, 1919-24. Bell Telephone Laboratories, 1924-. Mr. Beyer is concerned with the development of carrier telephone terminals and pilot channels for open-wire circuits.

E. C. BLESSING, B.S. in Electrical Engineering, Purdue University, 1922. Western Electric Company, Engineering Department, 1922-25; Bell Telephone Laboratories, 1925-. Mr. Blessing has been engaged in the development of carrier systems.

S. I. CORY, B.E.E., Ohio State University, 1916. American Telephone and Telegraph Company, Engineering Department and Department of Development and Research, 1916-1934; Bell Telephone Laboratories, 1934-. During this entire time Mr. Cory has been engaged in transmission development work, chiefly on telegraph systems and transmission-measuring methods.

FRANK A. COWAN, B.S. in Electrical Engineering, Georgia School of Technology, 1919. American Telephone and Telegraph Company at Atlanta, Ga., 1920; Long Lines Engineering Department, Special Service Group, New York, 1922; appointed Division Transmission Engineer, Division No. 1, New York, 1926; appointed to present position, Engineer of Transmission, Long Lines Department, 1928.

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, Pratt Institute, 1909, has taken an important part in much of the Bell System's invention and development in the field of radio and carrier currents; also in technical contacts and international conferences abroad. Previously, in the Department of Development and Research of the American Telephone and Telegraph Company; now, as Research Consultant of the Bell Telephone Laboratories, participates in a broad front of electric-communications development.

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.

WILLIAM H. HARRISON, Vice President and Chief Engineer, American Telephone and Telegraph Company. Pratt Institute, graduate in Industrial Electrical Engineering, 1915. New York Telephone Company, repairman, apparatus inspector, 1909-14. Western Electric Company, engineer, 1914-18. American Telephone and Telegraph Company, 1918-33, Department of Operation and Engineering, engineer; Equipment and Building Engineer; Plant Engineer. Bell Telephone Company of Pennsylvania, Diamond State Telephone Company, Vice President in charge of Operations, 1933-37. American Telephone and Telegraph Company, Department of Operation and Engineering, Assistant Vice President, 1937-38; Vice President and Chief Engineer, 1938-.

L. W. HUSSEY, A.B., Dartmouth, 1923; M.A., Harvard, 1924; B.S. in E.E., Union College, 1930; Mathematics Department, Union College, 1924-29. Bell Telephone Laboratories, 1930-. Mr. Hussey has been engaged principally in work on the stability of regenerative systems and on modulation in non-linear resistances.

B. W. KENDALL, S.B., Massachusetts Institute of Technology, 1906; Instructor in Physics at Massachusetts Institute of Technology, Barnard College, and Columbia University, 1906-13. Engineering Department of the Western Electric Company, 1913; Bell Telephone Laboratories, 1925-. As Toll Development Director Mr. Kendall has charge of the development of carrier, voice frequency, and telegraph

circuits. His early work was on repeaters in connection with the transcontinental line; he has also been connected with carrier-current development since its inception.

R. C. NEWHOUSE, B.E.E., Ohio State University, 1929; Guggenheim Fellow, Ohio State University, 1929-30; M.Sc., Ohio State University, 1930. Bell Telephone Laboratories, 1930-. During most of this period Mr. Newhouse has been engaged in the design and development of aircraft radio transmitters. In recent months his efforts have been confined to the development of the terrain clearance indicator, for which he has been given the 1938 Lawrence Sperry Award by the Institute of the Aeronautical Sciences.

J. T. O'LEARY, B.S. in Electrical Engineering, Villanova College, 1918. American Telephone and Telegraph Company, Department of Development and Research, 1919-34. Bell Telephone Laboratories, 1934-. Mr. O'Leary has been concerned with the transmission aspects of carrier systems.

E. PETERSON, Cornell University, 1911-14; Brooklyn Polytechnic, E.E. 1917; Columbia University, A.M. 1923; Ph.D. 1926. Electrical Testing Laboratories, 1915-17; Signal Corps, U. S. Army, 1917-19. Western Electric Company, Engineering Department, 1919-25; Bell Telephone Laboratories, 1925-. Lecturer in Electrical Engineering, Columbia, 1934-. As circuit research engineer, Dr. Peterson's work has been largely in theoretical studies of non-linear circuits and circuit elements.

J. J. PILLIOD, E.E., Ohio Northern University, 1908. American Telephone and Telegraph Company, Long Lines Department, 1908-11; General Engineering Department, 1912-13; Long Lines Department, Division Plant Engineer, 1914-17; Engineer of Transmission, 1918-19; Engineer, 1920-. Mr. Pilliod is the head of the Long Lines Engineering Department.

J. N. REYNOLDS, B.S. in Electrical Engineering, Purdue University, 1904; E.E. 1907. Western Electric Company, Engineering Department, 1904-25. Bell Telephone Laboratories, 1925-. Mr. Reynolds has been continuously associated with the development of machine switching apparatus. As Special Studies Engineer, he is now engaged in the development of improved forms of crossbar switch and allied apparatus.

FREDERICK J. SCUDDER, New York Telephone Company, 1905-10; Western Electric Company, Engineering Department, 1910-25; Bell Telephone Laboratories, 1925-. Mr. Scudder has been engaged in the development of machine switching systems since 1910, and in his present capacity as Systems Development Engineer is in charge of fundamental studies and circuit development of panel and crossbar systems.

R. B. SHANCK, B.E.E., Ohio State University, 1915. Railroad telegraph service, 1909-10. American Telephone and Telegraph Company, Plant Department, 1910-11 and summers 1912-14; Engineering Department and Department of Development and Research, 1919-34; Bell Telephone Laboratories, 1934-. As Telegraph Transmission Engineer, Mr. Shanck is engaged in development work on the transmission features of telegraph systems and the measurement of telegraph transmission.

THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS
OF ELECTRICAL COMMUNICATION

Some Ceramic Manufacturing Developments of the Western
Electric Company—*A. G. Johnson and L. I. Shaw* . . . 255

The Production of Ultra-High-Frequency Oscillations by
Means of Diodes—*F. B. Llewellyn and A. E. Bowen* . . . 280

A Representation of the Sunspot Cycle—*C. N. Anderson* . . . 292

The Number of Impedances of an n Terminal Network
—*John Riordan* 300

Copper Oxide Modulators in Carrier Telephone Systems
—*R. S. Caruthers* 315

Some Applications of the Type "J" Carrier System
—*L. C. Starbird and J. D. Mathis* 338

Line Problems in the Development of the Twelve-Channel
Open-Wire Carrier System
—*L. M. Ilgenfritz, R. N. Hunter and A. L. Whitman* 363

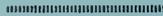
Abstracts of Technical Papers 388

Contributors to this Issue 391

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

THE BELL SYSTEM TECHNICAL JOURNAL

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



EDITORIAL BOARD

F. B. Jewett	H. P. Charlesworth	W. H. Harrison
A. F. Dixon	O. E. Buckley	O. B. Blackwell
S. Bracken	M. J. Kelly	G. Ireland
	W. Wilson	
R. W. King, <i>Editor</i>	J. O. Perrine, <i>Associate Editor</i>	



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, 1939
American Telephone and Telegraph Company

The Bell System Technical Journal

Vol. XVIII

April, 1939

No. 2

Some Ceramic Manufacturing Developments of the Western Electric Company

By A. G. JOHNSON and L. I. SHAW

A general picture is given of the development work involved in the introduction of manufacturing processes for vitreous enameled resistances, vitreous enameled iron and copper base number plates, pressed glass lenses, extruded and pressed porcelain parts, and close tolerance ceramic insulators for use in telephone apparatus. The reasons for undertaking the manufacture of these products, some of the major problems encountered in developing suitable processes, and the work done in overcoming these difficulties including several major contributions to commercial methods of manufacturing similar parts are described.

ORIGINALLY, the ceramic parts used in the telephone and associated equipment were not manufactured by the Western Electric Company because the technical requirements and volume of consumption of such parts did not warrant the development or establishment of processes or the facilities for manufacture. The later development of such manufacturing processes for some of the ceramic parts has been necessitated largely by inability to secure an adequate supply of parts meeting the close limits required for satisfactory functioning of the apparatus, although there have usually been other influencing factors. Such developments have, in most instances, been advantageous from an economic standpoint. The experimental work has been confined to that required for the above ends and only a very limited amount of research work has been done. Some of the major projects for which it was necessary to develop new methods of processing to obtain the desired quality at a satisfactory cost are outlined.

SWITCHBOARD LAMP CAP LENSES

The first major project undertaken was the development for manufacture of switchboard lamp cap lenses of the types shown in Fig. 1. One factor necessitating this undertaking was the difficulty experienced

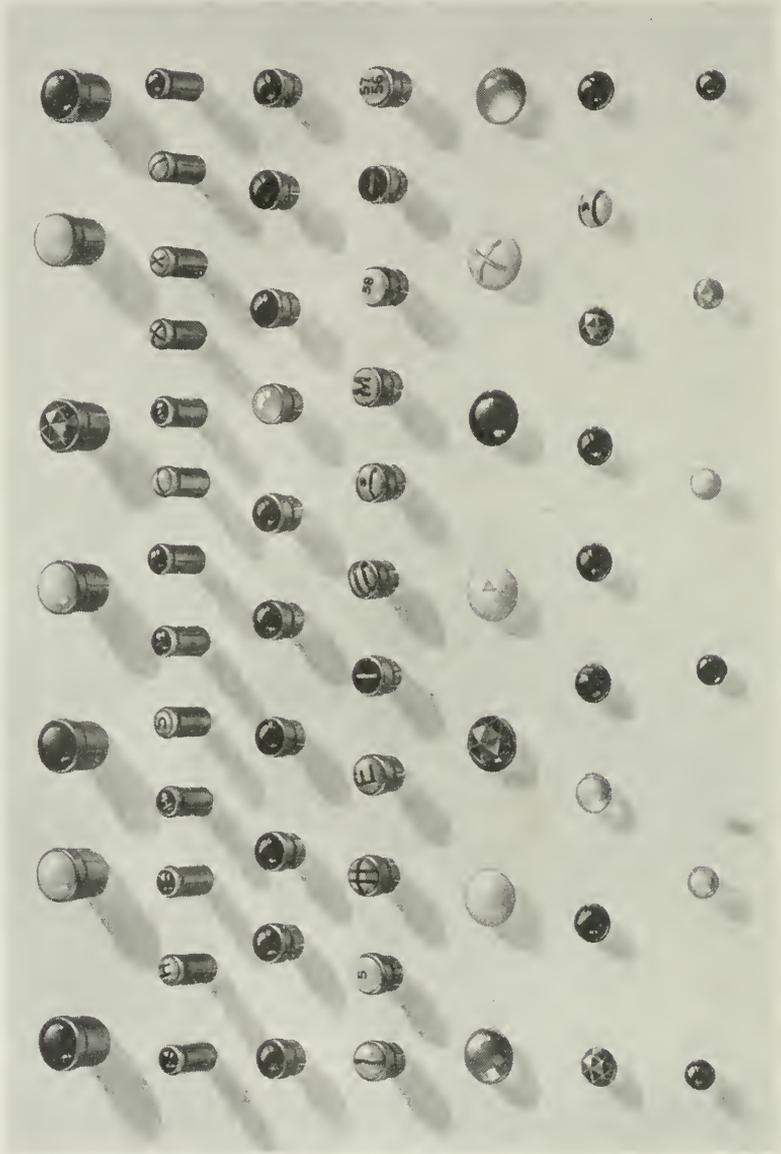


Fig. 1—Switchboard lamp cap lenses.

in obtaining a suitable supply since manufacture of parts of this general class was largely a side line incident to the production of artificial gems. Development of a technique for molding these lenses was therefore undertaken. The method adopted consists of forming the lenses from glass rod softened by heating in gas flames. The forming is performed by using steel dies in a small punch press on the front of which are mounted the gas fixtures.

Since the colored glass rod used in lens manufacture was obtainable at that time only from European sources, with the advent of the World War it was decided to undertake manufacture of the rod. This development required a careful consideration of the characteristics desired in lenses for use in switchboard lamp caps. These lenses must have sufficient light dispersion characteristics to make the lens visible at any angle from which it may be viewed in service. The degree of opacity must also be such as to make either weak light signals visible or to prevent glare with strong light signals. In view of the cost relation between lamp cap lenses and switchboard lamps, variations in the light diffusion and light transmission of lenses must be kept at a minimum since the degree of such variations frequently determines the useful life of the lamps. In addition the lenses must be able to withstand repeated impact shocks from plug tips and the relatively rapid heating in the operation of forming the lenses by punching.

In view of the characteristics desired, glasses having approximately twenty to thirty per cent lead oxide were investigated. Small gas-fired crucible furnaces were used in the preliminary work for batch melting and an impact machine was devised to simulate blows received in service in order that theoretical life tests could be made on the glasses developed. Comparative tests were also made on the light transmission qualities of the various glasses developed.

Originally, light dispersion in lamp cap lenses was obtained with sandblasted lenses of clear glasses. The first work therefore consisted of the development of suitable clear glasses of the desired colors. Silica, sodium oxide, and lead oxide combinations were investigated and approximate limits for these constituents were established to cover glasses of suitable viscosity, durability, and clarity. Final compositions were then evolved by progressive small changes in the amounts of the various constituents used. Since the color and working characteristics of the glasses were influenced by the rates and amounts of heating, the furnace conditions, and other variations encountered in melting and working the glasses, these progressive changes involved considerable time to evaluate properly the results of any composition

change. The resultant compositions are illustrated by the following batch which was developed and used for clear amber glass; and the functions of the various raw materials in this composition are given below:

Glass Sand.....	45
Red Lead.....	30
Sodium Nitrate.....	10
Sodium Carbonate.....	10
Manganese Dioxide.....	3
Ferric Oxide.....	2
	100

Scrap Glass—50 parts approximately.

As is common practice, glass sand was used as the most economical means of obtaining the desired silica content. The sodium content was introduced by the use of sodium nitrate and sodium carbonate. The oxidizing action of the nitrate and manganese dioxide assisted in (1) the prevention of lead reduction; (2) the oxidation of any organic materials present; and (3) the maintenance of the iron in ferric form. The liberation of gas during the decomposition of the sodium nitrate and carbonate tended to stir the glass during melting and in addition the escape of large gas bubbles during this decomposition assisted in the removal of small bubbles of occluded gas. Some of the sodium was introduced as sodium carbonate because it was cheaper than the nitrate. Red lead was used as an economical means of obtaining the desired lead oxide content and to lessen the possibility of any difficulties from unoxidized lead particles. A percentage of glass scrap from the punching and drawing operations was used in each batch as a means of reclaiming the scrap, facilitating melting and improving the working characteristics of the glass when drawn into rods. The amber color obtained in this glass was of course dependent on the predominance of the brown color of ferric iron. If sufficiently oxidizing conditions were not maintained during melting and working, the iron would be reduced to the ferrous state resulting in a greenish color. The color intensity obtained was very sensitive to changes in the amount of heating and to atmospheric conditions in the furnace. This complicated the problem of maintaining the glass within close limits for color and translucency.

After satisfactory glasses with twice the impact strength of the previously imported glasses were developed, open pot manufacture of clear glasses was started on a limited basis.¹ It was then found desirable in order to obtain better signaling characteristics to obtain

¹H. T. Bellamy *Patent* 1,271,652, "Method of Making Colored Glass," July 9, 1918.

the required light dispersion in certain colors of lenses without the use of sandblasted surfaces. Several methods of dispersing the light were tried including the application of a translucent layer of glass on the back of a clear lens, but as it was difficult to control economically the amount of light dispersion by these methods, it was decided to use opalescent glasses. Calcium phosphate and cryolite were found suitable as opacifiers and satisfactory compositions were developed by means of further progressive changes to suit the particular working conditions in the shop.

Several serious objections were found to the open pot method of manufacture, the most important of which were the long heating period required for new pots and their relatively short life. The manufacture of opalescent glasses increased these difficulties because of the more corrosive nature of these glasses as a result of which the maximum life of the pots was approximately twelve days. In view of this, a small 500-pound capacity gas fired melting furnace known as a day tank was designed and constructed. This tank consisted of a rectangular box shaped furnace lined with refractory blocks about twelve inches thick. With this equipment, a complete batch was melted each night and the resultant glass formed into rods during the next day. Under continuous operation, furnace life of about three months was obtained which was considered very satisfactory in view of the corrosive nature of these glasses.

Satisfactory compositions and methods of manufacture were finally developed for the production of glasses in the required colors. This development resulted in the elimination of an unsatisfactory supply situation, reduced the cost of lenses appreciably, and greatly improved the quality of lenses.

SPIRALLY GROOVED RESISTANCE CORES

At the same time that development work on glasses was being carried on, a preliminary survey was made of the advantages of manufacturing instead of purchasing the ceramic cores used in filament resistances. As the preliminary survey indicated that definite advantage would be realized, development for manufacture was undertaken. The part, shown in Fig. 2, consists of a thick-walled tube with a spiral groove on the outer surface in which a resistance filament is placed. Tests first were made on pressing blanks from sodium silicate and powdered slate mixtures. These parts adhered to the die, were difficult to dry, and were very weak in the fired state. Further work was done with talc and sodium silicate mixtures which were stronger but still had the objectionable feature of adhering to the dies. Addi-

tional compositions were then made of talc and clay with and without sodium silicate and feldspar. It was found that most satisfactory results were obtained with a ball clay, kaolin, and talc body in the proportions of forty per cent, ten per cent and fifty per cent, and this body was therefore adopted.

Originally, attempts were made to cut the groove in the fired core with a diamond tool but this resulted in excessive chipping of the

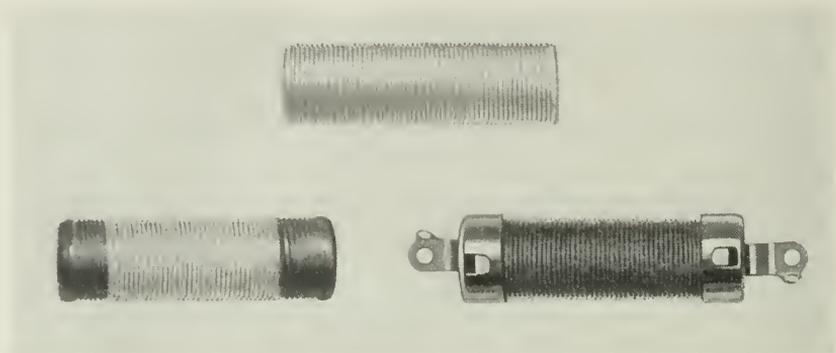


Fig. 2—Ceramic core and completed filament resistance.

groove. A chaser with alternate teeth was then tried out, with the thought that the gradual cutting action would prevent chipping. This also proved unsuccessful. The use of a circular saw, emery wheel, and a phosphor bronze disc charged with diamond dust were also considered. These methods were not completely satisfactory although better results were obtained. Rolling the thread in the core while in the leather hard state after extrusion was then tried with good results and a suitable machine was developed for performing this operation.²

With this machine, an extruded blank of slightly oversize diameter was placed on a revolving mandrel. An arm was provided to hold a shaving tool ahead of a disc which formed the thread. This arm was attached to a segment of a nut and the movement of the arm when the nut segment was engaged with a thread integral with the mandrel, shaved the core to exact diameter and carried the disc longitudinally across the core forming the spiral groove. An auxiliary arm carrying two knives was then engaged which cut the core to exact length and

² H. T. Bellamy *Patent* 1,384,587, "Manufacture of Composition Cores," July 12, 1921.

chamfered the ends. The finished core was then removed from the mandrel, dried and fired. This method of manufacture produced cores of superior quality at a greatly reduced cost.

PORCELAIN PROTECTOR BLOCKS

The next major development was that of the manufacture of protector blocks. These small porcelain blocks, used in open space cut outs and shown in Fig. 3, are illustrative of parts where it was necessary

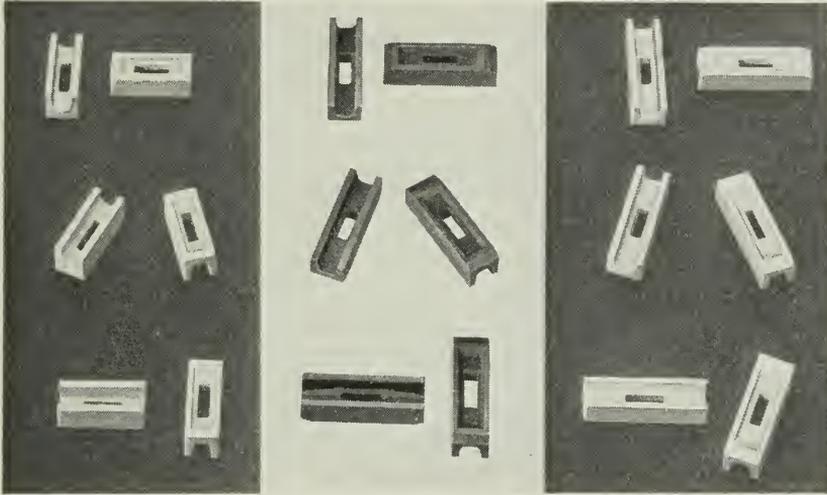


Fig. 3—Porcelain protector blocks.

to undertake manufacture for quality considerations and where such manufacture resulted in cost reduction. They were originally purchased from domestic sources which were unable to meet consistently, required limits as close as $+.020$ -inch and $-.015$ -inch on a $.390$ -inch dimension. The dimensional deviations encountered necessitated sorting to insure proper functioning in the field, and it was necessary to scrap a large percentage of the purchased parts. Difficulties were also experienced in the assembly of the blocks because of manufacturing defects such as small fins and low strength.

Common commercial practice on most porcelain parts of this type at that time was to form the parts on a hand screw press, remove fins by hand, and fire the blocks in refractory containers in intermittent furnaces. The amount of hand labor involved was large and early studies of the economics of manufacture showed it would be necessary to develop new methods of manufacture before

the development and plant expenditures associated with the installation of manufacturing facilities would be warranted. A survey of commercial practices indicated that mechanization of the forming operation and simplification of the finning, firing, and material preparation procedures were possible.

The two general methods of processing first considered were: (1) extrusion of a plastic column having the end cross section of the block, cutting this column to block lengths and forming complete in the plastic state; and (2) automatic pressing of damp granules. The first method offered some advantages in forming because of the thin walls of the protector blocks, but the greater shrinkage from raw to fired states which would result from the use of plastic material would involve greater dimensional variations. Because of this factor it was decided to confine the development effort to a study of the possibilities of automatic pressing of damp granules.

The uses of the porcelain protector blocks required a body as highly vitrified as was consistent with dimensional requirements, to minimize moisture absorption in service and to prevent the adherence of carborundum particles during lapping operations in assembly. High vitrification was also required to insure sufficient mechanical strength to withstand handling during assembly and service. Accurate dimensions were essential for satisfactory functioning in service.

Two general types of bodies were considered, talc-clay combinations and feldspar-clay-silica combinations. An investigation of talc-clay mixtures indicated that the eutectic proportion of the two minerals was approximately sixty-five per cent talc and thirty-five per cent clay with small variations dependent upon various clay compositions. The fusion temperature of this eutectic was approximately cone 12 or 2390° F. This combination, however, was not satisfactory since it softened over an extremely small temperature range and formed a very fluid glass in the melted state. A longer temperature range for softening and greater melted viscosity was obtained by the addition of feldspar. A eutectic composition of twelve and one-half per cent talc and eighty-seven and one-half per cent spar was found which fused at cone 6 or 2174° F. Using ten per cent to twenty per cent of this flux, a well vitrified body was obtained at cone 8 or 2237° F. The firing range of this body was still much narrower than desired and any excess firing resulted in blistering. Although it was evident that commercial use of this body would require extremely close regulation of temperature, it was decided to investigate its pressing characteristics in view of the small amount of abrasive material it contained and the importance of abrasion on dies and equipment with automatic molding.

A study of the pressing behavior of the body under automatic molding speeds and conditions indicated that development work would be necessary to prevent the molded parts adhering to die surfaces. An investigation of this factor indicated that the sticking to dies was caused primarily not by adhesion between the metal and the molded clay surface but rather by the vacuum effect of a dense air-tight layer of material against the metal. This was shown by the facts that the tendency for sticking decreased with (1) a decrease in the plastic content of the body or a decrease in moisture content; (2) a decrease in the viscosity of the die lubricant which thereby tended to clog the pores of the molded surface to a less extent; and (3) an increase in the volatility of the die lubricant. Two methods of overcoming the sticking difficulties with molding compositions were therefore suggested: (1) opening up the structure of the molded part by the use of coarser material to provide capillaries for the escape of entrapped air, and (2) the use of an improved lubricating compound. Since it was not feasible to improve the lubricant sufficiently, an attempt was made to obtain much coarser talc. The talc normally available at that time was such that on sieve tests approximately five per cent to ten per cent remained on the 300-mesh screen. The availability of coarser talc was investigated and it was found that material coarser than eighteen per cent on 300-mesh was not available at an economical price. In view of the fact that the talc was very fine grained and non-plastic, it gave a very dense molded structure without contributing materially to the strength required to hold the molded part together. It therefore seemed advisable to use a clay, feldspar, and silica body and to minimize abrasion by the selection of suitable tool steels and the proper design of equipment.

In arriving at a suitable body composition of the feldspar type it was decided to use a composition which would mature at about cone 12 or 2390° F. Sufficient feldspar was used to obtain a low porosity when fired over a reasonably wide temperature range. The amount of clay used was governed by the raw strength required. Enough silica was used to obtain sharp definite outlines and to avoid warpage. The following composition was arrived at:

Flint	22.5
Feldspar	37.5
Ball Clay	20.0
Kaolin	15.0
China Clay	5.0
	100.0

Further development work was then confined to methods of processing this body to obtain satisfactory results on an automatic machine. A survey of available commercial pressing equipment indicated that machines of the type used in the manufacture of various pharmaceutical tablets or pellets offered the most promise for adaptation to molding protector blocks. The development of suitable equipment was complicated by the extremely thin walls of the parts and the necessity for rapidity of operation. In the hand molding method commonly used in the industry, a slow application of the molding force was possible at the end of the stroke and likewise a gradual withdrawal of the top die was possible after completion of the forming operation.

After some preliminary work with various types of tableting machines, we concentrated our efforts on single-plunger-type machines with double dies. One of the major problems was a satisfactory method of die lubrication since with the machine operating at twenty-eight strokes a minute, the die surfaces were exposed for oiling only an instant during each cycle. The use of an atomizer-type device with a mixture of lard oil and kerosene was finally adopted with the amount of lubricant closely controlled by oil sight cups. Exact timing of the application of the spray to dies was obtained by automatically operating air check valves. This method proved more satisfactory than wiping with saturated felt or incorporating a lubricant in the body particles before molding.

The various stages of the molding cycle are shown in Fig. 4. The cycle of operation at twenty-eight strokes per minute was as follows: as the bottom die reached the lowest position, a feed hopper was vibrated over the cavity. The withdrawal of this hopper removed excess material, after which the top punch descended forming the part. The bottom and top punches then moved upward until the bottom of the part was flush with the top of the die. A projection on the hopper then pushed the part free. Before the hopper reached a position over the cavity, any particles of clay adhering to the dies were blown off and the lubricant was sprayed over each die. The lower and upper dies were then returned to the original positions.

A large amount of work was also necessary to adjust the size and moisture characteristics of the pressing material not only to secure well formed parts and prevent sticking but also to secure fired parts meeting the desired requirements. A mixture of colored and uncolored particles was used in the study of these characteristics in order that the flow movements in the die during compression could be studied. As a result of this work, it was found that most satis-

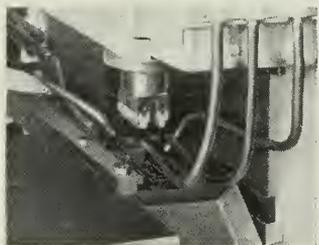
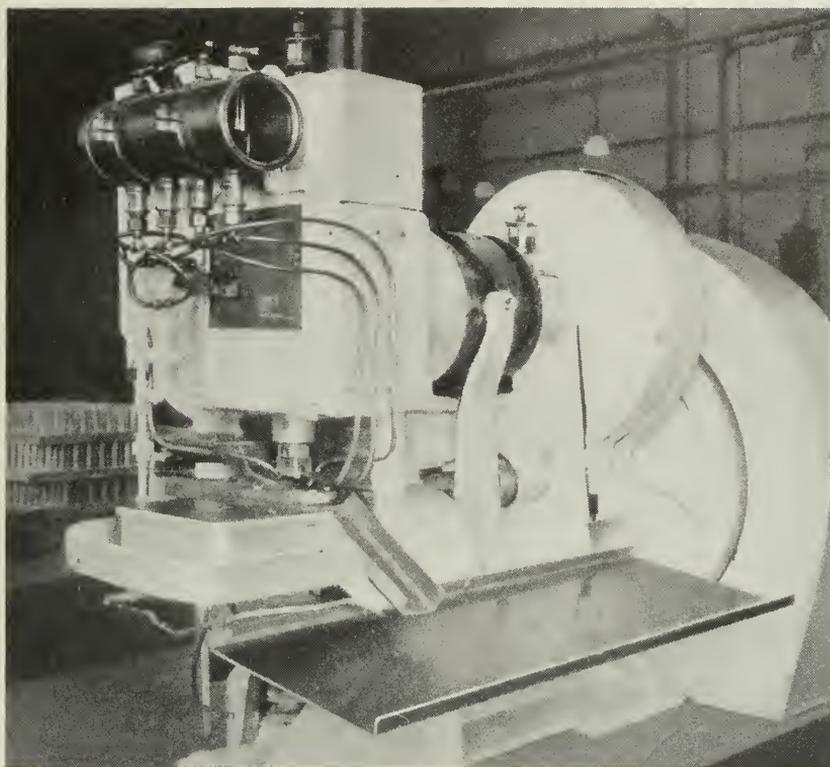


Fig. 4—Block forming press.

factory results could be obtained with 10–24-mesh material and a 12.75 per cent \pm .75 per cent moisture content. This moisture content was sufficient to give a compact part and body mix still could be fed satisfactorily to the dies. Methods of processing the body mix were then worked out to hold it within these limits.

The use of material within these close size limitations involved considerable effort to establish economical methods of production. Common practice consisted of slaking the clay in water, adding the other body ingredients, mixing thoroughly with water, filter pressing, complete drying, addition of water to obtain the desired moisture content, aging and screening. The effect of aging was investigated and found negligible with the moisture content to be used and it was therefore decided to dry the material after filter pressing to the required approximate twelve per cent moisture content before the disintegrating and sizing operations. Methods of handling were evolved to obtain a maximum percentage of material between 10 and 24-mesh and to regranulate the fines without again mixing them with excess water.

Various methods of economically removing fins after forming were investigated and initially the fired parts were tumbled with small porcelain balls. This method removed fins and produced smooth surfaces. Another advantage of the method was the automatic elimination of any weak or flawed parts by breakage during the tumbling. Later, further developments in methods of firing described hereafter made it more economical to remove the fins in the raw state by vacuum brushing the parts in multiple after they were arranged on trays at the pressing machines.

Initially the parts were fired using the practice then commonly followed in the industry. With this method, the parts were placed in saggars and fired in an intermittent kiln. This method involved costly handling, heat losses due to heating and cooling the furnace at each firing, and considerable expense from sagger replacements. A small continuous kiln was therefore installed in which the parts were carried in layers on top of cars through successive preheating, firing, and cooling zones which were continuously maintained at definite temperatures, the heat from the cooling fired ware being used to heat the incoming ware.

Summarizing, the method of manufacture finally developed for porcelain blocks consisted in mixing feldspar, clay and flint with water to get an intimate mixture, filter pressing, drying to proper moisture content, sizing, automatically molding the parts, removing fins in multiple, and firing in a continuous kiln. This method resulted in a

marked improvement in quality and reduced the cost of parts. The method of automatic pressing developed constituted a major contribution to existing commercial methods of manufacturing small porcelain parts.

VITREOUS ENAMELED COPPER BASE NUMBER PLATES

Manufacture of parts similar to the vitreous enameled number plates used on calling dials and shown in Fig. 5 was limited to producers

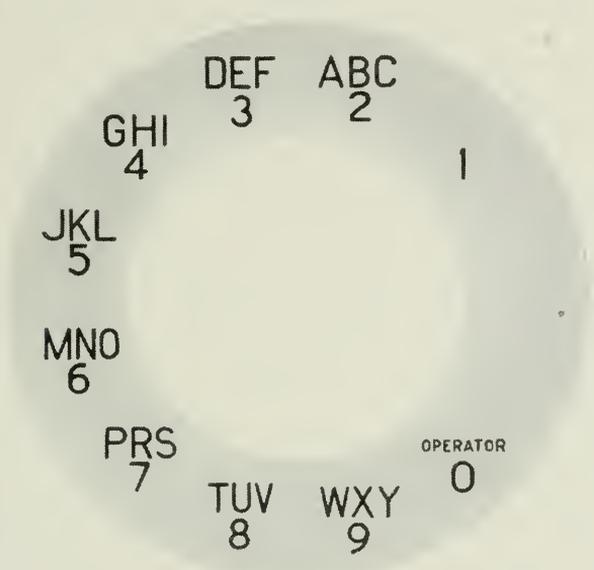


Fig. 5—Copper base number plate.

of enameled parts of the watch dial type. Because of our unusual requirements for dimensions and character location together with the need for a collar and locating pins, only one source of supply could be developed. With rapidly increasing schedules for calling dials and because of the possibility of conditions beyond the control of the supplier interfering with the continuity of supply, this situation was unsatisfactory and made it advisable to undertake manufacture to remove possibilities of any embarrassment from a supply standpoint. As sources of supply for copper enamels were also limited, various enamel compositions were investigated. It was found that the following composition would give an enamel satisfactory for color, texture, gloss, fusibility, and durability when fired on copper blanks:

Red Lead.....	40
Pearl Ash.....	6
Sodium Nitrate.....	9
White Arsenic Oxide.....	6
Flint.....	31
Borax.....	8
	100

In practice constituents of this enamel were first thoroughly melted to a homogeneous glass, giving on cooling a glass magma having high opalescence. The dead white opacity of this enamel could only be developed by slow cooling through the range necessary to precipitate the arsenic compounds. Manufacturing considerations, such as the necessity of an enclosed room for commercial smelting to avoid contamination as well as to avoid the possible health hazards involved in the smelting of arsenic-lead combinations, led us to purchase the required enamel. The fact that a suitable composition had been developed and was available for manufacture if necessary was an advantage from a supply standpoint.

Various enameling procedures were considered. In order to cover the vertical surface of the collar satisfactorily, it was essential that this surface be coated either by dipping or by spraying. It was equally important to apply the enamel coating to the flat surface of the plate by dusting on a thick coat of dry powdered enamel. This dust coat was necessary because of the thickness of enamel required on the flat portion to strengthen the number plate and also to obtain the desired quality of finish on the surface bearing the numerals and characters. From an economic standpoint, it was also imperative that only one enamel fire be used. Initially, efforts were made to dust enamel on a blank already completely coated with a thin coat of enamel slip consisting of finely divided enamel frit suspended in water by means of clay or bentonite. It was found on firing that the added refractoriness of the enamel slip containing the clay or bentonite resulted in a roughened fired surface over the dusted area. This was caused by the formation of gases in the decomposition of the clay or bentonite while the enamel was in a viscous state. It was therefore necessary to protect the flat portion of the plates by templates during spraying. As this was costly, a study was made of other means of floating the enamel frit for collar application.

In order to overcome these process difficulties, it was desirable to find a material which would (1) satisfactorily hold the heavy lead enamel particles in suspension and prevent packing, (2) not attack the enamel or impair its durability, (3) decompose before the enamel started to fuse, and (4) be inexpensive. Soluble alginates appeared to

possess these properties and excellent results were obtained from their use.³ These substances were made from kelp. Their most interesting property as a suspending medium was the ability of the alginates when added to water even in small percentages to make solutions of high viscosity. For example, water solutions of ten per cent ammonium alginate would stand stiff. Some of the advantages in our use of alginates for suspending number plate enamel were: (1) uniformity of composition resulting from the alginates being a manufactured product rather than a natural mineral; (2) the fact that dried sprayed coats of alginate suspended enamel were less subject to damage from handling; (3) a low decomposition temperature which resulted in the material being driven off before fusion of enamel, thus avoiding bubbles in the enamel; and (4) increased resistance of the finished enamel surfaces to chemical attack and their ability to withstand greater mechanical shock and distortion without damage, since any refractory materials present when the enamel was fired would not be completely fused or incorporated into the glass, leaving points more readily attacked chemically as well as lines of mechanical weakness.

Using alginate suspended enamels, suitable manufacturing processes were developed for the application and firing of enamel and the application of characters to the fired plates. A machine was devised for the application of the sifted coating, and rotary continuous furnaces were installed for the firing operations.

Originally the decalcomania method was used for character application. In this process, the enameled parts were first coated with a thin coat of sizing and, after partial drying, they were placed in a locating fixture mounted on a small arbor press and pressure was applied to a properly located transfer by means of a soft rubber pad. The paper backing of the transfer was then removed by soaking in water and, to insure contact, the characters were repressed with a silk covered pad. The sizing was then baked off before firing to remove organic materials and eliminate shadows around the characters. This method was costly and even well trained, careful operators did not produce satisfactory plates.

To eliminate these defects, an offset printing and dusting method was developed in which an electrotype printing plate was covered with printer's ink and an impression was transferred to the number plate by means of a rubber transfer pad. Powdered vitrifiable colors were then dusted over the entire surface of the plate and the unprinted areas of the part brushed clean with a camel's hair brush. In printing two color plates, the black letters were printed and dusted first, after

³ L. I. Shaw *Patent* 1,806,183, "Suspension," May 19, 1931.

which the red numerals were printed and dusted. By using a special black powder which would give an intense black in combination with a thin film of red powder, one firing for both colors was possible. This method required close control of temperature and humidity of the air in the room which was therefore air conditioned.

Even under good conditions considerable difficulty was experienced at times with the adherence of the powder to unprinted areas. In addition, the application of the powder and the brushing operation required the installation of a special well exhausted unit and involved some problems in the recovery of the ceramic dust which were quite expensive. Efforts were therefore made to incorporate the glass powder directly in the printing vehicle. The development of a

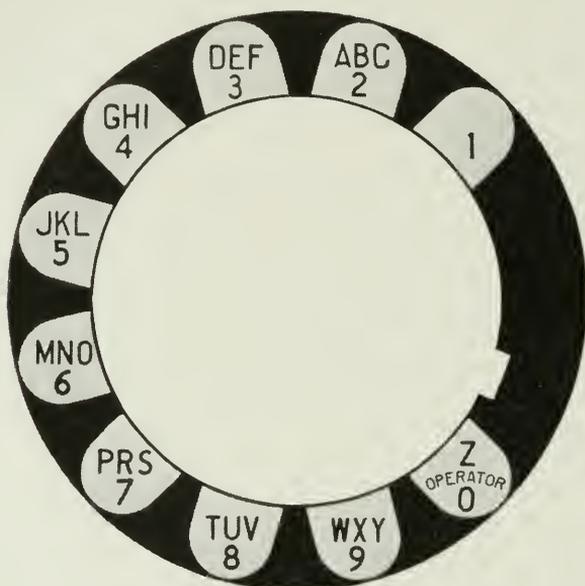


Fig. 6—Iron base number plate.

ceramic printing ink covered tests on various printing vehicles to determine what vehicle or mixture of vehicles would be most suitable. Difficulty was encountered in incorporating a sufficient amount of inert, finely pulverized, intensely colored glasses into a vehicle and still retaining the properties essential for offset printing by transferring an impression from an electrotpe plate to a vitreous enamel. This problem was finally solved by the use of a relatively large percentage of uncalcined ceramic material in combination with light and heavy

ink varnishes.⁴ Motor driven presses using this ink were also developed to facilitate the printing operations.⁵

The manufacture of these parts was undertaken primarily to eliminate an undesirable supply situation but Western Electric manufacture resulted in improvements in quality of enameling, quality of printing, and in mechanical strength. This latter characteristic was important since it reduced assembly losses from cracked plates.

While vitreous enameled copper base number plates have been replaced by other types, the developments outlined were the basis of subsequent enameling developments.

VITREOUS ENAMELED IRON BASE NUMBER PLATES

The low level of illumination at some pay stations led to the design by the Bell Telephone Laboratories of a large iron base number plate, shown in Fig. 6, to be mounted flush with the finger wheel of the dial. Since the demand for these plates was relatively small, they were originally made by the usual process followed in the industry in enameling similar articles. This process consisted of applying and firing one ground coat for adherence and then applying and firing two sprayed cover coats to obtain the whiteness and opacity desired; all being felspar enamels. The whiteness was not as good as that obtained on the copper base plates with lead enamels and in addition considerable difficulty was experienced in the field due to the fading of the characters as a result of chemical action on plates exposed to corrosive gases such as sulphurous fumes in certain locations. The process was also costly.

Since maximum whiteness and opacity was obtainable in the lead-arsenic type of enamels previously described when applied by dusting on dry, it was desirable that the coating be applied in this manner. In order to avoid several enamel applications and firings, it was also desirable that other portions of the plate be protected by some corrosion resistant coating other than vitreous enamel which would necessarily have to retain such corrosion resisting properties after exposure to a temperature of 1500° F. for six minutes and also be capable of being enameled with satisfactory results. Numerous coatings were tried and it was found that a Western Electric black oxide finish on iron would satisfactorily meet all requirements.⁶ Using this finish, it was possible to fuse the enamel directly on the upper surface of plates, to retain corrosion resistant qualities on all other exposed surfaces, and to reduce the number of process operations. A number plate of greatly improved appearance and durability also resulted. In addition, the curved

⁴ L. McLaughlin *Patent* 2,030,999, "Ink," February 18, 1936.

⁵ L. McLaughlin *Patent* 1,951,430, "Printing Apparatus," March 20, 1934.

⁶ W. J. Scott *Patent* 1,962,751, "Ceramic Coated Articles," June 12, 1934.

surface obtained in dusting a base plate having a groove around the edge prevented the entrapment of air between the plate and the printing pad during the printing operation, thereby resulting in a simplification of that process.⁷

As a result of our development of enameling over the black oxide finish it would have been possible to replace the previously described copper-base number plate by one employing a sifted coat of enamel over such finish on a steel blank. However, the application of the black oxide finish on enameling iron was so costly that manufacture of number plates by this process was not competitive. We therefore continued our developments and found that it was possible to enamel directly over an electroplated copper-nickel finish consisting of a minimum of 25 m.s.i. each of copper and nickel on a mild steel blank and get a smooth enamel coat having very good adherence.⁸ As this finish had the necessary rust resistance and the blank was relatively flat, the enamel could be applied in a single sifted coat on the face only to produce a satisfactory number plate. Also with the steel base it was not necessary to have a thick coating of enamel for strength as was the case with the copper base number plate. In fact, due to the good adherence of the enamel coat, if the thickness of enamel after firing was less than 0.010 inch the plate could be flexed considerably without chipping the finish. On the other hand, it was necessary to have a minimum of 0.007 inch of enamel to hide sufficiently the gray color of the nickel surface. Additional refinements of the enameling process were effected by improvements in the uniformity of enamel distribution and in the printing of characters; and the process was generally automatized. These developments produced a number plate of superior quality and appearance at a reduced cost. As all final details for commercial manufacture have not been completed further details of this process will not be given here.

VITREOUS ENAMELED RESISTANCES

With the increased use of panel-type machine switching, the demand for vitreous enameled resistances for controlling the current for operating relays and switches increased materially and manufacture of these parts was undertaken. These resistances were required to dissipate a considerable amount of heat in service and to reach a high operating temperature without being damaged. The units therefore consisted of a suitable resistance wire wound on a ceramic core and covered with a vitreous enamel. Some of the types now being manufactured at Hawthorne are shown in Fig. 7.

⁷ W. J. Scott *Patent* 2,020,476, "Ceramic Articles," November 12, 1935.

⁸ S. R. Mason and W. J. Scott *Patent* 2,020,477, "Ceramic Article," November 12, 1935.

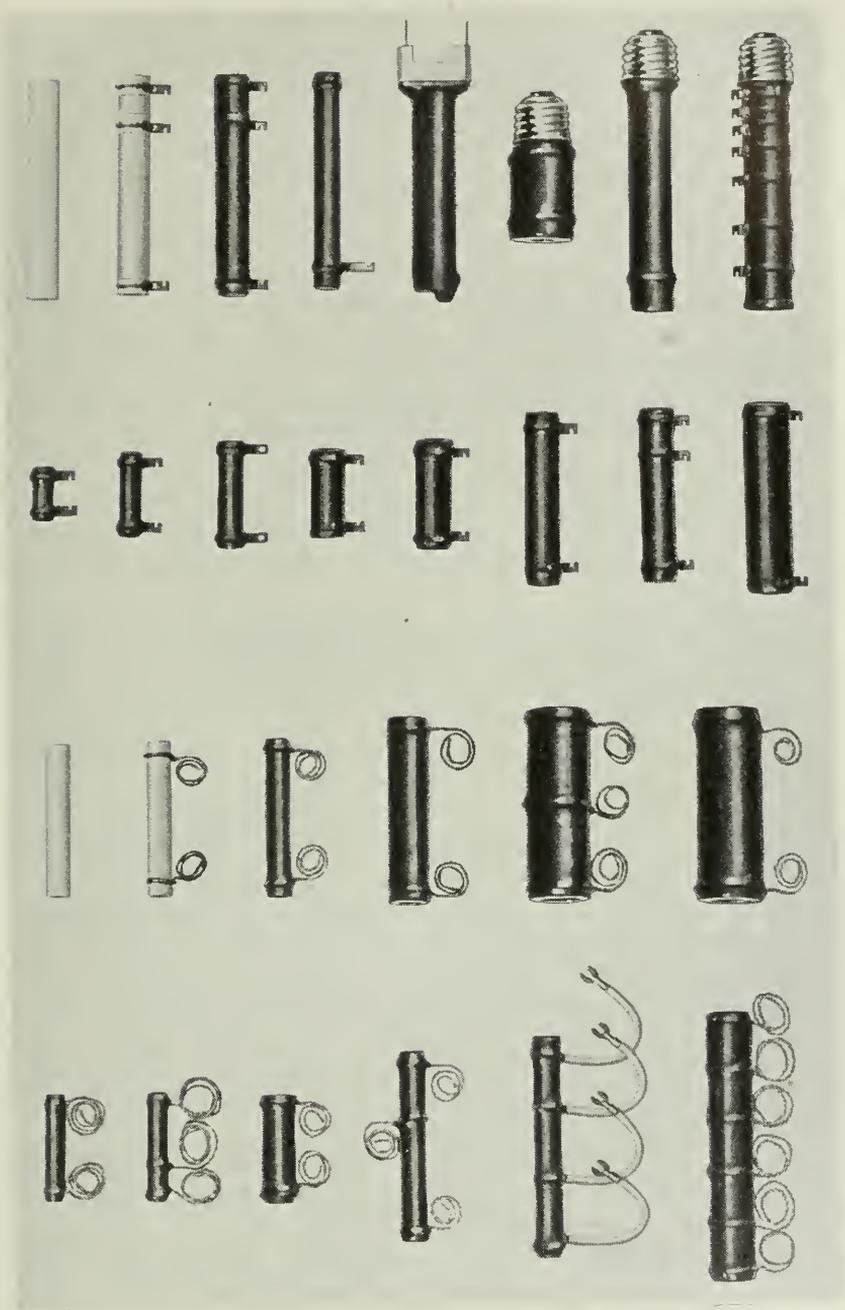


Fig. 7—Ceramic cores and vitreous enameled resistances.

Since the application of vitreous enamel to the resistances involved plunging the porcelain cores into a hot furnace and their removal into cold air without cracking, and since in use they were also subject to considerable heat shock, it was necessary to develop a body with thermal shock resistance characteristics that would still have the necessary fired strength. It was also required that this body be suitable for the extrusion of round cores and for the molding of more complicated shapes from a granular body. A somewhat porous fired structure was also desirable to facilitate the application of the enamel.

These desired characteristics indicated that a clay-talc combination would be suitable. Mixtures containing a greater percentage of clay than the eutectic mixture of the two minerals were investigated to avoid vitrified parts at the temperatures then used in the tunnel kiln for other products. Tests were made of extrusion and molding properties, fired, breaking and impact strengths, and resistance to thermal shocks. On the basis of these tests a talc-clay body composition was selected. Dies were then constructed based on the fired shrinkage of this body of about fifteen per cent in extruded and twelve per cent in molded forms.

Using these cores, suitable sizes of wire were selected considering the dimensions of the core, the resistance value desired, heat to be dissipated at certain wattages, and the maximum operating temperature permissible, and satisfactory winding methods were established. These methods were based on the use of motor driven machines in which the wire was spaced on the revolving cores by the transverse movement of a guide controlled by lead screws of various pitches. To facilitate any necessary minor resistance adjustments, methods were provided for checking the resistance of the units before connecting the wire to the second terminal. Since the heating received by the wire during the enamel firing increased its resistance value, tests were made to establish resistance value factors for winding use.

Difficulty was experienced with the resistance becoming open in the firing operation and it was shown that this condition was caused by the formation of a film of glass between the resistance wire and the terminal during the firing operation. Various methods of attaching the resistance wire to terminals were tried and it was found that the use of lead as solder would prevent the formation of a film of glass between the wire and the terminal even though the fusion temperature of the lead was considerably below the enameling temperature. Ordinary soft solder could not be used due to the tin content embrittling the standard copper lead wires during the enamel firing. While silver solder could be used it was costly both for material and in application.

In developing an enamel to be used for resistances, it was desirable that the melting temperature be as low as possible consistent with good durability in order to maintain at a minimum the thermal shocks received by the porcelain cores and any changes in the resistance of the wire during firing. A very high viscosity during fusion was also desirable in order to avoid running of the enamel during firing, an undesirable feature which would result in exposed wires and unsightly lumps unless the enamel was applied in numerous thin coats. Conversely to these requirements, it was necessary that the enamel coating be glassy in appearance, smooth, free from blisters and pin holes, and capable of being fired in a relatively short time.

These factors indicated the desirability of investigating lead-boron-silica mixtures and the elimination of any raw clay or similar refractory substance in the enamel slip. The enamel finally developed was as follows:

Red Lead	48.0
Boric Acid	24.0
Flint	10.0
Soda Ash	3.7
Cryolite	6.0
Tin Oxide	1.6
Manganese Dioxide	0.5
Cobalt Oxide	0.6
Iron Oxide	4.0
Zinc Oxide	1.6
	100.0
White Lead	10.0
Light Calcined Magnesia	1.3

All of the materials other than the white lead and light calcined magnesia of the above composition were fritted or melted to a glass and then quenched in water. The fritting of these materials was done to insure complete formation of stable compounds and to permit more rapid firing of the enamel coating on resistances to a smooth homogeneous glass. The proportions of sodium, lead, boron and silica were selected to obtain a stable coating with the desired viscosity characteristics at as low temperature as possible. Sufficient opacity of the coating was obtained through the use of cryolite and tin oxide. The cryolite also functioned as a flux. A pleasing dark color was obtained economically with the iron, cobalt and manganese contents. The zinc oxide functioned as an additional flux and also aided considerably in the formation of a smooth coating. Slight variations in the sodium content of this enamel affected its viscosity markedly and also affected its expansion characteristics.

After fritting, the resultant glass was sized and suspended in a water suspension of white lead and light calcined magnesia in a tank provided with mechanical agitation. This method of suspension aided in increasing the fired viscosity without the formation of blisters and pin holes. Smooth glassy resistances were obtained with this enamel in a ten minute firing at 1150° F. without appreciable bubbling or flowing of the coating. This eliminated the necessity of three or four thin fired coats and resultant greater variations in resistance values after firing.

CLOSE TOLERANCE CERAMIC BARRIERS AND INSULATORS

In the design of the handset type of telephone transmitter, it was found desirable to use a thin washer type insulator, shown in the lower portion of Fig. 8, as a barrier to control the path of the current between



Fig. 8—Close tolerance insulators.

electrodes. This necessitated very close dimensional tolerances, unusual freedom from surface and edge defects, and reasonable strength to withstand the clamping force used in assembly.

Various materials such as fiber, lava and metal coated with vitreous enamel, were tried and lava gave the most promising results. In view of the cost of lava parts, experiments were made using the usual process of dry pressing a porcelain body in which the clay content furnished the raw strength. The difficulties inherent with this process were the fragility of the raw part and the variable dimensions resulting from uneven drying and firing shrinkages. Because of the fragility of the parts, it was necessary to mold them 0.050 inch thick and then lap the

fired parts to the desired thickness of 0.030 inch. This operation was costly and losses from breakage were high. The narrow dimensional limits of ± 0.002 inch on thickness were also hard to maintain because of the difficulty of keeping the lapping surfaces parallel. In view of this, it was decided to machine the parts from natural talc rod or lava.

The mineral talc or lava, being soft, was easy to machine and the firing shrinkage was only one per cent as compared to about ten per cent with dry pressed porcelain. While less difficulty with warpage and dimensional variations was experienced, the machined surfaces, while reasonably smooth and accurate, were not equal in quality to surfaces obtainable with molded parts. The chief difficulty with the process was in obtaining a satisfactory raw material free from flaws and fissures. The first work was done with domestic lava which was somewhat granular in structure but large rejections resulted from pitted surfaces and chipped edges. A survey of domestic lavas showed that only a small percentage was sufficiently dense. Chinese white lava was found to be homogeneous and fine grained but of uneven shrinkage. Best results were obtained with Italian green lava and this material was used in commercial production. Due to breakage because of fissures, the number of good insulators per foot of rod was very low and the manufacturing cost was therefore excessive.

In view of this, various domestic manufacturers of glass, porcelain, lava and other types of ceramic parts were canvassed but no source of supply that could meet the required quality limits could be located. It was therefore decided to make a thorough investigation of new molding compositions for the job. As a first step in this study, it was necessary to do away with drying shrinkage which required a binder which would give sufficient strength in the raw state to withstand the various fining and handling operations prior to firing. It was also desirable that such a binder should not affect the fired structure of the parts. Various organic substances such as pitches, phenolic resins, asphalts, paraffins, and waxes were tried in both hot and cold molded bodies. It was found that a large percentage of these binders could be incorporated into a body without deformation during firing.⁹ As a mixture of paraffin and carnauba wax was found satisfactory for cold molding and in addition possessed sufficient hardness to furnish the necessary molded strength, this combination of materials was chosen for the binder.¹⁰

⁹ W. J. Scott *Patent* 1,847,102, "Ceramic Material," March 1, 1932. W. J. Scott *Patent* 1,977,698, "Ceramic Material and Method of Making the Same," October 23, 1934.

¹⁰ L. I. Shaw and W. J. Scott *Patent* 1,847,197, "Ceramic Material and Method of Making the Same," March 1, 1932.

Another major factor in the development of a suitable molding compound was the abrasive effect of the molding body on the die parts. Because of the close tolerances required, this factor was important in order to avoid excessive tool expense. Talc was therefore chosen as the chief body constituent to obtain a long die life. The balance of the body was made up of twenty-five per cent clay which gave the desired density in both molded and fired states. With the talc-wax compound, a long die life was obtained even with the close tolerances required. As a result of the use of a combination of waxes as a binder this composition had a low uniform shrinkage of approximately four per cent as compared to about ten per cent with most dry pressed porcelains. In addition, variable shrinkage and warpage resulting from drying strains were eliminated.

In molding this body, the lubrication of die surfaces was found to be critical because of the extreme thinness of the part. It was impracticable to apply a sufficiently exact amount of a liquid lubricant to prevent the parts from either adhering to the dies or being weakened from the absorption of the liquid. This problem was overcome by tumbling the granulated molding material with a fraction of a per cent of zinc stearate.¹¹ The stearate coated grains of material were then molded without any additional die lubricant.

Using the above composition, the process developed was as follows: The talc and clay were thoroughly milled in a carbon tetrachloride solution of the waxes. After drying, this mixture was disintegrated and sized, after which the particles of compound were coated with zinc stearate. The parts were then molded four at a time in a commercial self-contained hydraulic press within an accuracy of ± 3 per cent of the total thickness and ± 1 per cent of the inside diameter. After molding, any fins were removed and the parts trimmed within ± 1.5 per cent of the total thickness in a finning machine which was an adaptation of a commercial automatic indexing head drill press. In this machine, the parts were fed to a rotating end cutter by a revolving indexing head and were held under this cutter by a vacuum applied to the underside of the parts. Tungsten carbide cutters were used to obtain long tool life. After finning, the parts were fired in small trays in a continuous kiln. The parts were then individually gauged for thickness, roundness and inside diameter and individually inspected for cracks, flaws, and burrs. They were then examined under a 10 to 1 glass for smoothness and regularity of inner edge before being used in the assembly of the transmitter.

¹¹W. J. Scott *Patent* 1,847,196, "Ceramic Article and Method of Making the Same," March 1, 1932.

This development permitted the manufacture of ceramic insulators within limits not feasible with other methods of manufacture at that time except by machining from mineral talc and to closer dimensional tolerances than ever before attained in molded ceramic parts. The cost of the parts was reduced to a fraction of that of machined parts and their quality was greatly improved. Since that time the process has been used in the manufacture of other close tolerance ceramic parts for telephone use such as the insulator shown in the upper part of Fig. 8.

Although the outline of miscellaneous manufacturing developments given herein does not include all of the engineering development effort on glass, porcelain and vitreous enamel problems it gives a general picture of the type and scope of past engineering work in the production of ceramic articles for telephone apparatus. The miscellaneous ceramic parts used in telephone apparatus were described in an earlier publication.¹²

¹² A. G. Johnson and L. I. Shaw, "Ceramics in the Telephone," *Industrial and Engineering Chemistry*, Vol. 27, pp. 1326-1332, November, 1935.

The Production of Ultra-High-Frequency Oscillations by Means of Diodes

By F. B. LLEWELLYN and A. E. BOWEN

The general problem of obtaining oscillations by the use of diodes with critical electron transit time is outlined. Some of the properties of a 10 cm. oscillator tested experimentally are included. Extraneous losses were reduced when the oscillator was enclosed within a wave guide.

THE theory of the production of negative impedance by means of an electron discharge between two parallel planes has been known for some years.¹ The negative resistance appears whenever the electron transit time is approximately $1\frac{1}{4}$, $2\frac{1}{4}$, $3\frac{1}{4}$, etc. cycles of a given high-frequency current. Using this property, Müller was able to construct tubes giving 100 cm. oscillations.² The operating efficiencies were quite low, and in the frequency range covered by these tubes it seems fairly conclusive that other methods of producing oscillations are more effective than the critical transit time diode. However, there is promise in the application of diode operation to much higher frequencies than those of Müller.

In a diode where the electron discharge occurs between two parallel planes where one performs the function of electron emitting cathode and the other constitutes an anode biased at a positive potential, the effective impedance presented to an external source is inherently low in magnitude. This is because of the capacitance between the two planes which causes the decrease in impedance at high frequencies. For the production of oscillations, the capacitance must be combined with a resonant structure having the proper inductance to resonate at the desired frequency and having a resistance which effectively is less in magnitude than that of the electron stream. Because of the low losses thus required of the coupling or tuning circuit the properties of concentric lines and of tuned cavities offer a favorable method of attack. These structures also have the property that the impedance presented to the diode proper may be made low to match its capacitive reactance at the high frequencies desired.

The two most important sources of circuit resistance are ordinary ohmic loss modified in the usual way by skin effect in the conducting

¹ For numbered references see end of paper.

material forming the resonant system and secondly the losses caused by radiation of energy. These latter are extremely important where the negative resistance is only a few ohms as in the present instance and necessitate the use of nearly closed structures. This again directs attention to the properties of cavities and concentric lines tuned by internal capacitive resonance, the low capacitance being formed by the electrodes between which the electron discharge flows. It was on the basis of these principles that the actual diode models were constructed.

The general aspect of these tubes is shown in Fig. 1 which presents a section through the axis of revolution. The cylinders of radii r_1 and r_2 respectively constitute the outer and inner conductors of a concentric line. At one end of the inner conductor a flange partly closes the system thus confining most of the energy within the cavity. At the other end of the inner conductor the flat surface of the inner conductor constitutes an emitting cathode while the opposing surface of the outer conductor constitutes the positively biased anode which also completely closes the end of the cylinder. The system is tuned by the capacitance between cathode and anode and the effective inductance of the coaxial line of length h . The emitter was coated in the experiments with an oxide of the uncombined type and was heated by a filament located within the inner cylinder. The spacers for separating the inner cylinder from the main body of the outer conductor were composed of fused quartz in order to obtain low losses and good mechanical rigidity. A water jacket was supplied to assist in cooling the anode.

In reference to Fig. 1, the tuning relation between the cathode-anode capacitance and the inductance of the resonant circuit connected to it requires the following relation to be satisfied,

$$\frac{1}{\lambda} \tan \frac{2\pi h}{\lambda} = \frac{x}{\pi r_2^2 \log_e \frac{r_1}{r_2}}. \quad (1)$$

Here λ stands for the free space wave-length. The other quantities in the formula are illustrated in Fig. 1 and all dimensions are in centimeters. The radii r_1 and r_2 refer respectively to the inner surface of the outer cylinder and the outer surface of the inner cylinder. Improved formulas for the resonant frequencies of cavities of this type have recently been published by Hansen.³

The formula (1) is based on the approximation that the presence of electrons between cathode and anode does not affect the dielectric

constant of the resulting capacitance. This approximation is a good one when the electron transit time is greater than a cycle, as is the case here.

The next design formula required is the resistance of the electron

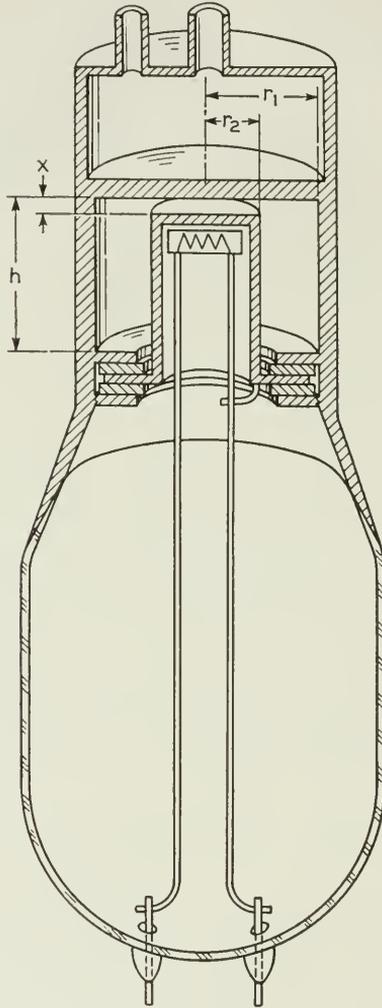


Fig. 1—Ten-centimeter diode used in tests.

Diode	r_1	r_2	x	h
No. 24	1.270	0.635	0.203	1.870
No. 37	1.220	0.635	0.105	1.870

Centimeters

stream. Reckoned per square centimeter of area this may be written,^{4*}

$$r_p = \frac{1.78 \lambda^4 I_0}{10^4} [2(1 - \cos \theta) - \theta \sin \theta] \text{ ohms for cm.}^2, \quad (2)$$

where I_0 is the direct current density in amperes per square centimeter flowing to the anode and θ is the electron transit angle, given by

$$\theta = \frac{Ax}{\lambda \sqrt{V_0}} \text{ radians.} \quad (3)$$

Here V_0 is the constant potential difference in volts between the cathode and anode and A is a numerical factor which depends upon the amount of space charge within the electron discharge, being equal to 6300 for negligible space charge and to 9500 for complete space charge with intermediate values for intermediate space charge. As an alternative the resistance (2) may be written

$$r_p = \frac{12 r_0}{\theta^4} [2(1 - \cos \theta) - \theta \sin \theta] \text{ ohms for cm.}^2, \quad (4)$$

where r_0 is the low-frequency series resistance of the device. With space charge, r_0 is the slope of the static characteristic derived from Child's equation

$$I_0 = \frac{2.33}{10^6} \frac{V_0^{3/2}}{x^2} \text{ amperes/cm.}^2. \quad (5)$$

More generally r_0 is given by the expression

$$r_0 = \frac{1.48}{10^5} I_0 \frac{x^4}{V_0^2} A^4 \text{ ohms for cm.}^2, \quad (6)$$

where A is the same as was defined under (3).

Figure 2 shows a graph of the electron stream resistance as a function of transit angle and is repeated from previous papers.¹ However, it may not have been emphasized in the literature that the graph as well as equations (2) and (4) apply not only with complete space charge but with intermediate values when interpreted correctly, namely in terms of the d-c. current density I_0 rather than in terms of the applied potentials.

Whenever the transit angle is equal to $2\pi n + \frac{\pi}{2}$ where n is 1, 2, 3,

* Equation 41 in this reference applies where the initial velocities are very small. With complete space charge $q = J$ and $a_a = 0$ whereas without space charge $q = 0$. Either condition gives the same series resistance in terms of I_0 .

etc. then the electron stream exhibits a negative resistance. From this it may be inferred that oscillations are possible not only for values of n equal to unity but also for larger values, thus yielding the possibility of higher order oscillations when the circuit coupled to the electron stream is properly proportioned. If we start with a given cathode temperature with space charge and attempt to obtain the longer transit times by a decrease in applied potential, the smaller currents obtained will decrease the negative resistance. Hence, with space charge it is advisable to employ the smallest value of n possible in actual circuit design.

For computation work the general formula (2) may be greatly simplified because we need to know only the maximum values which

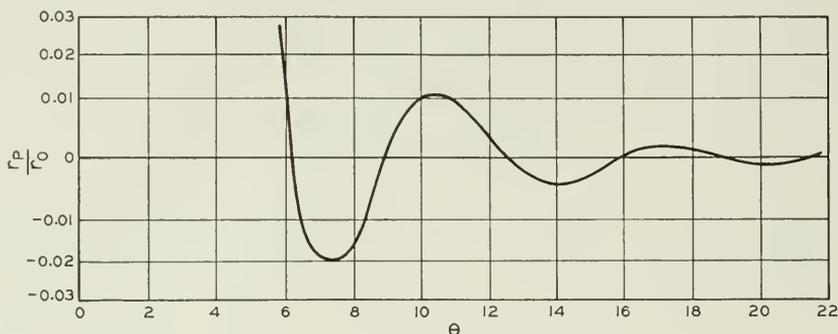


Fig. 2—Relation between transit angle θ and diode resistance.

the negative resistance attains. These occur in the neighborhood of transit angles given by

$$\theta = 2\pi n + \frac{\pi}{2}, \quad n = 1, 2, 3, \dots \quad (7)$$

and under these conditions the effective negative resistance is

$$R_p = \frac{r_p}{\text{area}} = -\frac{1.4 \lambda^4 I_0}{10^3 \pi r_2^2} \left(\frac{4n + 1}{5} \right) \text{ ohms.} \quad (8)$$

The detailed steps in circuit design are these: First the allowable value of current density I_0 must be determined. This depends upon the ability of the cathode to emit electrons and as a practical limit something in the neighborhood of 300 mils per square centimeter cannot be exceeded. When this current has been decided upon then the value x of the separation between cathode and anode may be found from (3) and (7) with the lowest value of n which will give practical figures. The space charge condition (4) also gives the lowest allowable potential for which the required current can flow and hence

the best efficiency in the simple diode of Fig. 1. From these values the negative resistance may be computed from (8).

As a next step the remainder of the circuit must be proportioned. The tuning relation (1) yields the values of height h for a given diameter. The next consideration is to insure that the sum of all positive resistances is less than the negative resistance of the diode. For a circuit with dimensions small compared with the wave-length, approximate formulas for the resistances associated with the losses in the circuit conductors can be readily derived from classical circuit analysis.

A most important resistance, not so readily computed, is caused by radiation of energy through the gap between the insulating flanges which separate cathode and anode. In most uses of the device, this radiated energy constitutes the useful load on the oscillator but care must be taken that the load is not so heavy as to stop the oscillations altogether. An important distinction must be made as to whether the tube is to radiate into free space or into some enclosure such as a hollow wave guide, for example. In the latter case the radiation may be regulated to a large extent by the geometry of the enclosure. For values of radiation resistance when energy is directed into free space an article by S. A. Schelkunoff⁵ may be referred to.

For oscillation, as pointed out, the sum of all these positive resistances must be less than the negative resistance of the electron discharge and for high efficiency the radiation resistance should be much greater than the sum of all of the other positive resistances. This is usually found to be the case, and in fact the radiation resistance is likely to be so great as to stop oscillations unless the gap is made sufficiently small.

In designing a hollow wave guide mounting for diodes of the sort pictured in Fig. 1 it was recognized that since the high-frequency wave energy issues from the coaxial resonator as a wave guided along the heater leads, the natural and probably most effective thing to do was to dispose these leads so that the field associated with them would conform as nearly as possible to one of the wave types which can be supported in a hollow wave guide. Of these wave types, the so-called H_1 type⁶ is readily generated by high-frequency current in a wire extending across a diameter of the guide, and the wave guide mounting shown in cross section on Fig. 3 is such as to give rise to this type of wave. For mechanical reasons a brass pipe of circular cross sections was chosen for the guide, and its diameter ($3\frac{7}{8}$ inches) was chosen large enough so that it would freely transmit an H_1 wave of the expected frequency. In the mounting, the high frequency circuit is completed from the anode to the wall of the guide through a stopping condenser.

Preliminary experiments with diode no. 24 had shown that when wave power issuing from the tube was allowed to radiate into free space, a space current of 500 milliamperes with anode voltage of about 300 volts was required to maintain oscillations. An interesting and instructive experiment is then to determine by how much this current

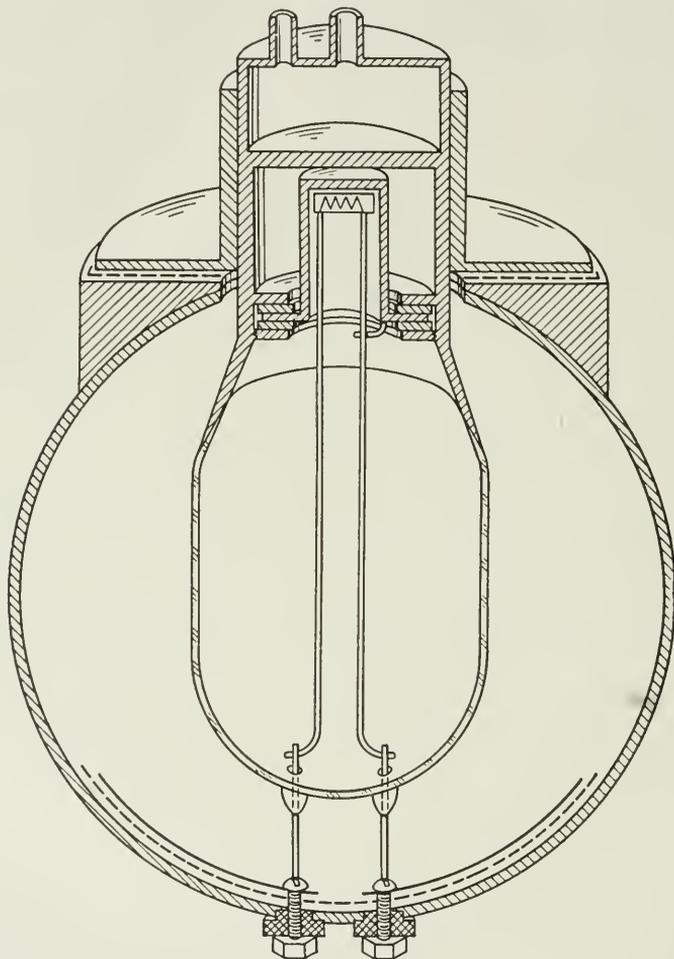


Fig. 3—Hollow wave guide mounting for diodes.

is reduced when radiation is held to as low a value as possible. For this purpose an assemblage as in Fig. 4 was used. Here a wave guide mounting of form comparable to Fig. 3 is clamped into sections of wave guide closed at the two ends by closely fitting but longitudinally ad-

justable reflecting pistons. By thus completely enclosing the tube, escape of energy into free space is avoided, and the losses external to the tube are reduced to the ohmic losses (including dielectric losses) incident to the existence of the wave within the guide. The presence of wave power within the guide is indicated by a crystal detector-microammeter combination connected to an antenna extending a short distance within the guide, as shown in Fig. 4. Adjustment of the pistons closing the

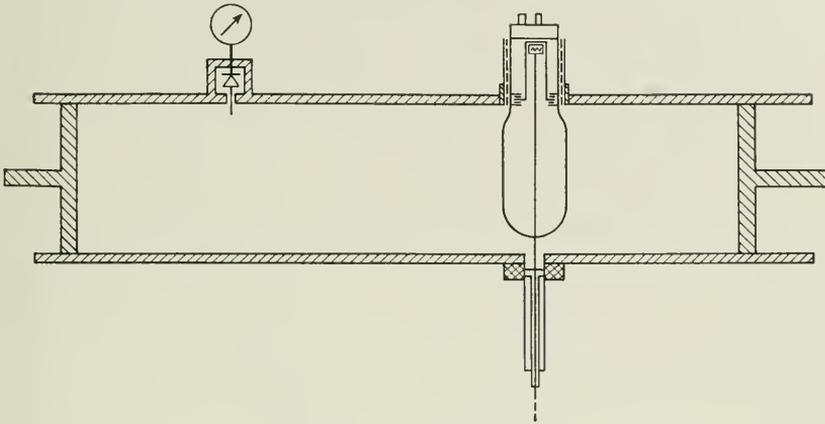


Fig. 4—Apparatus arrangement for "No Load" test of diode.

ends of the wave guide system allowed the attainment, for each value of anode current and voltage, of the most favorable impedance conditions for abstraction of energy from the diode.

The results of this experiment are shown on Fig. 5, which gives the boundaries of the domain of oscillation of the two tubes whose dimensions are given on Fig. 1. The large gain in extent of the oscillation region of tube no. 24 is immediately apparent; the free space oscillation limit of $E_p = 300$ volts, $I_p = 500$ ma. has been lowered to $E_p = 210$ volts, $I_p = 110$ ma. For tube no. 37 oscillations occur at much lower voltages, as is to be expected from the smaller anode-cathode distance, and the minimum plate current required to maintain oscillations is also somewhat smaller.

In the arrangement of Fig. 4 no useful power is extracted from the diode. To examine the oscillation domain of the diodes when delivering useful power, the arrangement of Fig. 6 was used. Here in a section of wave guide closed at both ends by tightly fitting but longitudinally adjustable pistons there are placed the diode mounting shown on Fig. 3 and a power absorbing and measuring element. The assemblage constitutes in effect a wave guide transformer, for by

suitably adjusting the positions of the pistons with respect to the source and the power absorber, and by a proper choice of the distance between the source and the power absorber, the impedance of the latter can be matched to that of the source, so as to ensure the maximum delivery of power.

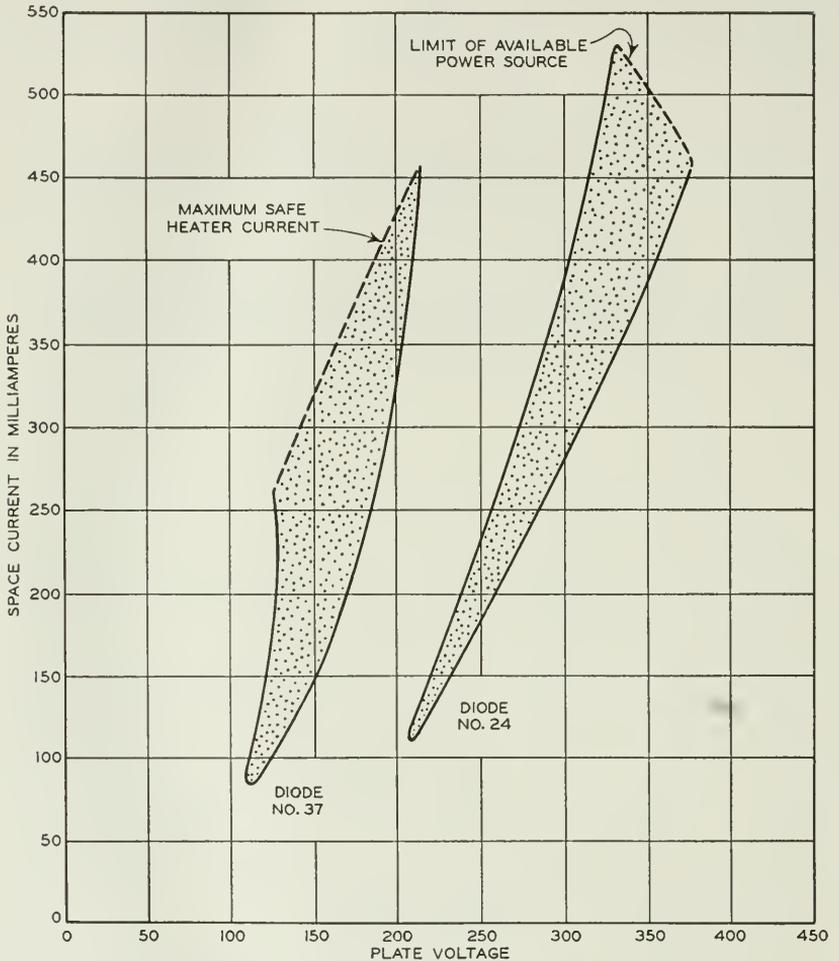


Fig. 5—"No Load" oscillation domain of diodes no. 24 and no. 37.

The power absorbing and measuring element shown on Fig. 6 represents a new and useful way of measuring power at the very high frequencies involved in this investigation. It makes use of the high negative temperature coefficient of resistance of boron. In the middle

of a wire extending across a diameter of the wave guide, parallel to the lines of electric force in an H_1 wave, there is placed a small crystal of boron. Connection to the crystal is made by fine platinum wires, melted into two small globules on opposite sides of the crystal.* By virtue of its small size and the fine leads connected to it, small amounts of power dissipated in the resistance of the crystal will raise its temperature materially, with a consequent large change in its resistance. With a stopping condenser, an ohmmeter connected as shown in Fig. 6 serves to indicate the resistance of the crystal when absorbing high-frequency power, and calibration curves showing resistance as a function of power absorption can be obtained with direct current.

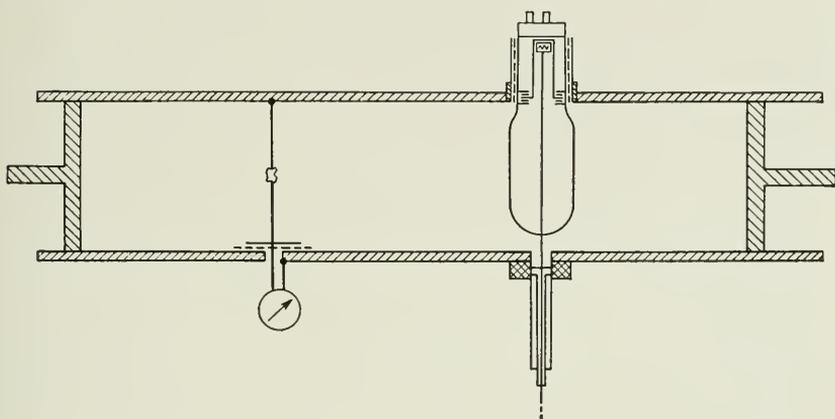


Fig. 6—Apparatus arrangement for "Loaded" test of diode no. 24.

Power output and efficiency data obtained during these measurements are shown on Fig. 7. Power outputs of a few tenths of a watt at efficiencies ranging from one to two tenths of a per cent are obtainable.

In consideration of the wave-length of the oscillations generated by these two diodes, it will be recalled that they were designed nominally for a wave-length of about 10 centimeters. For diode No. 24 the wave-length was close to 10.6 cm. (2830 mc.) and for diode no. 37 it was somewhat higher, about 11.55 cm. (2600 mc.). This difference is of the order to be expected from the difference in the dimensions of the two tubes.

While the wave-length should be fixed largely by the dimensions of the coaxial resonant circuit built into the diode, it is to be expected that it will be affected to a small extent by the applied voltage and by

* These were developed by Mr. G. L. Pearson of the Bell Telephone Laboratories.

the position of the piston closing one or both of the ends of the wave guide. In the case of diode No. 37 the wave-length was found to vary over a range between 11.50 and 11.65 cm. with plate voltage and over a range between 11.52 and 11.56 cm. with piston position.

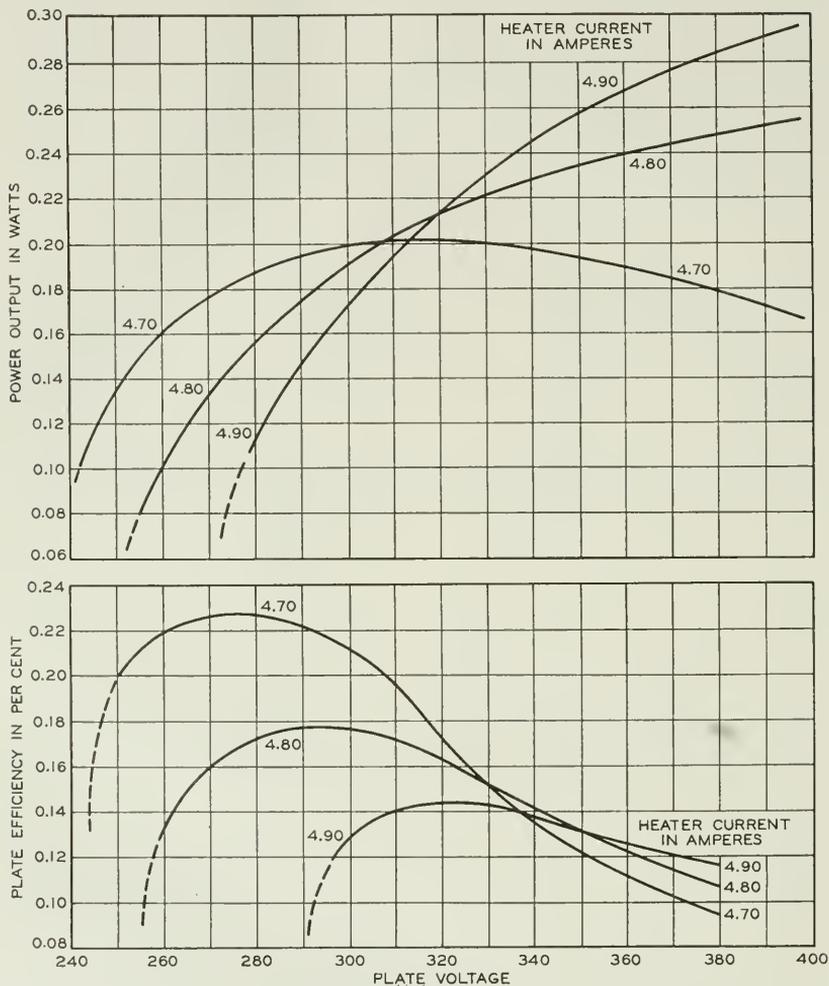


Fig. 7—Power output and plate efficiency for diode no. 24.

In conclusion, the writers wish to mention the work done by Mr. C. A. Bieling of the Bell Telephone Laboratories in working out suitable mechanical design features and in the actual assembly and processing of tubes which were built and tested.

REFERENCES

1. W. E. Benham, "Theory of the Internal Action of Thermionic Systems at Moderately High Frequencies," *Phil. Mag.*, March 1928 and *Phil. Mag. Supplement*, Vol. 11, February 1931.
2. Müller, "Experimentelle Untersuchungen über Elektronen-Schwingungen," *Hoch. u. Elek.*, Vol. 43, No. 6, June 1934.
3. W. W. Hansen, "On the Resonant Frequency of Closed Concentric Lines," *Jour. App. Phys.*, Vol. 10, No. 1, January 1939.
4. F. B. Llewellyn, "Operation of Ultra-High-Frequency Vacuum Tubes," *Bell System Technical Journal*, Vol. 14, No. 4, October 1935.
5. S. A. Schelkunoff, "Some Equivalence Theorems of Electromagnetics and Their Application to Radiation Problems," *Bell System Technical Journal*, Vol. 15, No. 1, January 1936.
6. G. C. Southworth, "Hyper-Frequency Wave Guides—General Considerations and Experimental Results," *Bell System Technical Journal*, Vol. 15, No. 2, April 1936.

A Representation of the Sunspot Cycle *

By C. N. ANDERSON

ALTHOUGH sunspots had been observed occasionally back to ancient times, their study may be said to date from their rediscovery by Galileo in the spring of 1610 with the then newly invented telescope. Since then much has been written about their nature, their periodicities and possible influence on human affairs.

The purpose of the study reported on in this paper was to analyze the components of the sunspot data and thereby to reconstruct a curve which would not only represent the variation in sunspot numbers from 1749 to date but would also be consistent with times of maxima and minima from 1610 to 1749. A number of attempts along this line have been made in the past,^{3, 4, 5, 6, 9, 11} all of which have neglected the data previous to 1749 and all have used a slightly different method of analysis. It is believed that the agreement in the present study is somewhat better than in those of the past; nevertheless, no claim is made for any great accuracy in predicting future sunspot activity. Harmonic analysis based on a fraction of a period is always a source of danger and, furthermore, we have no assurance that all the components of the sunspot curve are periodic functions.⁸ In fact, some papers have appeared in which each cycle was treated as a more or less independent outburst.^{2, 10, 12, 13} Nevertheless, because of the long base line over which agreement is obtained in this present study, it is hoped that the results may not be too much in error for at least a few cycles.

The data used are the series of relative sunspot numbers begun by Rudolph Wolf,¹ Professor of Astronomy at Zurich and continued by his successors Wolfer and Brunner. Wolf began his systematic observations of sunspot numbers in 1849. He endeavored to make some allowance for the area of the spots and to avoid having a small spot of short duration count as much as a large group. With this in mind, he applied the following formula to his observation:

$$\text{Relative sunspot number} = k(10g + f),$$

where g and f are the group and total spot numbers respectively,

* Presented before the Astronomy Section of A.A.A.S. at Richmond, Va., December 28, 1938.

and k is a constant depending on the type of telescope and other factors affecting the observation. The figure 10 is an arbitrary one arrived at by Wolf from investigation of a number of individual cases and which seemed to give him the proper relationship.

A careful study was then made by Wolf of all existing records of prior data. Hofrat Schwabe of Dessau supplied data for the period 1826 to 1855 to which a correction was applied determined from a study of the overlapping data and from the percentage of spotless days. Johann Casper Staudacher of Nürnberg had made a total of 1131 observations (from one to ten every month) by means of a helioscope during the period 1749 to 1799. He often gave detailed descriptions and included many sketches. Imagine a man making observations for fifty years without any attempt at analysis and then have the data resurrected fifty years after his death to form an important contribution to the record. Flaugergues (1794–1830), Tevel (1816–1836), Adams (1819–1823) and Arago (1822–1830) supplied most of the data for the intervening period between the observations of Staudacher and Schwabe. Wolf lists about a hundred references (225 up to 1866) to sunspots prior to 1850. In most cases, the observations were incidental to other solar observations such as culminations, solar diameter, eclipses, transits, or on the nature of sunspots rather than their number. Each record was carefully studied, and from them all Wolf obtained a representation of sunspot numbers for as far back as 1749 and established the times of maxima and minima with an accuracy of ± 2 years or better back as far as 1610 A.D. Although the data prior to 1849 include a certain element of unreliability and all the data represent *relative* numbers which have been obtained from the observations by applying a weighting factor, it is, however, not only the best record but the only one for such a long period of time. In the aggregate it is probably a good indication of the variation of solar activity.

The method employed in the present analysis is briefly as follows:

(a) The yearly averages of sunspot numbers from 1749 to the end of 1937 were first plotted in the conventional way as shown in Fig. 1; it was noted that in certain sections of the curve, notably after 1840, the maximum amplitudes of alternate eleven-year periods were higher than the intervening ones.

(b) The data were redrawn with alternate eleven-year periods above and below the time axis; this not only smoothed out the envelope of the maxima but also simplified the analysis by eliminating a computed mean value base line which has been employed in previous analyses; the maximum-amplitude component becomes approximately

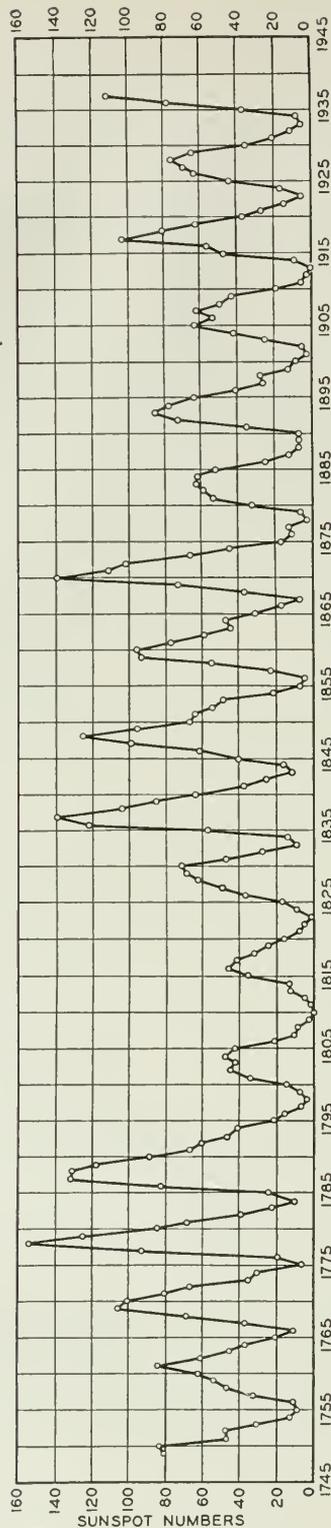


Fig. 1—Yearly averages of monthly sunspot numbers. This series of relative sunspot numbers was begun by Wolf of the Zurich Observatory and continued by his successors, Wolfer and Brunner.

twenty-two years instead of eleven years and the physical justification is the similarity in the polarity of the leading spots in alternate eleven-year periods.

(c) Quarterly values were obtained from the curve in order to obtain a smaller unit and to approximate better such periods as 22.25 years, 17.33 years and other periods not an integral number of years; the values were assigned + or - signs in accordance with whether they were above or below the time axis.

(d) A periodogram was computed and the component of maximum amplitude appeared to be 22.75 years; this component was eliminated, a new periodogram computed and so on; after many trials it was finally decided that no solution could be found which would reduce the residue satisfactorily with 22.75 years as the main component.

(e) Inspection of the analysis indicated an improvement would be obtained by a decrease in the period and accordingly a change was made to 22.5 years and the computations repeated; a solution was finally obtained which fit the 1749-1937 curve of sunspots fairly well but, when extrapolated back to 1600 A.D., did not fit particularly well the observed times of maxima and minima for the period 1610 to 1749; a still further reduction in the chief component was necessary.

(f) The computations were repeated with 22.25 years as the chief component; after the computation of the series was completed, it was discovered that the components were either harmonics or nearly integral harmonics of 312 years; many of these components had been in use since the original periodogram.

(g) A search was made for the 312-year period, since it should be possible to check in the cases of two minima and two maxima, with the following results:

Minima	1610.8 ± 0.4	1923.1	Diff. 312.3 ± 0.4
Maxima	1615.5 ± 1.5	1928.6	313.1 ± 1.5
Minima	1619 ± 2	1933.6	314.6 ± 2
Maxima	1626 ± 0.5	1937.7	311.7 ± 0.5

giving a weighted average of 312.5 years or an average of 312.0 years using the two most reliable values.

(h) Computations were again repeated using harmonic components of 312 years. This gave a very good representation of the data from 1749 to date and also agreed with the times of maxima and minima for the period 1610 to 1749. The resultant curve is shown in Figs. 2 and 3.

The components of the reconstructed curve are shown in Fig. 4. It is of interest to note that the 22.25-year component which is largely

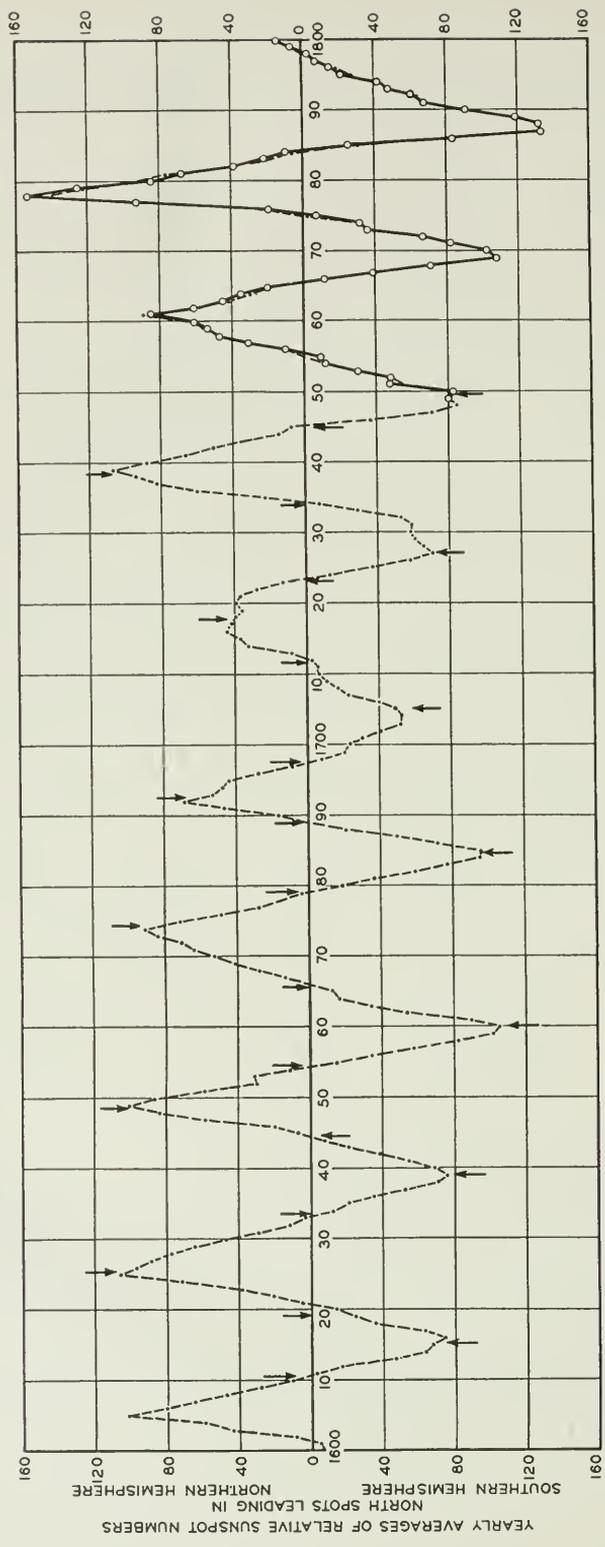


Fig. 2—Measured and computed sunspot numbers, 1600–1800 A.D. Solid line indicates measured values; light dashed line indicates values obtained by adding up the components computed from the measured values; components are harmonics of a 312-year period; arrows indicate the years when maxima and minima were observed.

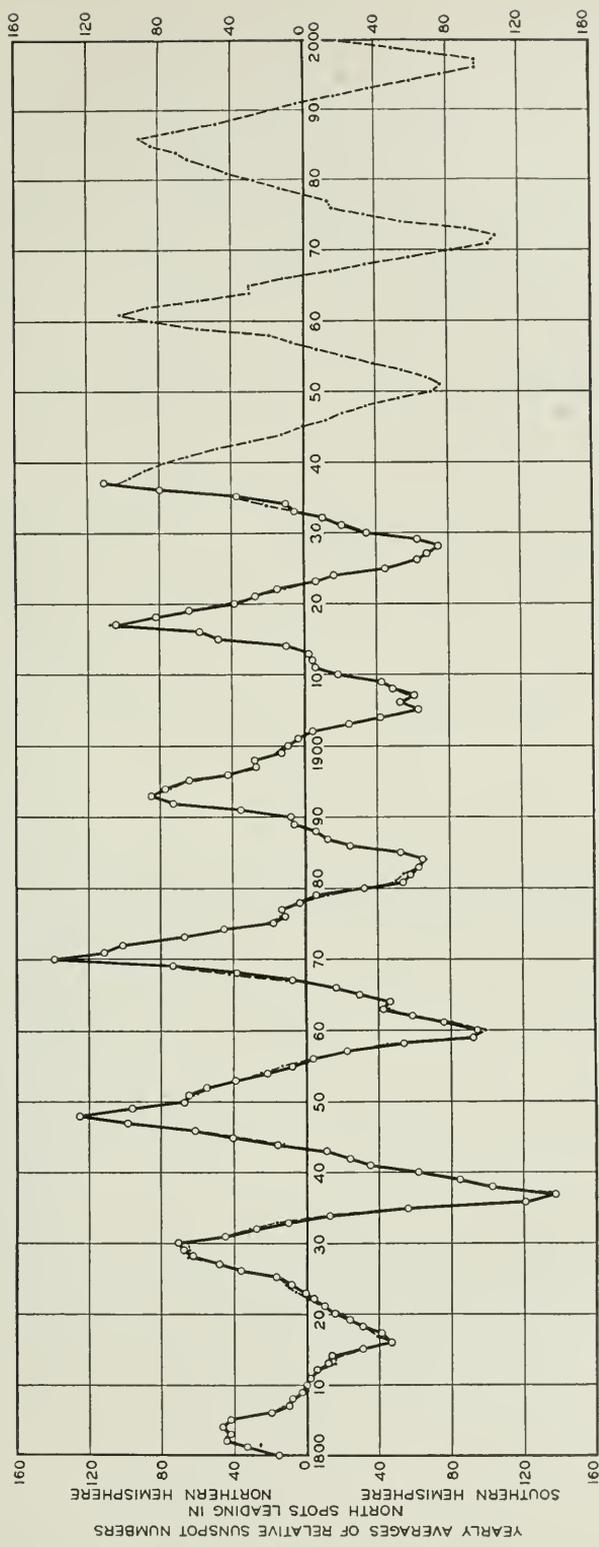


Fig. 3—Measured and computed sunspot numbers, 1800-2000 A.D. Solid line indicates measured values; light dashed line indicates values obtained by adding up the components computed from the measured values; components are harmonics of a 312-year period.



Fig. 4—Analysis of components of the sunspot cycle. Vertical lines indicate amplitudes of harmonic components of an apparent 312-year cycle.

responsible for the eleven-year periodicity of sunspots has an amplitude of only about $2/5$ of the greatest amplitudes of the resultant maxima. Next in importance are the two adjacent periods at 17.3 and 18.4 years, respectively.

In conclusion, the study has resulted in a representation of yearly sunspot averages which agrees as well as could be desired with the data and which is also consistent with the reported times of maxima and minima back as far as 1610 A.D. The chief component has been treated as a 22-year (22.25 years) component instead of eleven years. In the course of computation the components appeared to be harmonics, or nearly so, of a 312-year period. A substantiation of this 312-year cycle was found in a check of the overlapping data (maxima and minima from 1923 to date).

It had been hoped that the resultant distribution of amplitudes versus frequency of the components might be capable of simple interpretation and that a rather simple explanation of the phenomena of sunspot periodicity might result, such as one or two forces acting upon a nonlinear system with a given fundamental period. Further studies may still indicate this to be the case.

The author wishes to express his appreciation of the assistance of Miss Helen Grant with the rather tedious computations.

REFERENCES

1. "Mitteilungen über die Sonnenflecken," Rudolph Wolf, *Naturforschende Gesellschaft in Zurich Vierteljahrsschrift*, 1856-1865.
2. "Über Eine Neue Theorie zur Erklärung der Periodicität der Solaren Erscheinungen," J. Halm, *Astronomisches Nachrichten*, Vol. 156, 44, 1901.
3. "On the Periodicities of Sunspots," A. Schuster, *Phil. Trans. Royal Soc.*, London, Vol. 85, 1911.
4. "Determination of Periodicities by Harmonic Analyzer with Application of the Sunspot Cycle," A. A. Michelson, *Astrophysical Journal*, Vol. 38, 1913.
5. "On the Harmonic Analysis of Sunspot Relative Numbers," Hisashi Kimura, *Monthly Notices of R.A.S.*, Vol. 73, 1913.
6. "On the Harmonic Analysis of Wolf's Sunspot Numbers," H. H. Turner, *Monthly Notices R.A.S.*, Vol. 73, 549, 1913.
7. "On the Expression of Sunspot Periodicity as a Fourier Sequence," H. H. Turner, *Monthly Notices R.A.S.*, Vol. 73, 714, 1913.
8. "On a Method of Investigating Periodicities in Disturbed Series with Special Reference to Wolfer's Sunspot Numbers," G. Udny Yule, *Phil. Trans. Royal Soc.*, London, Ser. A226, 1926-27.
9. "A New Analysis of the Sunspot Numbers," Dinsmore Alter, *Monthly Weather Review*, October 1928.
10. "Untersuchungen über die Häufigkeitskurve der Sonnenflecke," H. Ludendorff, *Zeitschrift für Astrophysik*, Vol. 2, May 1931.
11. "Über Perioden der Sonnenflecken," S. Oppenheim, *Astronomisches Nachrichten*, Vol. 232, 369, 1932.
12. "Forecasting Sunspots and Radio Transmission Conditions," A. L. Durkee, *Bell Laboratories Record*, Vol. XVI, December 1937.
13. "The Mathematical Characteristics of Sunspot Variations," John Q. Stewart and H. A. A. Panofsky, *Astrophysical Journal*, Vol. 88, November 1938.

The Number of Impedances of an n Terminal Network

By JOHN RIORDAN

This paper gives the enumeration of impedances measurable at the n terminals of a linear passive network. The enumeration supplies background for the study of network representations and the numerical results which are given up to ten terminals are perhaps surprising in the rapidity of the rise of the number of impedances with the number of terminals; almost 126,000,000 impedances, e.g., are measurable for ten terminals.

A LINEAR passive network having n accessible terminals may be completely represented by an equivalent direct impedance network,¹ consisting of branches, devoid of mutual impedance, connecting the terminals in pairs. The number of elements (branches) in this representation is equal to the number of combinations of n things taken two at a time, i.e., $\frac{1}{2}n(n-1)$. Each of the elements is defined by an impedance measured by energizing between one of the terminals it connects and the remaining terminals connected together and taking the ratio of the driving voltage to the current into the other terminal it connects. The network then is represented by a particular set, of $\frac{1}{2}n(n-1)$ members, of impedances measurable at its terminals; as will appear later, the set is of short-circuit transfer impedances.

The direct impedance network is one among many network representations; it is taken as illustrative of two aspects, (i) the necessity of a certain number of elements $\frac{1}{2}n(n-1)$ and (ii) the expression of these elements in terms of measurable impedances. It is well known that any linearly independent set, of $\frac{1}{2}n(n-1)$ members, of the measurable impedances of an n -terminal network will serve as a network representation; hence the enumeration of representations may be taken in two steps, the first of which, the enumeration of measurable impedances, is dealt with in the present paper.

The number of measurable impedances for two to ten terminal linear passive networks is given in Table I, which lists the driving-point impedances, D_n , transfer impedances (open or short circuit), T_n , certain additional transfer impedances to be described later, U_n , and the total N_n . As mentioned below, this total counts once only

¹ Item (b) in the list of equivalent networks given by G. A. Campbell "Cisoidal Oscillations," *Trans. A.I.E.E.* 30, pp. 873-909 (1911), p. 889; or p. 81, "Collected Papers of George Ashley Campbell," Amer. Tel. & Tel. Co., New York, 1937.

impedances which are equal by the reciprocity theorem; the doubling of T_n in forming the total is due to the equality in number of open-circuit and short-circuit transfer impedances. The numbers increase rapidly with n , reaching almost 126,000,000 for ten terminals. The number of representations, which is the number of combinations of the measurable impedances $\frac{1}{2}n(n - 1)$ at a time less the number of non-independent sets, at a guess increases even more rapidly, indicating a variety of equivalents, few of which seem to have been investigated.

TABLE I
MEASURABLE IMPEDANCES OF AN n -TERMINAL NETWORK

n	D_n	T_n	U_n	$N_n = D_n + 2T_n + U_n$
2	1	0	0	1
3	6	3	0	12
4	31	33	60	157
5	160	270	1,050	1,750
6	856	2,025	12,540	17,446
7	4,802	14,868	129,570	164,108
8	28,337	109,851	1,257,060	1,505,099
9	175,896	827,508	11,889,990	13,720,902
10	1,146,931	6,397,665	111,840,180	125,782,441

Because the field of the work is somewhat unusual, considerable space is given to details in the formulation of the problem before proceeding to the enumeration proper. The enumerating expressions obtained are found susceptible of some mathematical development which, though subsidiary to the main object of the paper, seems of sufficient interest to justify the relatively brief exposition given. The arrangement is such that readers not interested in this mathematical half may obtain the substance of the paper without it.

FORMULATION OF THE PROBLEM

The enumerating problem is essentially one of combinations, as indicated schematically in Fig. 1, which shows the n terminals of a linear passive network together with the apparatus required for impedance measurement, that is, a source, a voltmeter and an ammeter, each supplied with two terminals (shown solid to distinguish them from the network terminals). Each of these latter may be connected across any pair of the n terminals except that the ammeter, which constitutes a short circuit, may not be connected to terminals to which either the source or voltmeter is connected; in the former case no current will be supplied to the network and in the latter the voltmeter will read zero. The ammeter may be connected in series with the source to read the source current, of course.

Although but one source, voltmeter, and ammeter are shown, as many of each as will produce distinct impedances should of course be included. Multiple sources are not required because if the source voltages are in defined proportions, as is necessary to determine impedances independent of source voltage, the corresponding measurable admittances are linear combinations of single-source admittances, by the principle of superposition; a similar requirement on source currents produces impedances which are linear combinations of single-source impedances. A single voltmeter is sufficient because it has no effect on network currents or voltages and it is immaterial whether

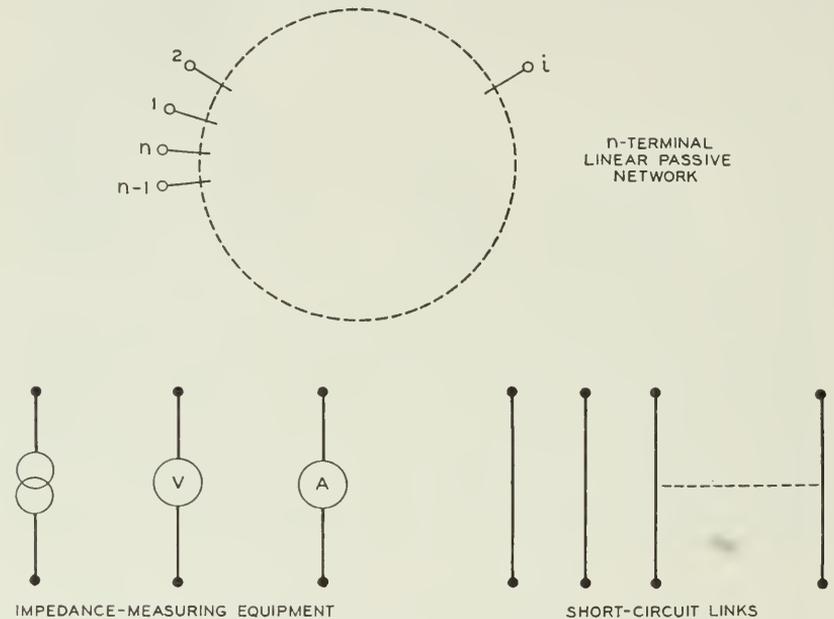


Fig. 1.—Elements involved in impedance enumeration.

impedances are supposed measured by successive positions of a single voltmeter or by many voltmeters. The connection of an ammeter is equivalent to a short circuit (except of course when in series with a source) across the terminals the ammeter connects; this alters network voltages and currents and the impedances measured without the ammeter differ from those with it. Hence a plurality of ammeters or its equivalent is required; for convenience, all ammeters except that one determining a specific impedance under consideration are supposed replaced by the short-circuiting links on the right of Fig. 1, thus focussing attention on the single items of the enumeration.

The classification under which the enumeration is conducted is illustrated by Fig. 2, which shows typical positions of source, voltmeter and ammeter for measuring impedances of three classes. In the first of these, the ammeter reads the source current, the voltmeter source voltage (across some pair of the network terminals) and the class is that of driving-point impedances, D_n . In the second class, that of transfer impedances T_n , there are two types of connection: in the first the ammeter reads the source current, the voltmeter a non-source voltage, the voltage-current ratios being open-circuit transfer impedances; in the second the voltmeter reads the source voltage and

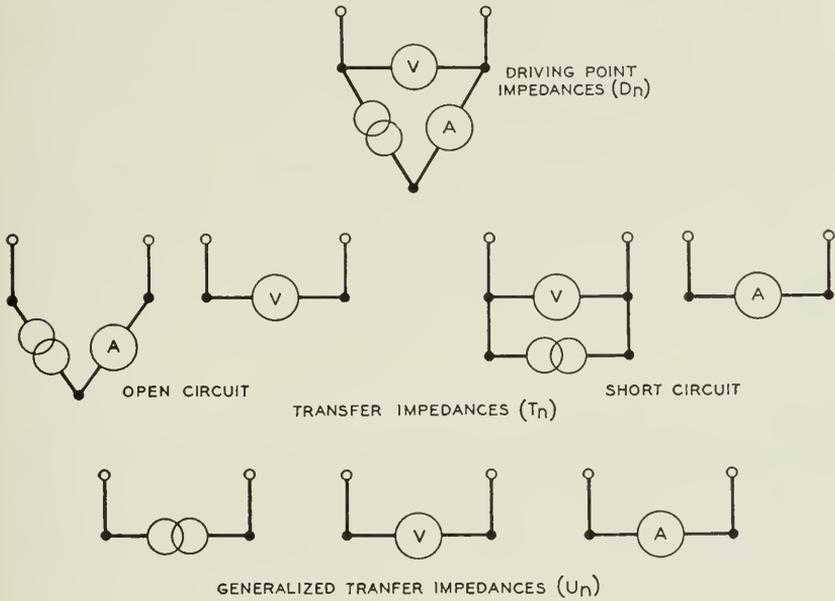


FIG. 2—Arrangement of apparatus for measuring impedances of three classes.

the ammeter a non-source current, the voltage-current ratios being short-circuit transfer impedances. It will be noted that the two connections differ only in that the ammeter and voltmeter are interchanged. The third class is that of generalized transfer impedances U_n , in which both voltmeter and ammeter are across non-source terminals.

The last class, of course, might be supposed to include the two preceding ones but the separation proves convenient not only for numerical work, as will appear, but also for keeping distinct both well-recognized and formally different classes.

The most important of these differences in classes is that arising from the reciprocity theorem. Of the three classes, only the T (open-circuit and short-circuit transfer impedances) includes members which are equal by the reciprocity theorem; this follows because the reciprocity theorem requires interchange of voltmeter (or ammeter) and source with associated ammeter (or voltmeter) and the T class alone permits this. It is a matter of taste whether such duplicates should be counted separately or as one; in the interest of keeping large figures as low as possible they are here counted as one, since the classification is such that the other alternative may be taken merely by doubling the T_n .

Another reason for keeping the T class distinct is that total open-circuit and short-circuit transfer impedances for a given number of terminals are equal in number. This is proved immediately by observing that the two connections shown in Fig. 2 for this class are in one-one correspondence: each may be obtained from the other by interchanging voltmeter and ammeter. Moreover, if $T_{x, n}^o$ and $T_{x, n}^s$ are the numbers of open-circuit and short-circuit transfer impedances measurable when short circuits have been placed across the n terminals in all possible ways so as to realize x terminals, each merged group of terminals counting as a single terminal, the correspondence leads to the relation

$$T_{x+1, n}^o = T_{x, n}^s, \quad (1)$$

since the interchange of voltmeter and ammeter in the measuring arrangement for open-circuit transfer impedances results in one less available terminal, two terminals being merged by the ammeter short circuit. Note that $T_{2, n}^o = T_{n, n}^s = 0$, since with just two terminals, no non-source voltages and with n terminals no non-source currents are measurable.

Equation (1) is important in determining enumerating expressions in the section following.

ENUMERATING EXPRESSIONS

The laws of enumeration appear most simply exposed by examining the simplest cases first.

For two terminals, there is but one measurable impedance, the driving-point impedance between the terminals.

For three terminals, with the terminals distinct, there are three driving-point and three open-circuit transfer impedances, for there are three ways of selecting driving pairs of the three terminals and for each selection two ways of selecting pairs for open-circuit voltage

measurement, the total of six transfer impedances being halved to eliminate reciprocity theorem duplicates. With two of the terminals connected by an ammeter, there are again three driving-point and three transfer impedances, the latter being short-circuit transfer impedances, for there are three ways of connecting pairs of terminals and one driving-point and two transfer impedances for each, the total of transfer impedances again being halved to eliminate duplicates.

There are no generalized transfer impedances because with an ammeter connected, there is only one measurable voltage, the driving-circuit voltage.

With terminals designated by t_1, t_2 and t_3 , the conditions arising from connection of terminals may be exhibited as follows:

$$\begin{array}{l} \text{Terminals distinct} \quad t_1|t_2|t_3 \\ \text{Pairs connected} \quad t_1t_2|t_3 \quad t_1t_3|t_2 \quad t_1|t_2t_3 \end{array}$$

the lines of separation dividing the terminals into groups such that the terminals in any group are merged into a single terminal. Paying attention only to the number of terminals in each group, the groups illustrated may be designated by the partition notation (111) or (1³) and (21), the numbers in the designation being partitions² of the number 3.

The enumeration for three terminals may then be exhibited as follows:

MEASURABLE IMPEDANCES

Group	Driving Point	Open-Circuit Transfer	Short-Circuit Transfer	Total
(1 ³)	3	3	0	6
(21)	3	0	3	6
	6	3	3	12

It will be noted that the open-circuit and short-circuit transfer impedances satisfy equation (1), that is, $T^o_{3,3} = T^s_{2,3}$.

This table and its correspondents for larger values of n show that the impedances may be expressed as sums with respect to x , where x is the number of terminals defined as in equation (1), from 2 to n ; thus e.g., $D_n = \sum D_{x,n}$ where $D_{x,n}$ is the number of driving-point impedances measurable for all conditions of merging of n terminals such that the resulting number of terminals is x . Moreover, con-

²A partition of a number n is any collection of positive integers whose sum is equal to n . It may be noted that the number of parts of a partition is the number x of equation (1); the partition (1³) has three parts corresponding to the three distinct terminals; (21) has two parts corresponding to two terminals, each merged pair of terminals counting singly.

sidering for the moment only the driving-point and open-circuit transfer impedances, the numbers $D_{x, n}$ and $T_{x, n}^o$ are the products of two factors: (i) the number of such impedances measurable for x terminals, which is independent of n and (ii) the number of ways the n terminals may be merged so as to result in x terminals, which is independent of the impedance classes. By equation (1) this result applies also to $T_{x, n}^s$ and, as $U_{x, n}$ is related to $T_{x, n}^s$ by a factor independent of n , as will be shown, it applies generally.

This leads to the following equation:

$$\left\{ \begin{array}{c} D_n \\ T_n \\ U_n \end{array} \right\} = \sum_{x=2}^n \left\{ \begin{array}{c} d_x \\ t_x \\ u_x \end{array} \right\} S_{x, n}. \quad (2)$$

The small letters are the several factors of the first kind and $S_{x, n}$ is the common second factor.

The small letters are determined as follows: A driving point impedance may be measured between every pair of terminals; hence d_x is the number of combinations of x things taken two at a time, that is:

$$d_x = \binom{x}{2} = \frac{1}{2}x(x-1) = \frac{1}{2}(x)_2, \quad (3)$$

where $(x)_i$ is the factorial symbol $x(x-1) \cdots (x-i+1)$.

For a given pair of driving terminals, there are $\binom{x}{2} - 1$ measurable open-circuit transfer impedances since a voltmeter can be connected to every pair of the x terminals except the driving pair; hence, multiplying by the number of driving terminals and by the factor one-half to eliminate reciprocity theorem duplicates:

$$\begin{aligned} t_x &= \frac{1}{2} \binom{x}{2} \left[\binom{x}{2} - 1 \right], \\ &= \frac{1}{8} [4(x)_3 + (x)_4]. \end{aligned} \quad (4)$$

The second, factorial, form is given for convenience of later development.

By equation (1) this serves for enumeration of both open-circuit and short-circuit transfer impedances; the direct enumeration of the latter appears more difficult.

Considering, for the generalized transfer impedances, a fixed source and an ammeter in a fixed (non-source) position, the voltmeter may be connected across $\binom{x}{2}$ pairs of terminals when x terminals are

available; one of these pairs is the source pair measuring a short-circuit transfer impedance which must be excluded; hence, remembering that reciprocity theorem duplicates are eliminated in the latter:

$$\begin{aligned}
 U_{x, n} &= 2 \left[\binom{x}{2} - 1 \right] T_{x, n}^s \\
 &= 2 \left[\binom{x}{2} - 1 \right] T_{x+1, n}^o = 2 \left[\binom{x}{2} - 1 \right] t_{x+1} S_{x+1, n}.
 \end{aligned}$$

Degrading x by unity to obtain the form of equations (2), the third of the lower case factors is reached as follows:

$$\begin{aligned}
 u_x &= \binom{x}{2} \left[\binom{x}{2} - 1 \right] \left[\binom{x-1}{2} - 1 \right] \\
 &= \frac{1}{8} [20(x)_4 + 10(x)_5 + (x)_6].
 \end{aligned} \tag{5}$$

The common factor $S_{x, n}$ remains for determination.

Returning to the connection conditions illustrated for three terminals, this number is the number of ways separators may be placed between letters of the collection $t_1, t_2 \dots t_n$ symbolizing the terminals so as to produce x compartments, symbolizing merged terminals. The terminal symbols $t_1 \dots t_n$ may be thought of as the prime distinct factors (excluding unity) of some number and the number $S_{x, n}$ is then identically the number of ways a number having n distinct prime factors may be expressed as a product of x factors. The enumeration for this latter problem is given by Netto,³ who gives the recurrence relation

$$S_{x, n+1} = xS_{x, n} + S_{x-1, n}$$

with

$$S_{n, n} = 1, \quad S_{x, n} = 0, \quad x > n, \quad S_{0, n} = 0, \quad n \neq 0.$$

This is the recurrence relation for the Stirling numbers of the second kind,⁴ the notation for which has been adopted in anticipation of the result. These numbers are perhaps better known as the "divided differences of nothing," that is, as defined by the equation:

$$S_{x, n} = \lim_{z \rightarrow 0} \frac{1}{x!} \Delta^x z^n = \frac{1}{x!} \Delta^x 0^n,$$

where Δ^x denotes x iterations of the difference operator with unit

³ "Lehrbuch der Combinatorik," Leipzig, 1901, pp. 169-170; Whitworth, "Choice and Chance," Cambridge, 1901, Prop. XXIII, p. 88, gives a generating function for the solution of this problem which, it is not difficult to show, leads to the same answer.

⁴ Ch. Jordan, "Statistique Mathematique," Paris, 1927, p. 14.

interval, that is, of the operator defined by

$$\Delta f(z) = f(z + 1) - f(z).$$

For convenience of reference, a short table of the numbers follows:

$x \backslash n$	$S_{x, n}$					
	0	1	2	3	4	5
0	1					
1	0	1				
2	0	1	1			
3	0	1	3	1		
4	0	1	7	6	1	
5	0	1	15	25	10	1

The table may be verified and extended readily by the recurrence relation.

With this table (extended to $n = 10$) and corresponding tables of d_x , t_x and u_x running to $x = 10$, the values given in Table I may be calculated by equations (2) and in this sense this paper is completed at this point. The sections below contain an algebraic and arithmetical examination of the numbers.

GENERATING IDENTITIES

The generating identity for the function

$$\sum_{x=0}^n a^x S_{x, n}$$

is ⁵

$$\exp [a(e^t - 1)] = \sum_{n=0}^{\infty} \frac{t^n}{n!} \sum_{x=0}^n a^x S_{x, n}.$$

This leads, by differentiating s times with respect to a and setting a equal to unity, to the generating identity:

$$(e^t - 1)^s \exp (e^t - 1) = \sum_{n=0}^{\infty} \frac{t^n}{n!} \sum_{x=0}^n (x)_s S_{x, n}.$$

This relation may be rendered more summarily by introducing the notation of the symbolic or umbral calculus ⁶ of Blissard; the expression on the right is written $\exp t\delta$ where δ is an umbral symbol standing for the sequence $(\delta_0, \delta_1, \dots, \delta_n, \dots)$ in this case infinite, through the relation $\delta^n = \delta_n$ and:

⁵ E. T. Bell, "Exponential Polynomials," *Annals of Math.* 35, 2 (April, 1934) p. 265; or J. Riordan, "Moment Recurrence Relations . . .," *Annals of Math. Statistics* 8, 2, pp. 103-111 (June, 1937), eq. 3.4.

⁶ Cf. Bell, l.c. p. 260 where further references are given.

$$\delta_n = \sum_{x=0}^n (x)_s S_{x, n}.$$

All algebraic operations on umbral symbols are carried out as in ordinary algebra except that the degrading of subscripts must not be performed until operations are completed. It must be noted that $\delta^0 = \delta_0$, hence is unity only when $\delta_0 = 1$, as in the present case and not always as in ordinary algebra.

The umbrae for the impedance numbers are written D , T and U , and by use of the generating identity above have the following generating identities:

$$\begin{aligned} \exp tD &= \frac{1}{2}(e^t - 1)^2 \exp(e^t - 1), \\ \exp tT &= \frac{1}{8}[4(e^t - 1)^3 + (e^t - 1)^4] \exp(e^t - 1), \\ \exp tU &= \frac{1}{8}[20(e^t - 1)^4 + 10(e^t - 1)^5 + (e^t - 1)^6] \exp(e^t - 1). \end{aligned} \tag{6}$$

These follow immediately from the base generating identity and the factorial expressions for d_x , t_x and u_x .

Expanding these expressions in powers of e^t gives alternate expressions as follows:

$$\begin{aligned} \exp tD &= \frac{1}{2}(e^{2t} - 2e^t + 1) \exp(e^t - 1), \\ \exp tT &= \frac{1}{8}(e^{4t} - 6e^{2t} + 8e^t - 3) \exp(e^t - 1), \\ \exp tU &= \frac{1}{8}(e^{6t} + 4e^{5t} - 15e^{4t} + 35e^{2t} - 36e^t + 11) \exp(e^t - 1). \end{aligned} \tag{6.1}$$

To recapitulate, these expressions mean that D_n , T_n and U_n are the coefficients of $t^n/n!$ in the expansions of the right-hand sides; taking D_n , for example, the first equation of (6) is equivalent to the equation:

$$D_n = \lim_{t \rightarrow 0} \frac{d^n}{dt^n} \left[\frac{1}{2}(e^t - 1)^2 \exp(e^t - 1) \right],$$

which may be shown to be equivalent to the first of equations (2).

The generating identities lead immediately to recurrence relations, as will now appear.

RECURRENCE RELATIONS

Recurrence relations to be derived are all obtained by differentiation with respect to t . Under this operation umbrae behave like ordinary variables; thus

$$\begin{aligned} \frac{d}{dt} \exp tD &= D \exp tD \\ &= D_1 + D_2 t + D_3 \frac{t^2}{2!} + \cdots + D_{n+1} \frac{t^n}{n!} + \cdots, \end{aligned}$$

as may be verified readily.

In the first type of recurrence only successive values of the numbers themselves appear. The derivation is illustrated for the D_n , the simplest case. Differentiating the first of equations (6) leads to the relation:

$$D \exp tD = \frac{1}{2}(e^{3t} - e^t) \exp (e^t - 1),$$

or

$$\begin{aligned} (e^t - 1)D \exp tD &= e^t(e^t + 1)\frac{1}{2}(e^t - 1)^2 \exp (e^t - 1) \\ &= (e^{2t} + e^t) \exp tD. \end{aligned}$$

Equating coefficients of $t^n/n!$ in this relation gives the umbral recurrence:

$$D(D + 1)^n - D_{n+1} = (D + 2)^n + (D + 1)^n,$$

which in ordinary form is:

$$(n - 2)D_n = \sum_{i=1}^n \left[\binom{n}{i} (2^i + 1) - \binom{n}{i+1} \right] D_{n-i}.$$

The process is common to the three classes of numbers and produces similar results which may be put in general form as follows:

$$a_n A_n = \sum_{i=1}^n \left[\binom{n}{i} b_i - \binom{n}{i+1} c_{i+1} \right] A_{n-i}, \quad (7)$$

where A_n , a_n , b_i and c_i are defined for the three cases by the following table:

A_n	a_n	b_i	c_i
D_n	$n - 2$	$2^i + 1$	1
T_n	$4n - 12$	$3^i + 6 \cdot 2^i + 5$	$2^i + 2$
U_{n-3}	$480 \left[\binom{n}{4} - \binom{n}{3} \right]$	$7^i + 10 \cdot 6^i + 5 \cdot 5^i - 60 \cdot 4^i + 35 \cdot 3^i + 34 \cdot 2^i - 25$	$6^i + 4 \cdot 5^i - 15 \cdot 4^i + 35 \cdot 2^i - 36$

Somewhat more convenient recurrences may be obtained by allowing the presence of numbers other than those for which the recurrence is sought. For this purpose it is expedient to introduce the exponential numbers ϵ_n of E. T. Bell.

These are defined by the generating identity:

$$\exp t\epsilon = \exp (e^t - 1)$$

or by the equivalent formula:

$$\epsilon_n = \lim_{t \rightarrow 0} \frac{d^n}{dt^n} \exp (e^t - 1) = \sum_{x=1}^n S_{x, n},$$

which shows their close relation with the impedance numbers. They have the recurrence relation:

$$\epsilon_{n+1} = (\epsilon + 1)^n$$

and

$$\epsilon_0 = \epsilon_1 = 1.$$

Now, returning to the first of equations (6) and again differentiating:

$$\begin{aligned} D \exp tD &= \frac{1}{2}[2(e^t - 1)e^t + (e^t - 1)^2e^t] \exp (e^t - 1), \\ &= (e^t - 1)e^t \exp (e^t - 1) + e^t \exp tD, \\ &= 2 \exp tD + (e^t - 1) \exp (e^t - 1) + \exp t(D + 1), \\ &= 2 \exp tD + \exp t(\epsilon + 1) - \exp t\epsilon + \exp t(D + 1), \end{aligned}$$

from which, passing to the coefficient relation, comes the umbral recurrence:

$$D_{n+1} = 2D_n + (D + 1)^n + \epsilon_{n+1} - \epsilon_n.$$

Similar recurrences for the T and U numbers are derived in the same way; writing $\Delta\epsilon_n = \epsilon_{n+1} - \epsilon_n$, the results may be summarized as follows:

$$\begin{aligned} D_{n+1} &= 2D_n + (D + 1)^n + \Delta\epsilon_n, \\ T_{n+1} &= 4T_n + (T + 1)^n + 3D_n, \\ U_{n+1} &= 6U_n + (U + 1)^n + 46T_n - 4T_{n+1} \\ &\quad + 30D_n - 6D_{n+1} + 6\Delta\epsilon_n. \end{aligned} \tag{8}$$

The expressions in parentheses, it will be remembered, are short-hand binomial expansions; thus:

$$(D + 1)^n = \sum_{i=0}^n \binom{n}{i} D_i.$$

RELATIONS WITH THE EXPONENTIAL INTEGERS

The generating identities in equations 6.1 furnish immediate relations with the exponential integers, ϵ_n . Writing $\exp (e^t - 1)$ as $\exp t\epsilon$, as above, and passing from generating relations to coefficient relations, these results are as follows:

$$\begin{aligned} D_n &= \frac{1}{2}[(\epsilon + 2)^n - 2(\epsilon + 1)^n + \epsilon_n], \\ T_n &= \frac{1}{8}[(\epsilon + 4)^n - 6(\epsilon + 2)^n + 8(\epsilon + 1)^n - 3\epsilon_n], \\ U_n &= \frac{1}{8}[(\epsilon + 6)^n + 4(\epsilon + 5)^n - 15(\epsilon + 4)^n \\ &\quad + 35(\epsilon + 2)^n - 36(\epsilon + 1)^n + 11\epsilon_n]. \end{aligned} \tag{9}$$

Expanding internal parentheses by the binomial theorem, the general result is as follows:

$$A_n = \sum_{i=1}^n \binom{n}{i} \alpha_i \epsilon_{n-i},$$

where the coefficients α_i for the three cases are as follows:

A_n	α_i
D_n	$2^{i-1} - 1$
T_n	$(2^{i-1} - 1)(2^{i-2} - 1)$
U_n	$\frac{1}{8}[6^i + 4 \cdot 5^i - 15 \cdot 4^i + 35 \cdot 2^i - 36]$

Note that in the first case $(D_n)\alpha_1 = 0$, in the second $(T_n)\alpha_1 = \alpha_2 = 0$, in the third $(U_n)\alpha_1 = \alpha_2 = \alpha_3 = 0$. Thus a given table of values of ϵ_n up to $n = k$ determines D_n up to $k + 2$, T_n up to $k + 3$, and U_n up to $k + 4$.

Somewhat simpler relations may be derived as follows. Repeated differentiation of the generating identity of the ϵ_n with respect to t , and passage from the generating relations to coefficient relations leads to the following:

$$\begin{aligned} \epsilon_{n+1} &= (\epsilon + 1)^n, \\ \epsilon_{n+2} &= (\epsilon + 1)^n + (\epsilon + 2)^n, \\ \epsilon_{n+3} &= (\epsilon + 1)^n + 3(\epsilon + 2)^n + (\epsilon + 3)^n, \end{aligned}$$

or, in general:

$$\epsilon_{n+m} = \sum_{x=1}^m (\epsilon + x)^n S_{x, m}.$$

This formula may be inverted by the reciprocal relations for the Stirling numbers of the first and second kinds⁷ which run as follows: If

$$a_m = \sum_{x=1}^m b_x S_{x, m}$$

then

$$b_m = \sum_{x=1}^m a_x s_{x, m}$$

where $s_{x, m}$ is the Stirling number of the first kind defined by the recurrence relation

$$s_{x, m+1} = s_{x-1, m} - m s_{x, m}$$

and the boundary conditions $s_{m, m} = 1$, $s_{x, m} = 0$ $x > m$, $s_{0, m} = 0$, $m > 0$.

⁷ Nielsen: "Handbuch der Gamma Funktion," Leipzig, 1906, p. 69.

The inverted formula⁸ is:

$$(\epsilon + m)^n = \sum_{x=1}^m \epsilon_{n+x} S_{x, m}$$

A short table of the Stirling numbers of the first kind follows:

$m \backslash x$	$S_{x, m}$						
	0	1	2	3	4	5	6
0	1						
1	0	1					
2	0	-1	1				
3	0	2	-3	1			
4	0	-6	11	-6	1		
5	0	24	-50	35	-10	1	
6	0	-120	274	-225	85	-15	1

The three equations resulting from applying this transformation to equations (9) are as follows:

$$\begin{aligned} D_n &= \frac{1}{2}[\epsilon_{n+2} - 3\epsilon_{n+1} + \epsilon_n], \\ T_n &= \frac{1}{8}[\epsilon_{n+4} - 6\epsilon_{n+3} + 5\epsilon_{n+2} + 8\epsilon_{n+1} - 3\epsilon_n], \\ U_n &= \frac{1}{8}[\epsilon_{n+6} - 11\epsilon_{n+5} + 30\epsilon_{n+4} + 5\epsilon_{n+3} \\ &\quad - 56\epsilon_{n+2} - 5\epsilon_{n+1} + 11\epsilon_n]. \end{aligned} \tag{10}$$

For computing purposes, values of ϵ_n and $\Delta\epsilon_n$ up to $n = 10$ are given in Table II.

TABLE II
EXPONENTIAL NUMBERS

n	ϵ_n	$\Delta\epsilon_n$
0	1	0
1	1	1
2	2	3
3	5	10
4	15	37
5	52	151
6	203	674
7	877	3,263
8	4,140	17,007
9	21,147	94,828
10	115,975	562,595

⁸ Noting that $\sum_{x=1}^m a^x S_{x, m} = (a)_m$, where $(a)_m$ is the factorial symbol used throughout, the inverse relation may also be written:

$$(\epsilon + m)^n = \epsilon^n (\epsilon)_m$$

In this notation, the inverses to equations (2) for the impedance numbers have the following simple forms which are worth noting:

$$\begin{aligned} (D)_n &= d_n \\ (T)_n &= t_n \\ (U)_n &= u_n \end{aligned}$$

CONGRUENCES

For numerical checks, it is convenient to note the simplest congruences⁹ for the three numbers. These follow from the Touchard congruence for the ϵ numbers¹⁰ which runs as follows:

$$\epsilon_{p+n} \equiv \epsilon_{n+1} + \epsilon_n \pmod{p},$$

where p is a rational prime greater than 2.

Since by equations (10) each of the impedance numbers is a linear function of the ϵ numbers, each has a similar congruence as follows:

$$\begin{aligned} D_{p+n} &\equiv D_{n+1} + D_n \pmod{p}, \\ T_{p+n} &\equiv T_{n+1} + T_n \pmod{p}, \\ U_{p+n} &\equiv U_{n+1} + U_n \pmod{p}. \end{aligned} \tag{11}$$

Special values for the first few congruences are as follows:

n	Remainder, mod p		
	D_{p+n}	T_{p+n}	U_{p+n}
0	0	0	0
1	1	0	0
2	7	3	0
3	37	36	60

These are sufficient for checking every value in Table I at least once and the values for $n = 5, 6, 7, 8$ are checked twice.

ACKNOWLEDGMENT

This paper arose as a result of a suggestion made by R. M. Foster on a former paper¹¹ and thanks are also due him for continuous counsel and critical scrutiny which have enlarged the boundary and sharpened the outline of the problem.

⁹ The congruence $D_n = r \pmod{p}$ is equivalent to the equation $D_n = mp + r$, where m is an integer; that is, r is the remainder after division by p (or the remainder plus some multiple of p).

¹⁰ See E. T. Bell, "Iterated Exponential Integers," *Annals of Math.*, 39, 3 (July, 1938), eq. 1.101, p. 541.

¹¹ "A Ladder Network Theorem," *Bell System Technical Journal* 16, pp. 303-318 (July, 1937); see especially footnote 3; I take this opportunity to draw attention to an error in that footnote: for four terminals (see Table 1) there are 157, not 64, measurable impedances; hence the upper bound to the number of representations is 18,883,356,492, not 74,974,368.

Copper Oxide Modulators in Carrier Telephone Systems *

By R. S. CARUTHERS

Copper oxide modulators are widely used in telephone systems for translating either single speech channels or groups of speech channels to carrier frequency locations on the lines. A number of simple circuit arrangements have been developed that enable suppression of certain undesired frequencies to a degree that is impractical in tube modulators. These modulators transmit equally well in either direction and the modulating elements are more non-linear than in tube modulators. As a result numerous effects are found that ordinarily are not important in the tube arrangements. Analytical studies have been considerably simplified by the use of a small signal, and a large carrier controlling the impedance variation of the copper oxide. It is found in this case that the superposition and reciprocity theorems hold for all the circuits that it has been possible to analyze even though the modulator is made up of non-linear elements. Open and short-circuit impedance measurements can be made use of as in four-terminal linear networks, and a generalized reflection theory developed. Performance data are given for an idealized modulator under a variety of operating conditions.

INTRODUCTION

AT least as early as 1927, copper oxide rectifiers were being tried as modulators for the speech channels of carrier telephone systems in this country. At this time only a rather large type of rectifier was available, better adapted for power use rather than in modulating the few milliwatts of a speech signal. Largely because of instabilities these early units were found to be unsatisfactory for modulator use. Further developments in copper oxide rectifiers made in various laboratories extended the variety and improved the quality of the product available, so that by about 1931 they began to be promising as serious competitors for vacuum tubes in modulators. Since 1931 continued improvements in copper oxide rectifiers have rapidly increased their field of application until now they are employed in practically all modulators of the latest types of carrier telephone systems.

In the new systems a copper oxide modulator is used instead of the previous push-pull arrangement of two vacuum tubes. In cable,¹

* Presented at Winter Convention of A.I.E.E., New York, N. Y., January 23-27, 1939.

¹ The general features of carrier telephone terminals have been described in a paper "Cable Carrier Telephone Terminals" by R. W. Chesnut, L. M. Ilgenfritz and A. Kenner, *Electrical Engineering*, January 1938.

open wire and coaxial carrier systems, from twenty-six to twenty-eight of these modulators are needed in each direction for translating each twelve-channel group of speech bands from voice to carrier and back again. These copper oxide modulators have no power costs, tube replacements or possibilities of power failures. In Fig. 1 the four $3/16$ inch diameter copper oxide discs generally used in a carrier telephone modulator are shown individually, assembled with connections, and potted in a can.

The carrier terminals have tended to become increasingly complex as it became their function to place more and more channels on a single pair of wires. The extreme simplicity and reliability of copper oxide modulators have been of great value in helping to overcome this tendency. Copper oxide modulators have been used from zero fre-

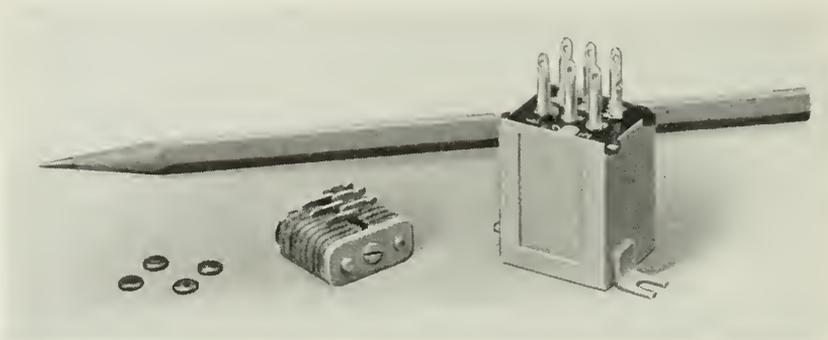


Fig. 1—Four disc copper oxide modulator.

quency to nearly four million cycles. Certain modulators for coaxial carrier systems have been designed to modulate simultaneously as many as sixty speech channels spaced over a 240,000-cycle band of frequencies.

Copper oxide modulators probably differ most from tube modulators because the simplicity of the rectifier elements allows a much greater variety of circuit arrangements to be used. Although the underlying principles of operation are not new, it has become necessary to investigate numerous transmission effects that could be neglected in tube modulators. This has resulted not only from the newer circuit arrangements with their smaller losses, but also from higher transmission standards for the overall system along with the greatly increased numbers of modulators in long circuits. Copper oxide modulators, unlike tube modulators, transmit signals equally well in either direction. While this is a simplification in allowing a modulator also to

be used as a demodulator, the modulator becomes complicated by the effects of reflections back and forth into the signal bands of numerous frequency bands of modulation products.

CIRCUIT ARRANGEMENT

The circuit arrangements used in copper oxide modulators generally are concerned either with carrier suppression, with carrier transmission along with the signal, or with balancing action to suppress certain unwanted bands of signal frequencies. In most carrier telephone systems economy of frequency space and amplifier load capacity demands the use of single-sideband, carrier-suppressed transmission.

In Figs. 2(a), 2(b), and 2(c) three types of copper oxide modulators are shown, each arranged to suppress the carrier in both the signal input and the signal output circuits. In Figs. 2(d) and 2(e) the carrier is balanced out in only one signal branch. In the usual arrangements a signal band selective filter must be used in each signal branch to restrict transmission to that of the wanted frequency band. Largely in this way interferences are guarded against, not only into other channels or systems to which the modulator output circuit is connected on the line or at the distant end, but also back into the complex array of facilities to which the input circuit may be connected.

In any of the circuits shown, modulation results from either the reduction or reversal of the current flow between the input and output signal circuits at periodic intervals as the carrier varies the copper oxide resistance back and forth between high and low values. In Fig. 2(a) where the connections of the input and output signal circuits are periodically short-circuited by the carrier-actuated copper oxide, transmission of the modulated signal into the input circuit or the unmodulated signal into the output circuit is prevented by filters, each of high impedance at the frequency of the other signal. In Fig. 2(b) the connections between the signal and modulated signal circuits are open-circuited periodically by the carrier. In this case each filter should have a low impedance at the other signal frequency. In Figs. 2(c), 2(d) and 2(e) the copper oxide rectifiers are made to become alternately low and high resistance in pairs as the polarity of the carrier is either in the same direction as the arrows or in the opposite direction. As a result, current flow from the input signal circuit into the output is periodically reversed by provision of a periodically reversing low impedance path. In effect each signal is balanced from the other's circuit.

Although an indefinite number of other circuit configurations can be used, no novel transmission feature would be found which was not

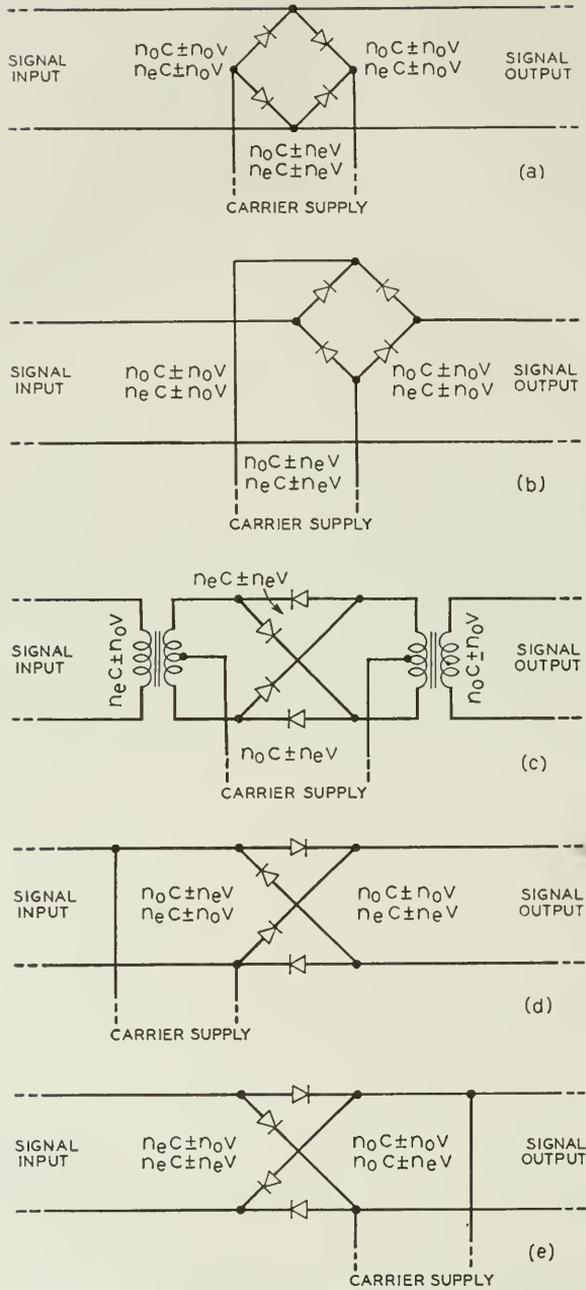


Fig. 2—Types of copper oxide modulator circuits.

present in the five circuits already shown. Third order modulators in which the copper oxide is arranged to give equal interruptions to the signal in both positive and negative half cycles of the carrier are exceptions not considered here. In addition, circuits like Hartley's, in which phase discriminations have been obtained in the sideband outputs from two modulators by altering both the carrier and signal input phase of one, can be viewed as composed of two modulators of any of the types illustrated.

In these copper oxide modulators all modulation product frequencies can be grouped into four classes:

$$\begin{aligned} n_0c \pm n_0v, \\ n_0c \pm n_e v, \\ n_e c \pm n_0v, \\ n_e c \pm n_e v, \end{aligned}$$

in which c and v are the carrier and input signal frequencies and n_0 is any odd number 1, 3, 5 etc., while n_e is any even number 0, 2, 4, 6, etc. If c and v contain more than one frequency each, n_0 and n_e are respectively the odd and even combinations of all multiples of the c and v frequencies. All frequencies of one of these four types appear together in a specific branch of the modulator circuit; and they will not appear in another branch unless from a dissymmetry among the copper oxide units or unless inherent in the circuit configuration. The branches in which the modulation products appear are shown in the circuits illustrated. It is apparent that only in the case of the double-balanced circuit of Fig. 2c, are all of these types of products completely separated in different parts of the circuit. In the other circuits shown the classes of products appear together in combinations of two types. In any balanced circuit that can be drawn the above relationships will be found to hold.

Modulation products will be of a type to which the circuit offers some degree of balance, of a type that can be made to vary in importance relative to the signal by adjustment of either the carrier or signal voltage, or of a type to which neither balance nor level adjustment is of any benefit. Satisfactory operation of such modulators requires large carrier voltages relative to those of the signal, so that products like $c \pm v$, $2c \pm v$, $3c \pm v$, etc. tend to be of large magnitude while products like $c \pm 2v$, $c \pm 3v$, etc. tend to be small. Furthermore, the former types can be made to predominate even more over the latter types either by increasing the carrier amplitude or by decreasing the signal levels. A 6 db reduction in signal results in 12 db reduction of $c \pm 2v$ and 18 db reduction in $c \pm 3v$. In any circuit

application interferences of this type lend themselves to reduction in so far as carrier power is available for high signal level operation, or in so far as noise does not limit for low signal levels. Laboratory measurements of some of these modulation products made during the development of a group modulator for a twelve-channel open-wire carrier system are shown in Fig. 3 for a double-balanced modulator

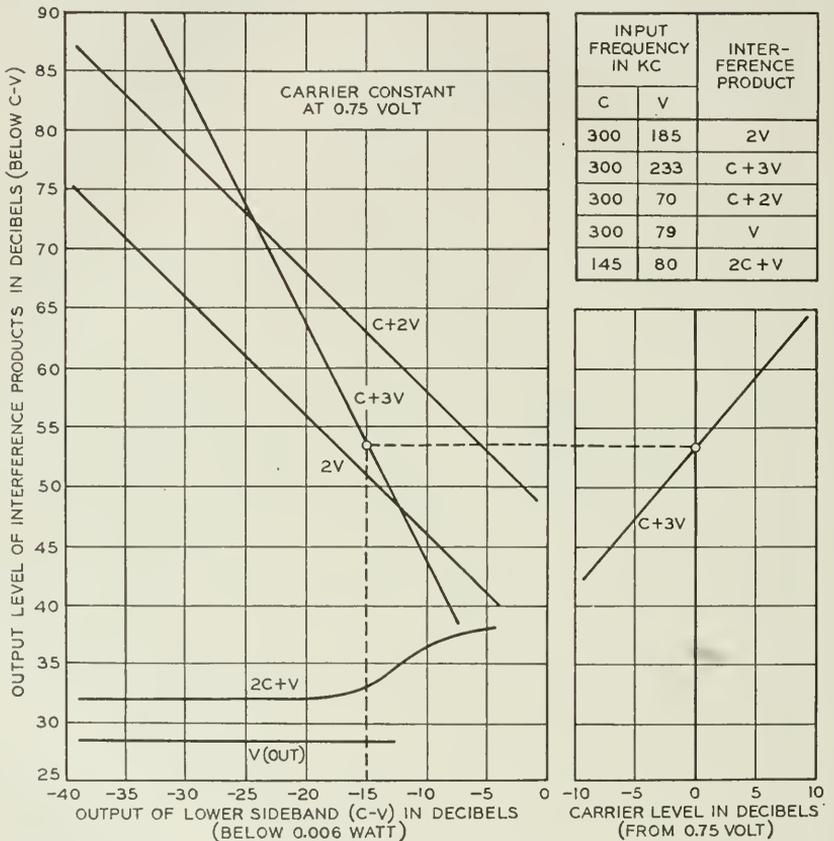


Fig. 3—Intensity of modulation products in a representative double-balanced copper oxide modulator.

like that of Fig. 2(c). Single 3/16 inch diameter discs as shown in Fig. 1 were used in each bridge arm. About 20 to 30 db reduction in interference by balance alone is quite readily obtained in the normal run of manufactured copper oxide rectifier elements, for those products to which the circuit arrangement offers a balance. Any further improvement must be obtained either by closer control of manufacture

or disc selection, by artificial balancing with some means such as condenser-resistance potentiometers, or by statistical averaging through use of numbers of discs in each bridge arm.

In single-channel modulators interferences caused by the signal into its own signal band will occur only in the presence of the signal. In such cases they need be only 20 to 30 db below the signal, except in special cases, as for example, modulators for broad-band program channels. In multi-channel systems interferences may be produced in the silent channels by the active channels. This kind of interference or crosstalk is ordinarily made to be 70 db or more below the wanted signals for commercial telephone service. In such cases overlapping bands of frequencies not improved by level adjustment are avoided by judicious choice of the carrier frequencies.

CIRCUIT IMPEDANCE AND LOSS

In all modulators the carrier serves merely as a means for obtaining a simple periodic variation of the impedance presented to the signals. It is not only immaterial to the signals how this variation is obtained, but the signals also are totally unaware of whether electrical, mechanical or other means are used, just so long as the signals themselves are unable to affect the time variation of this impedance. In a copper oxide modulator, only by making the carrier amplitude large compared to the signal amplitudes across the rectifier elements, can the impedance of the rectifiers be made to vary at carrier rather than signal rates. Too large a signal amplitude not only results in the production of undesired frequencies, but also the impedance and loss characteristics of the modulator vary with the signal amplitude. With small signals the carrier energy is used up in maintaining the copper oxide at prescribed impedance values at each instant of time, and none of the modulation products involving the signal receive more than a negligible amount of energy from the carrier. As a result the output signal energy will always be less than that of the input signal, partly because of i^2r losses within the copper oxide, and partly from the diversion of the input signal energy into the energies of the many modulation product frequencies.

The signal impedance of a copper oxide modulator is a combination of a characteristic impedance of the rectifier elements and the impedance of the connected circuits at all the modulation product frequencies. The characteristic impedance of the rectifier can be viewed crudely as an average of the impedance encountered by a small signal over a cycle of the carrier, treating each instantaneous value of the carrier voltage as a d-c. bias. If the impedance for small super-

imposed a-c. voltages is measured on a single copper oxide disc at various d-c. bias voltages, this impedance generally changes with both bias and frequency. Measurements to 200 kilocycles made on a 3/16 inch diameter disc are shown in Fig. 4. At all negative bias

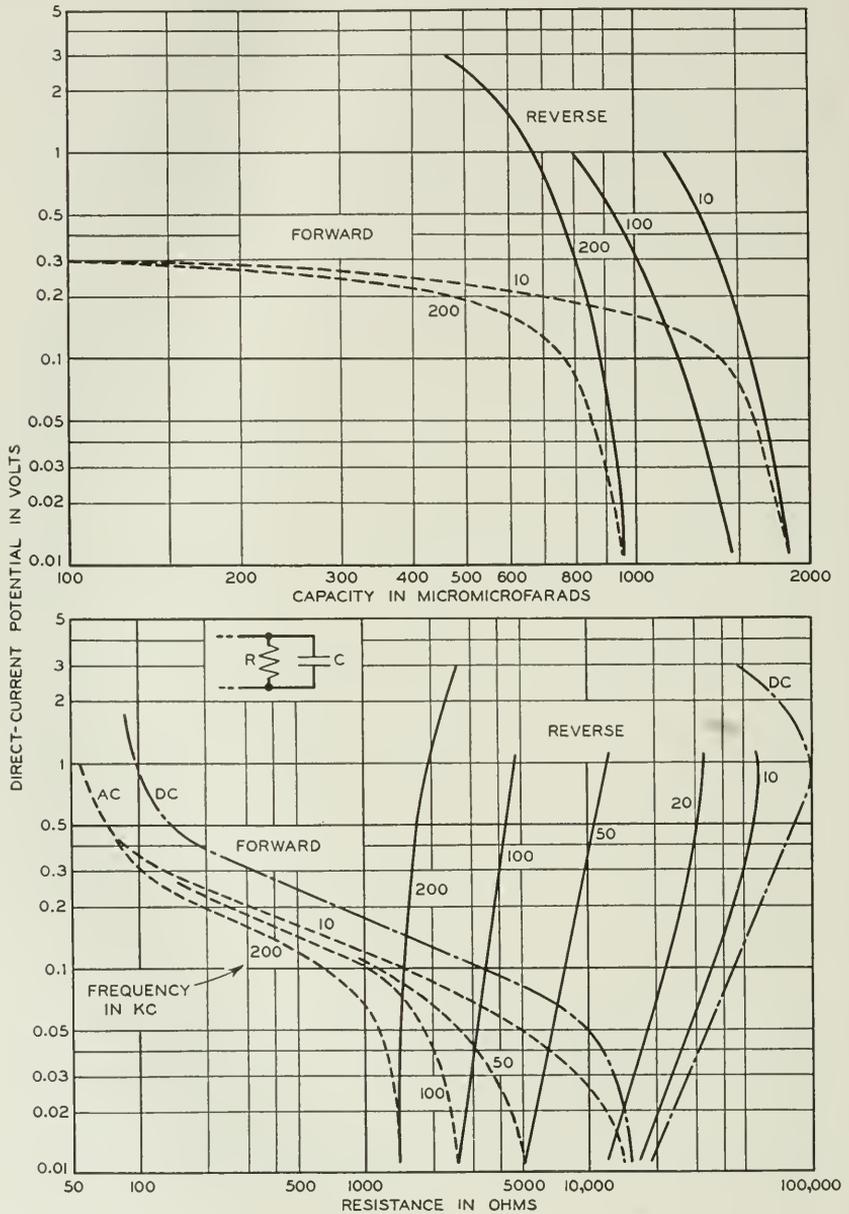


Fig. 4—Impedance of a copper oxide disc at various forward and reverse d-c. voltages for small superposed a-c. voltage.

voltages the impedance is a resistance in parallel with a condenser. This resistance decreases rapidly with increasing frequency while the shunt capacity decreases only a moderate amount. At moderate positive voltages (about 1/2 volt) the impedance becomes resistive and does not change appreciably with frequency. Experimentally it is found that the signal impedances in resistance terminated modulators can be made largely free of reactance at high frequencies by using either large carrier amplitudes, inductive tuning of the copper oxide capacities, or lower impedance connected signal circuits to accentuate the importance of the low-resistance part of the copper oxide characteristic. Very much lower circuit impedances must be used at the higher frequencies. Where 600 to 1000 ohms is a satisfactory impedance at speech frequencies, less than 50 ohms may be the best impedance to use, at three megacycles.

Impedance measurements on a double-balanced modulator designed to translate a twelve-channel group of frequencies for cable carrier systems from a band at 60 to 108 kilocycles to a 12- to 60-kilocycle band are shown in Fig. 5 for several resistance terminations. Absence of any impedance irregularities with frequency is apparent. Also, the tendency is shown for the modulator impedance to become less reactive with lower resistance terminations.

Inasmuch as copper oxide discs are available in sizes from 1/16 inch to more than an inch in diameter, a wide range of circuit impedances are possible varying from only a few ohms to thousands of ohms. Large area discs roughly are equivalent to small area discs in parallel. Thus by using a disc of n times the area of a small one or n of the small ones in parallel, the best circuit impedance becomes $1/n$ th at the same carrier voltage. Either discs in series or ones of smaller diameter enable the impedance to be raised in a corresponding manner. The lower impedance and greater energy dissipations of larger discs, or paralleled smaller ones at the same carrier voltage, obviously allow greater input signal energies before the signal impedance and loss begin to vary with the signal, and overload distortion appears. Similarly series stacks of discs, or series-parallel combinations, offer wide choice in both the signal levels that can be satisfactorily modulated and in the impedance levels. Usually r.m.s. carrier voltages across individual discs in the conducting direction will best be made somewhere between 3/10 and 3/4 volts.

The impedances of the connected circuits at the modulation product frequencies react back on the signal impedances in a way similar to the way that the two terminating impedances of a four-terminal linear network react on each other. In the case of the copper oxide modu-

lator a reaction from some modulation product back into the signal impedance is less and less as the product becomes of higher order, or as the circuit loss to it becomes greater. Where the impedances of the connected circuits have bothersome interactions with the copper oxide impedance either at the signal frequencies or the lower loss modulation products, resistance pad separation is usually the simplest solution if the increased loss can be tolerated.

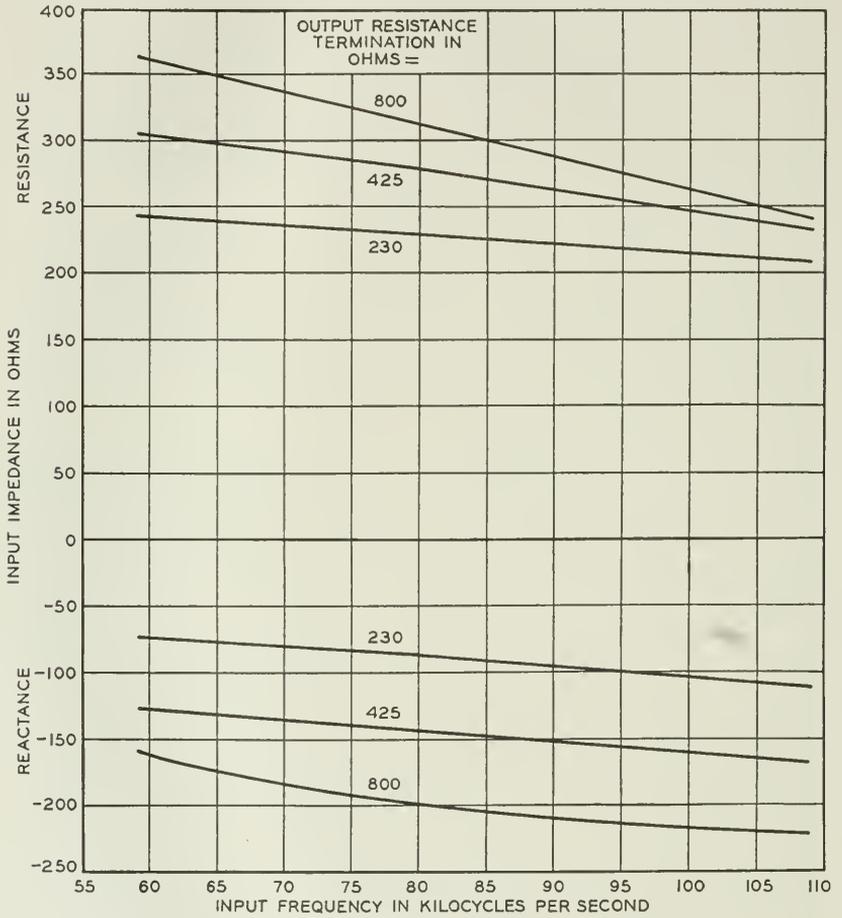


Fig. 5—Impedance of a representative double-balanced modulator.

Energy losses in copper oxide modulators between signal input and single sideband signal output have been found to be no greater than 8 or 9 db even at frequencies of 3 or 4 megacycles. At lower frequencies 5 or 6 db losses are normal, but losses as low as 2 db have been

obtained under less practical operating conditions. Experimental loss measurements are shown in Fig. 6 for a double-balanced modulator using single 3/16 inch diameter discs in each bridge arm. This modulator was designed to simultaneously modulate sixty speech channels occupying a 240,000-cycle band width. The modulator loss, like the impedance, depends on the impedance terminations of the modulator at all the modulation product frequencies as well as on the internal losses of the modulator. Short circuit, open circuit, or reactive terminations at the unwanted frequencies, permit energy losses only through reflections at the signal circuit junctions to the modulator or within the modulator. With proper terminations and loss-free copper oxide, 100 per cent efficiency frequency translations are theoretically possible. In a practical case, a larger carrier amplitude results in a smaller percentage of the time in which the rectifier

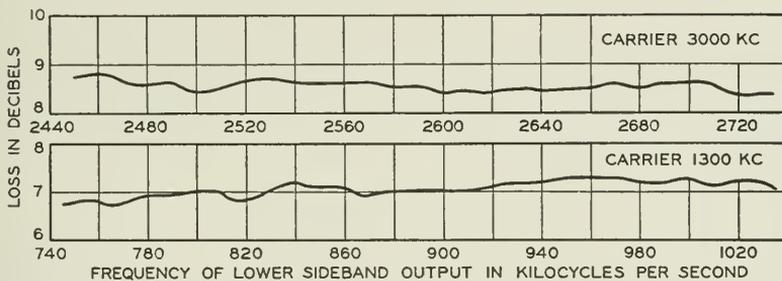


Fig. 6—Loss in a double-balanced group modulator for coaxial systems.

elements have impedances that are comparable to the connected circuits and that are neither blocking nor conducting. Signal energies are lost in this time interval, so a higher efficiency modulator results. The time spent on the intermediate resistance parts of the rectifier characteristic can be further reduced by introducing harmonics into the carrier wave, so that a square type of wave results. The resistance of the rectifier is abruptly switched back and forth between blocking and conducting values in this manner. When the connected circuit impedances at the unwanted frequencies are very high or low, best efficiencies result when transmission between the signal circuits is blocked most of the time. Thus in a circuit like that of Fig. 2(a) when the filters are high impedance at the unwanted products, highest efficiency results when the copper oxide is a low resistance short circuit for the major portion of the carrier cycle. In Fig. 2(b) an open circuit is desirable most of the time.

LINEAR MODULATOR THEORY

The analytical studies that have been of most benefit in the development of copper oxide modulators have made use of a variable resistance characteristic controlled by the carrier. This assumption has made it possible to investigate modulator performance² for a wide variety of characteristics under a great many operating conditions. Copper oxide modulator performance in particular cases as well as the effects of the circuit elements on this performance can readily be inferred from the data at hand about these idealized modulators.

In limited space it is not possible to discuss the varieties of resistance modulators that have been analyzed. However, certain viewpoints will be discussed that have been very useful not only for obtaining solutions for some of the hypothetical cases, but also in supplementing laboratory experiments on actual modulators.

All of these analytical studies have assumed a signal sufficiently small compared to the carrier, that it can be varied in magnitude without noticeable changes in the signal impedance or in the linearity between input and output signal amplitudes. This is in agreement with design procedure, as the circuit impedances and losses are determined on such a linear basis.

SUPERPOSITION PRINCIPLE

All of the modulator circuits with which we have been dealing, though composed of non-linear elements, have been resolved into the equivalent of linear systems by virtue of using a large carrier and small signal amplitude. We may simultaneously apply any number of signal frequencies, but all have negligible effect on the periodic changing of the non-linear element resistance by the carrier. These frequencies may be modulation product voltages, some applied at the output terminals and some at the input terminals, but in all cases, even though frequencies may coincide, it can be shown that the principle of superposition will hold without interaction between the applied forces and the responses. This permits a great simplification in the mathematical approach to modulator analysis, because the modulation product or signal voltages can be applied one at a time and the current responses summed. The voltage-current ratios at each frequency can then be replaced by equivalent impedances.

Any non-linear resistance like copper oxide will have a current-voltage characteristic that can be expressed as accurately as de-

² A physical picture of modulator performance in terms of linear networks is developed in a paper published in the January 1939 *Bell System Technical Journal*, "Equivalent Modulator Circuits" by E. Peterson and L. W. Hussey.

sired by

$$i = a_1e + a_2e^2 + \dots + a_n e^n. \tag{1}$$

If a carrier and signal voltage are applied to this non-linear element, each term beyond the first will independently produce currents of new frequencies composed of the intermodulation products of these two voltages. If in turn the external impedances at these new frequencies are not zero, new voltages will appear across the non-linear element to produce still more new frequencies. In this case the simplest consideration is to minimize the number of voltages by presenting zero impedance to the modulation products. If the carrier voltage is $C \cos ct$ and the input signal voltage $S \cos st$, the current flow in the n th term is

$$i = a_n(C \cos ct + S \cos st)^n. \tag{2}$$

In the binomial expansion of this expression it is obvious that linear response to the signal and freedom from distortion result when the ratio of carrier to signal is made sufficiently large so that only the first two terms are important.

$$i \approx a_n[(C \cos ct)^n + n(C \cos ct)^{n-1} \cdot S \cos st]. \tag{3}$$

The first term in equation (3) is the current flow at d-c. and harmonics of the carrier; it has no effect on the input signal and output signal except in so far as impedance termination presented across the non-linear element at the carrier harmonic frequencies may alter the carrier voltage harmonic content. The signal input current and the signal output current, as well as the unwanted modulation products of the signal, result from the second term. The even values of n produce the even order sidebands, second order being the output signal, while the odd values of n produce the input signal current and the odd order sidebands. These currents can be evaluated from

$$\cos^n b = \frac{K_n^n}{2^n} + \sum_{m=1}^{m=n} \frac{K_{n-m}^n}{2^{n-1}} \cos mb,$$

in which $K_{\frac{n-m}{2}}^n$ is equal to the combination of n things taken $\frac{n-m}{2}$ at a time for $\frac{n-m}{2}$ integral and is equal to zero for $\frac{n-m}{2}$ non-integral.

The signal input current is

$$i_s = \frac{a_n n K_{\frac{n-1}{2}}^{n-1}}{2^{n-1}} C^{n-1} S \cos st, \quad (4)$$

while the second order output signal sideband is

$$i_{c \pm s} = \frac{a_n n K_{\frac{n-2}{2}}^{n-1}}{2^{n-1}} C^{n-1} S \cos (c \pm s)t. \quad (5)$$

Similarly, if the output signal voltage at second order sideband frequency $(c \pm s)$ had been applied along with the carrier in place of the input signal, and of an equal amplitude, then the following currents of the output and input signal frequencies would result:

$$i_{c \pm s} = \frac{a_n n K_{\frac{n-1}{2}}^{n-1}}{2^{n-1}} C^{n-1} S \cos (c \pm s)t, \quad (6)$$

$$i_s = \frac{a_n n K_{\frac{n-2}{2}}^{n-1}}{2^{n-1}} C^{n-1} S \cos st. \quad (7)$$

If both signal input and output frequency voltages are applied simultaneously, equation (3) then becomes

$$i \approx a_n [(C \cos ct)^n + n(C \cos ct)^{n-1} \cdot S \cos st + n(C \cos ct)^{n-1} \cdot S \cos (c \pm s)t]. \quad (8)$$

The current responses obviously are the sum of the separate responses from independent application of the two frequencies. Even if a complex array of terminating impedances are supplied so that voltages appear across the non-linear element at all the modulation product frequencies, each new voltage will individually produce its own current response, quite independently of the responses that are being produced by the other voltages. It can readily be seen then that superposition does not depend on any assumptions about what the terminating impedances may be.

RECIPROCAL THEOREM

Equations (5) and (7) show that the sideband response to an input signal voltage is exactly equal in magnitude to the input signal response to the same amplitude sideband voltage. It can readily be seen that any two modulation products also bear such a reciprocal relation-

ship between their voltages and currents, as a result of using the same amplitude and frequency of carrier harmonic multiplier of their respective voltages to modulate between the two frequency positions. Although reciprocity has been proved valid here only for short-circuit terminations at the modulation product frequencies, it can also be proved under numerous other conditions of circuit operation. It seems that, regardless of modulator complexity of impedance terminations or frequency loss effects, *the reciprocal theorem is a necessary attribute of such a linear and bilateral system in which there are no internal energy sources.* Two-way systems in which an amplifier for example, is included as an internal energy source in one or both directions will, of course, violate the reciprocal theorem if the gains in the two directions are different. This arrangement is, however, both bilateral and linear.

COMPLETE PERFORMANCE CRITERIA

The laws for transmission between a signal input frequency and a signal output frequency can be completely specified from open and short circuit impedance measurements at the signal input and output frequencies, regardless of the complexity of the modulator. (From such measurements optimum impedance terminations can even be determined for linear-bilateral systems with internal energy sources.)

The four-terminal network of Fig. 7 is assumed to represent a modu-

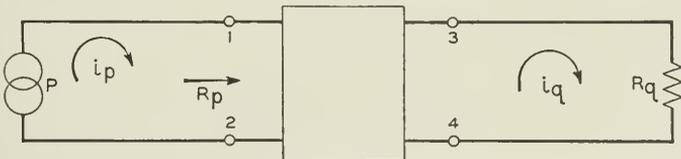


Fig. 7—Four terminal network equivalent of a linear modulator.

lator with a large carrier amplitude and having small signal voltage P of frequency p applied at the input terminals 1-2 at the left. Current of the output signal frequency q flows out of the terminals 3-4 into the impedance R_q . The generator P is assumed to have zero internal impedance at its own frequency. Impedance terminations at the 1-2 and 3-4 terminals at all other modulation product frequencies are perfectly general; whatever they are in a particular case, it is assumed that they are undisturbed as the terminations of the input and output at the signal frequencies are varied between open circuit and short circuit. The following symbols are used for the impedances looking into the modulator at the input terminals at input signal frequency p and at the output terminals at signal output frequency q .

Z_{p0} = impedance at p for open circuit at q ;

Z_{q0} = " " q " " " " p ;

Z_{ps} = " " p " short " " q ;

Z_{qs} = " " q " " " " p ;

R_q = impedance termination at frequency q ;

R_p = impedance of modulator at frequency p with R_q at 3-4 terminals for frequency q ;

K_a = transfer admittance between voltage at frequency p applied at the 1-2 terminals and short-circuit current at frequency q flowing from 3-4 terminals;

K_b = transfer admittance between voltage applied at 3-4 terminals at frequency q and short-circuit current at frequency p flowing from 1-2 terminals.

For R_q first short-circuited

$$i_{q1} = PK_a, \quad (9)$$

$$i_{p1} = \frac{P}{Z_{ps}}. \quad (10)$$

If the short circuit on R_q is removed, the following two additional currents will flow due to superposition of a new voltage $-i_q R_q$

$$i_{q2} = \frac{-i_q R_q}{Z_{qs}} \quad (11)$$

and

$$i_{p2} = -i_q R_q K_b, \quad (12)$$

$$i_q = i_{q1} + i_{q2} = \frac{PK_a}{1 + \frac{R_q}{Z_{qs}}}, \quad (13)$$

$$i_p = i_{p1} + i_{p2} = \frac{P}{Z_{ps}} - \frac{PK_a K_b R_q}{1 + \frac{R_q}{Z_{qs}}}. \quad (14)$$

The efficiency of the frequency translation, measured by the ratio of the power delivered to R_q to the power into terminals 1-2, is

$$\eta = \frac{i_q^2 R_q}{i_p P} = \frac{\left(\frac{K_a}{1 + \frac{R_q}{Z_{qs}}} \right)^2 R_q}{\frac{1}{Z_{ps}} - \frac{K_a K_b R_q}{1 + \frac{R_q}{Z_{qs}}}}, \quad (15)$$

which is maximum for

$$R_q = \frac{Z_{qs}}{\sqrt{1 - K_a K_b Z_{ps} Z_{qs}}} \tag{16}$$

The maximum efficiency is then

$$\eta_{max.} = \frac{K_a^2 Z_{ps} Z_{qs}}{(1 + \sqrt{1 - K_a K_b Z_{ps} Z_{qs}})^2} \tag{17}$$

When equation (16) is substituted in (14) it is found that

$$R_p = \frac{P}{i_p} = \frac{Z_{ps}}{\sqrt{1 - K_a K_b Z_{ps} Z_{qs}}} \tag{18}$$

In order to evaluate $K_a K_b$, open circuit impedance measurements must also be made. By superposition methods like those used in obtaining equations (11) and (12) it can readily be shown that

$$\frac{1}{Z_{p0}} = \frac{1}{Z_{ps}} - K_a K_b Z_{qs}, \tag{19}$$

$$\frac{1}{Z_{q0}} = \frac{1}{Z_{qs}} - K_a K_b Z_{ps}. \tag{20}$$

From these two equations, it follows that

$$\frac{Z_{ps}}{Z_{p0}} = \frac{Z_{qs}}{Z_{q0}}, \tag{21}$$

$$K_a K_b = \frac{\frac{1}{Z_{ps}} - \frac{1}{Z_{p0}}}{Z_{qs}}, \tag{22}$$

and $K_a K_b$ can be determined from any three of the open-short measurements by using (21) and (22). It follows that

$$K_a K_b Z_{ps} Z_{qs} = 1 - \frac{Z_{ps}}{Z_{p0}} \tag{23}$$

Upon substitution in (16) and (18) it is found that

$$R_q = \sqrt{Z_{q0} Z_{qs}} \tag{24}$$

and

$$R_p = \sqrt{Z_{p0} Z_{ps}} \tag{25}$$

when 3-4 is terminated in R_q for maximum efficiency.

Open and short-circuit measurements enable us to compute the optimum efficiency from equation (17) only if the transfer admittance

K_a is known. If the reciprocal theorem holds, $K_a = K_b$ and K_a can be determined from (23). The optimum efficiency is then

$$\eta_{\max.} = \frac{1 - \sqrt{\frac{Z_{ps}}{Z_{p0}}}}{1 + \sqrt{\frac{Z_{ps}}{Z_{p0}}}} \quad (26)$$

If the input signal generator has an internal impedance, most efficient energy delivery to the modulator will, of course, result if this impedance is made equal to R_p .

Equivalent T , π and bridge networks can obviously be drawn from the open and short-circuit measurements as in four-terminal linear networks.

It appears that even in a plate or grid circuit modulator the formulae of equations (24), (25) and (26) can be applied to the plate or grid circuit, respectively, where the signals are small compared to the carrier, inasmuch as the modulating parts of these circuits are linear and bilateral with no internal energy sources.

DOUBLE-BALANCED OR REVERSING-SWITCH MODULATOR³

A number of interesting conclusions can be reached about copper oxide modulators by assuming that the copper oxide acts like a switch having a low-resistance value when the positive half-cycle of the carrier voltage is across the disc and a high-resistance value during the negative half-cycle. The circuits of Fig. 2(c), 2(d) or 2(e) can then be represented by the equivalent circuit of Fig. 8.

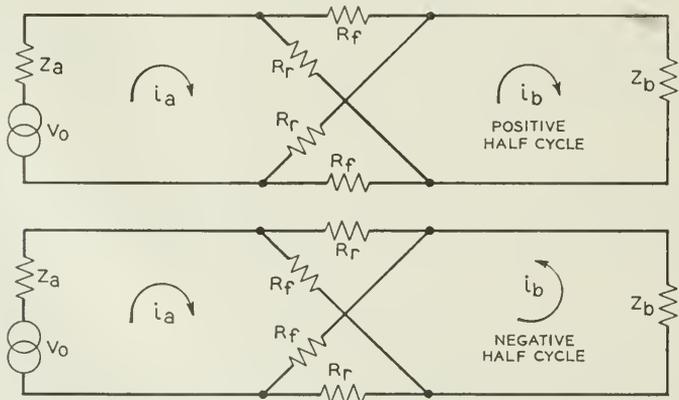


Fig. 8—Equivalent circuit of a double balanced modulator.

³ Referred to in the German literature as the "ring modulator."

sented by i_{3+} , i_{3-} , i_{5+} , etc. No currents flow at the sideband frequencies of the even harmonics of the carrier, i_{2+} , i_{2-} , \dots .

If V_0 should be replaced by an equal amplitude generator at any of the sideband frequencies, V_{1+} , V_{2-} , etc., then the input current would in any case be

$$I_{1+} = \frac{V_{1+}}{2R}, \quad I_{2-} = \frac{V_{2-}}{2R}, \quad \text{etc.}$$

The correspondence between the magnitudes of these entering currents and the magnitudes of the output currents at the modulation product frequencies are shown in Table I. Reciprocal relations between the

TABLE I
CORRESPONDENCE BETWEEN CURRENTS ENTERING AND LEAVING A
REVERSING-SWITCH MODULATOR

Component on One Side of Modulator	Corresponding Components on Other Side of the Modulator			
	$\frac{2k^*}{\pi}$	$-\frac{2k}{3\pi}$	$\frac{2k}{5\pi}$	$-\frac{2k}{7\pi}$
I_0	$I_{1-} \quad I_{1+}$	$I_{3-} \quad I_{3+}$	$I_{6-} \quad I_{6+}$	$I_{7-} \quad I_{7+} \quad \dots$
I_{1-}	$I_0 \quad I_{2-}$	$I_{2+} \quad I_{4-}$	$I_{4+} \quad I_{6-}$	$I_{6+} \quad I_{8-} \quad \dots$
I_{1+}	$I_0 \quad I_{2+}$	$I_{2-} \quad I_{4+}$	$I_{4-} \quad I_{6+}$	$I_{6-} \quad I_{8+} \quad \dots$
I_{2-}	$I_{1-} \quad I_{3-}$	$I_{1+} \quad I_{5-}$	$I_{3+} \quad I_{7-}$	$I_{6+} \quad I_{9-} \quad \dots$
I_{2+}	$I_{1+} \quad I_{3+}$	$I_{1-} \quad I_{5+}$	$I_{3-} \quad I_{7+}$	$I_{6-} \quad I_{9+} \quad \dots$
I_{3-}	$I_{2-} \quad I_{4-}$	$I_0 \quad I_{6-}$	$I_{2+} \quad I_{8-}$	$I_{4+} \quad I_{10-} \quad \dots$
I_{3+}	$I_{2+} \quad I_{4+}$	$I_0 \quad I_{6+}$	$I_{2-} \quad I_{8+}$	$I_{4-} \quad I_{10+} \quad \dots$
I_{4-}	$I_{3-} \quad I_{5-}$	$I_{1-} \quad I_{7-}$	$I_{1+} \quad I_{9-}$	$I_{3+} \quad I_{11-} \quad \dots$
I_{4+}	$I_{3+} \quad I_{5+}$	$I_{1+} \quad I_{7+}$	$I_{1-} \quad I_{9+}$	$I_{3-} \quad I_{11+} \quad \dots$
I_{5-}	$I_{4-} \quad I_{6-}$	$I_{2-} \quad I_{8-}$	$I_0 \quad I_{10-}$	$I_{2+} \quad I_{12-} \quad \dots$
I_{5+}	$I_{4+} \quad I_{6+}$	$I_{2+} \quad I_{8+}$	$I_0 \quad I_{10+}$	$I_{2-} \quad I_{12+} \quad \dots$
I_{6-}	$I_{5-} \quad I_{7-}$	$I_{3-} \quad I_{9-}$	$I_{1-} \quad I_{11-}$	$I_{1+} \quad I_{13-} \quad \dots$
I_{6+}	$I_{5+} \quad I_{7+}$	$I_{3+} \quad I_{9+}$	$I_{1+} \quad I_{11+}$	$I_{1-} \quad I_{13+} \quad \dots$
I_{7-}	$I_{6-} \quad I_{8-}$	$I_{4-} \quad I_{10-}$	$I_{2-} \quad I_{12-}$	$I_0 \quad I_{14-} \quad \dots$
I_{7+}	$I_{6+} \quad I_{8+}$	$I_{4+} \quad I_{10+}$	$I_{2+} \quad I_{12+}$	$I_0 \quad I_{14+} \quad \dots$
I_{8-}	$I_{7-} \quad I_{9-}$	$I_{5-} \quad I_{11-}$	$I_{3-} \quad I_{13-}$	$I_{1-} \quad I_{15-} \quad \dots$

NOTE: A current of the frequency indicated in the first column will be modulated to produce the components written on the same line, the magnitudes of which are the magnitude of the generating current multiplied by the factors at the top of the columns.

$$*k = \frac{\sqrt{\frac{R_r}{R_f}} - 1}{\sqrt{\frac{R_r}{R_f}} + 1}$$

driving voltage at one frequency and the output current at another frequency are obvious.

TABLE II

PERFORMANCE OF DOUBLE-BALANCED MODULATOR (IDEAL REVERSING SWITCH) FOR VARIOUS INPUT AND OUTPUT TERMINATIONS

Modulator Terminations				Modulator Impedance		Modulator Loss or Efficiency	
Input Circuit		Output Circuit		Input Signal	Output Signal	Voltage Ratio	db
Signal	Others	Signal	Others				
R	R	R	R	R	R	$\frac{2}{\pi}$	3.9
R	any value	R	R	R	R	$\frac{2}{\pi}$	3.9
R	R	R	0	$\frac{2R}{\pi^2 - 2}$	R	$\frac{2}{\pi}$	3.9
R	R	R	∞	$\frac{(\pi^2 - 2)R}{2}$	R	$\frac{2}{\pi}$	3.9
R	$\frac{(\pi^2 - 2)R}{2}$	$\frac{(\pi^2 - 2)R}{2}$	0	R	$\frac{\pi^2(\pi^2 - 2)R}{6\pi^2 - 16}$	$\frac{\pi}{\sqrt{2(\pi^2 - 2)}}$	2
R	$\frac{2R}{\pi^2 - 2}$	$\frac{2R}{\pi^2 - 2}$	∞	R	$\frac{(6\pi^2 - 16)R}{\pi^2(\pi^2 - 2)}$	$\frac{\pi}{\sqrt{2(\pi^2 - 2)}}$	2
R	0	R	0	0	0		∞
R	∞	R	∞	∞	∞		∞
R	0	R	∞	$\frac{\pi^2 R}{4}$	$\frac{4}{\pi^2} R$	$\frac{4\pi}{4 + \pi^2}$.85
R	∞	R	0	$\frac{4}{\pi^2} R$	$\frac{\pi^2}{4} R$	$\frac{4\pi}{4 + \pi^2}$.85
R	0	$\frac{4}{\pi^2} R$	∞	R	$\frac{4}{\pi^2} R$	1	0
R	∞	$\frac{\pi^2}{4} R$	0	R	$\frac{\pi^2}{4} R$	1	0
R_s	R_s'	R_r	R_r'	Z_i^*	Z_0^*	η^*	

$$*Z_i = \frac{\pi^2 R_r'(R_r + R_s') + 4(R_r - R_r')R_s'}{\pi^2(R_r + R_s') - 4(R_r - R_r')}$$

$$Z_0 = \frac{\pi^2 R_s'(R_s + R_r') + 4(R_s - R_s')R_r'}{\pi^2(R_s + R_r') - 4(R_s - R_s')}$$

$$\eta = \frac{4\pi(R_r' + R_s')\sqrt{R_s R_r'}}{\pi^2(R_s + R_r')(R_r + R_s') - 4(R_r - R_r')(R_s - R_s')}$$

GENERALIZED REFLECTION THEORY

Superposition permits us to apply simultaneously driving forces of the frequencies tabulated above in any relative phases and amplitudes that we care to choose on either side of the modulator. If simultaneously I_0 is applied on one side of the modulator and $(I_{1+}) \frac{2k}{\pi} \cdot \frac{Z_{1+} - R}{Z_{1+} + R}$ is applied to the other set of modulator terminals, then the total current at the output terminals at the sideband frequency (1^+) will be

$$(I_{1+}) \frac{2k}{\pi} \cdot \left[1 - \frac{Z_{1+} - R}{Z_{1+} + R} \right]. \quad (34)$$

This is equivalent to saying that a resistance R at the sideband frequency (1^+) has been connected to the output terminals of the modulator and in this resistance is an internal zero impedance generator of voltage

$$2R(I_{1+}) \frac{2k}{\pi} \cdot \frac{Z_{1+} - R}{Z_{1+} + R}. \quad (35)$$

This resistance R at sideband frequency (1^+) must be infinite at all other frequencies, if in parallel we assume another resistance of R at all frequencies except (1^+) at which it is infinite.

The equivalent impedance at frequency (1^+) at the modulator terminals connected to the (1^+) resistance with its internal generator, is the ratio of (1^+) voltage to (1^+) current.

$$Z = \frac{\frac{2k}{\pi} I_{1+} \cdot 2R - \frac{2k}{\pi} \cdot I_{1+} \cdot \left[1 - \frac{Z_{1+} - R}{Z_{1+} + R} \right] R}{\frac{2k}{\pi} I_{1+} \cdot \left[1 - \frac{Z_{1+} - R}{Z_{1+} + R} \right]}, \quad (36)$$

which reduces to

$$Z = Z_{1+}. \quad (37)$$

Z_{1+} may be real or complex as it involves only the amplitude and phase of the superimposed voltage of upper sideband frequency. It can readily be seen then that the solution for current flow at this frequency of equation (34) is identical with the case of linear networks in which the current is expressed as that flowing in a matched circuit modified by a reflection factor. Reflection from any modulation product frequency can be similarly treated.

A number of cases have been worked out of efficiencies and impedances in such modulators for transmission between an input signal and a single-sideband output signal. The modulating element has been assumed perfect ($k = 1$) and the terminations pure resistances. The results are shown in Table II.

ACKNOWLEDGMENTS

The writer wishes to acknowledge his appreciation of the assistance of numerous associates in the Bell Telephone Laboratories in arriving at the views on copper oxide modulator performance recorded in this paper. In particular, acknowledgment is due to Mr. R. W. Chesnut, Dr. E. Peterson and Dr. G. R. Stibitz.

Some Applications of the Type "J" Carrier System*

By L. C. STARBIRD and J. D. MATHIS

Previous papers before the American Institute of Electrical Engineers describe the development of a twelve-channel type J Carrier System. This paper discusses some of the practical problems encountered in extending the circuit capacity of existing open-wire lines by the use of this carrier system.

The first systems of this type were placed in commercial operation late in 1938. One of these systems is discussed in detail from the standpoint of obtaining satisfactory operation with the most economical arrangement of new and existing facilities.

A TWELVE-CHANNEL carrier telephone system for open-wire lines was described before the American Institute of Electrical Engineers early this year,¹ and a discussion of the requirements of line facilities for its operation is being presented.² Since the first three systems to be placed in commercial operation are located in Texas, it seems appropriate to present to the Southwest District Convention the major problems arising from the practical application of this type system on existing open-wire plant.

In 1935 it became apparent that existing open-wire facilities on some of the major toll lines in Texas would soon be exhausted. In the case of the Dallas-Houston, Dallas-San Antonio, and Dallas-Longview lines, current growth and requirements for the future indicated that while a toll cable would probably have to be provided ultimately, the development of the open-wire twelve-channel J carrier system makes available an arrangement for obtaining a large number of additional circuits over the existing lines to provide for the immediate requirements and also permit postponement of more costly relief measures for a number of years.

The type J system operates in a frequency range above that of the three-channel type C carrier system and can be superposed on the same conductors with the type C, thereby providing a total of sixteen circuits from one pair of conductors. However, conductors suitable

* Presented April 18, 1939 before the A.I.E.E. in Houston, Texas.

¹ "A Twelve-Channel Carrier Telephone System for Open Wire Lines," by B. W. Kendall and H. A. Affel, Winter Convention, A.I.E.E., 1939. *Bell System Technical Journal*, January 1939.

² "Line Problems in the Development of the Twelve-Channel Open-Wire Carrier System," by L. M. Ilgenfritz, R. N. Hunter, and A. L. Whitman, District Convention, A.I.E.E., Houston, 1939. This issue of the *Bell System Technical Journal*.

for type C carrier operation are not necessarily satisfactory for the operation of the new system.

The three lines under consideration were practically of the same construction, being twelve-inch phantom lines originally built for voice frequency circuits only and later modified for the application of type C carrier systems. Over lines of this type, it is practicable to operate a single type J system without any material change in the line wire because no crosstalk considerations are involved, although it is necessary to select by transmission measurement pairs which are free from absorption effects. Where more than one system is required a transposition arrangement has been designed for use with line conductors of a non-phantomed pair spaced six inches apart and thirty inches between conductors of horizontally adjacent pairs. This design can be used either for new wire or for existing wire retransposed, and can be applied without regard to the existing phantom transposition design, thereby permitting respacing and retransposing any portion of the existing wire, a phantom group at a time if desired.

ADVANCE ENGINEERING

With these operating limitations a review of the circuit requirements established a plan to place a J carrier system on one of the phantom groups of the Dallas-Longview line during 1938. This system would not only provide sufficient circuits to meet the additional requirements but would furnish sufficient spare circuits to release one phantom group of twelve-inch wire for respacing and retransposing. This plan was not applicable to the Dallas-Houston and Dallas-San Antonio lines since circuit relief was required for the 1937 business, and the J carrier system would not be available until 1938. These lines each consisted of five crossarms of 104 mil wire over the greater portion of their length. An inspection showed that, although the poles were of sufficient strength to support additional crossarms, it would be difficult to maintain the necessary wire clearance with an additional crossarm below the existing wire and also that new wire so placed would be susceptible to interference from possible breaks in the wire above.

The solution of this problem was the addition of a crossarm two feet above the others on a simple extension fixture. This fixture shown in Fig. 1 consists of a four-inch steel "H" beam fastened to the pole by the through bolts which also support the two upper crossarms. By placing four pairs of six-inch spaced conductors on the new crossarm and by using four type C carrier systems, sixteen additional circuits were obtained to furnish the circuit relief for 1937 and, in addition to

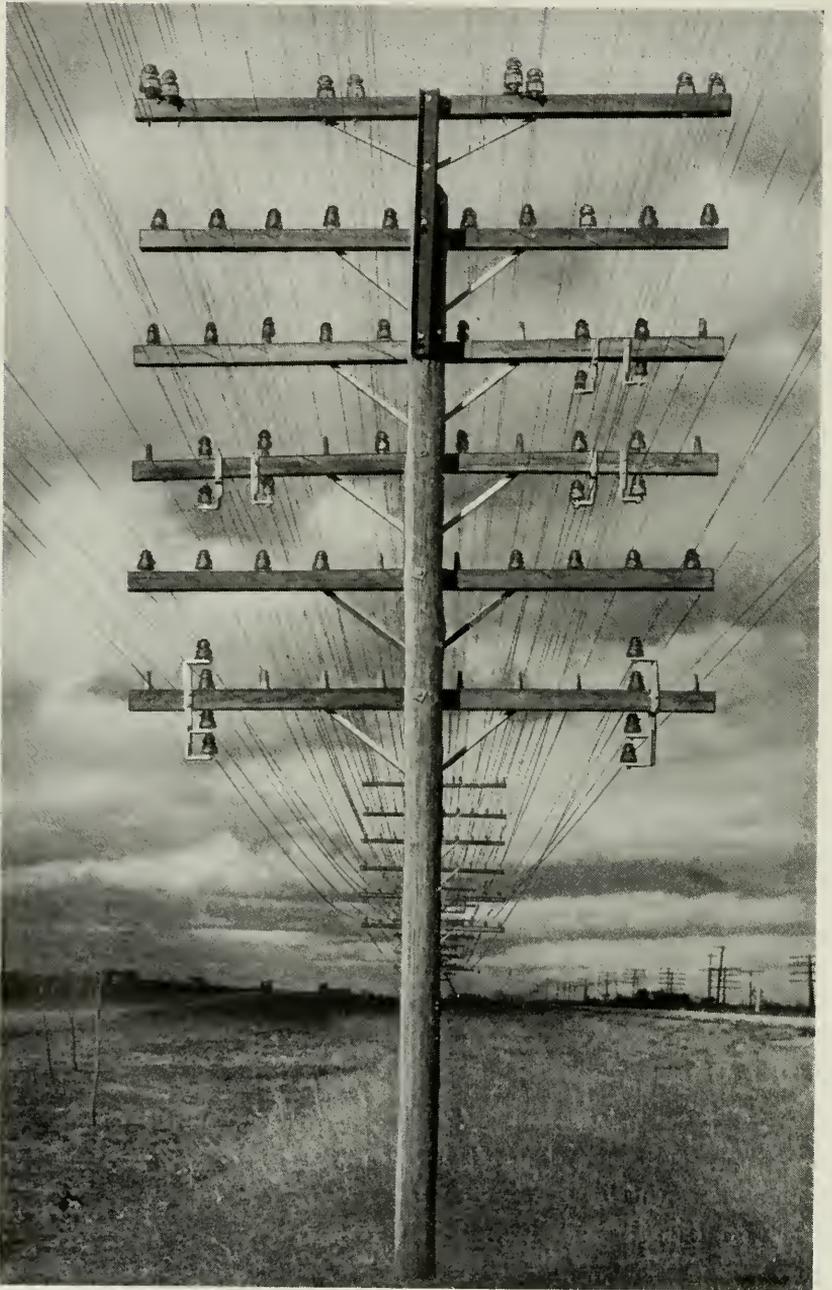


Fig. 1—Typical pole with extension fixture.

the immediate relief, four suitable J carrier paths were provided of which one on each line was needed in 1938. Figure 2 is a typical pole head and shows how the ultimate circuit capacity of this open-wire plant has been expanded from 69 to 133 circuits by the addition of one crossarm and eight conductors. The use of 128-mil wire instead of 104-mil wire provides greater strength and, considering the particular location, reduces the probability of interrupting sixteen circuits by a single wire break or other physical interference.

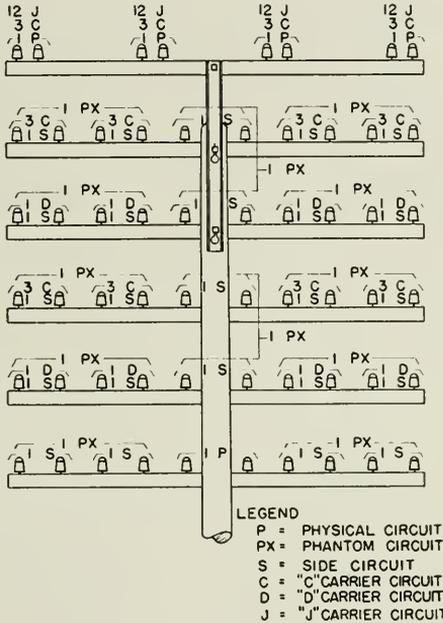


Fig. 2—Pole head diagram showing circuit capacity of the Dallas-Houston and Dallas-San Antonio lines.

The program of placing three type J carrier systems in service in Texas during 1938 was established. Figure 3 is a map of a portion of the state showing the routes of the lines and the principal cities along the routes. Since the length and attenuation of each of these lines are such that the carrier systems can not operate without intermediate amplification, it was necessary that the number and locations of repeater stations be determined.

TYPICAL SYSTEM

The layout of a particular system is largely controlled by available repeater gain, existing entrance cables, line attenuation under normal

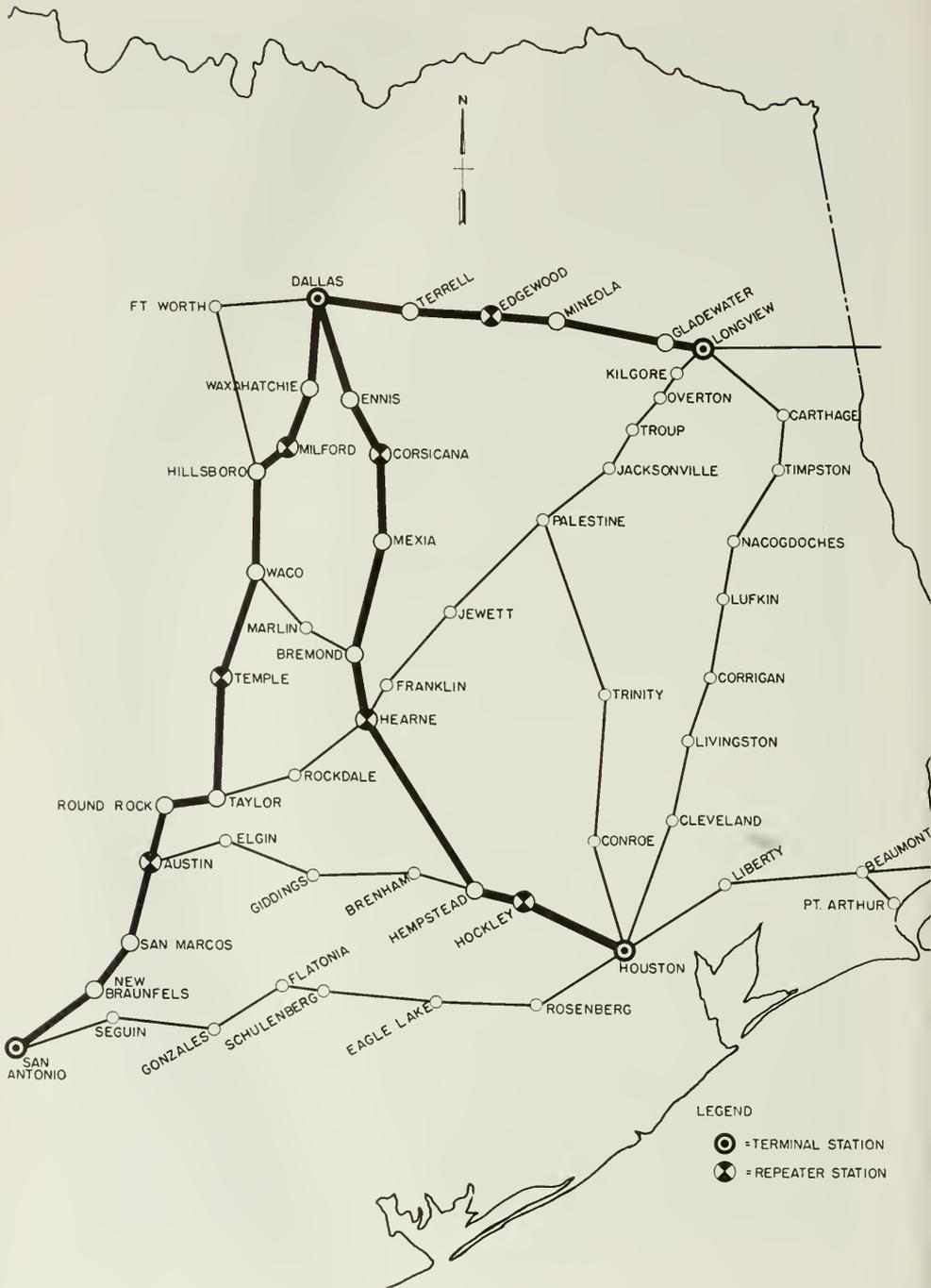


Fig. 3—Routes of toll lines on which the J carrier systems are applied.

and adverse weather conditions, location and availability of existing telephone buildings, and availability of commercial power for new buildings. Line attenuation is increased greatly by deposits of ice on the wire during sleet conditions. Although data are available regarding the frequency of large deposits of ice, there is very little information as to the amounts or frequency of occurrence of small deposits. Under normal wet weather conditions the maximum attenuation of six-inch spaced 128-mil facilities at 140 kilocycles is 0.35 db per mile.

On the Dallas-San Antonio line the facilities available consisted of 286 miles of six-inch spaced 128-mil copper wire and 42,000 feet of 16-gauge non-loaded paper insulated cable. Using repeaters having a maximum amplification of 45 db in each direction of transmission, the provision of two intermediate repeaters would provide sufficient gain to take care of the wet weather conditions with no extra margin; three repeaters would provide 45 db margin and four would provide 90 db margin for the overall system. Considering the location of this line and the small probability of obtaining large deposits of ice in accordance with past experience, it was decided to select tentatively three intermediate repeater stations which would provide sufficient gain to take care of attenuation up to about 0.5 db per mile as compared to the wet weather value of 0.35 db per mile.

For type C operation over this line, only one type C carrier repeater point is required, and it is at Austin. Considering the availability of power equipment and operating personnel and the possibility of future J carrier terminals being located at Austin, it is desirable that this be one of the repeater points on the J system. A division of the attenuation of the facilities north of Austin indicated that the other stations should be in the vicinity of Temple and Milford.

At these repeater stations amplification is needed only on the type J system and the other circuits on the line pass through these stations without amplification. Under these conditions energy may be transferred from the output of one type J repeater to the input of the same repeater or to the input of a repeater on another J system via crosstalk paths involving the wires which are not used for type J systems. The effect of this transfer of energy is accentuated by the fact that there is a large difference in transmission level between the output of one type J repeater and the input of the same or another repeater. In order to minimize these effects it is necessary that all wires on the line be given special treatment, including a gap in the toll line, longitudinal choke coils in all wires at terminal poles and crosstalk suppression filters in the non-J pairs in the repeater station itself. In selecting locations for repeater stations, consideration must also be

given to the possible coupling between type J systems by interaction paths involving other conductors adjacent to the toll line.

Before definite selection of repeater station locations may be made, it is necessary that each repeater section be checked in detail and in this check the entrance cable arrangement may be controlling. The newly developed spacer insulated spiral-four cable, either loaded or non-loaded, or non-loaded pairs of the conventional paper insulated cable may be used between the open wire and equipment. Generally the existing voice and C carrier circuits use loaded entrance cable pairs and in most cases a change to non-loaded facilities would require extensive rearrangements in these circuits. In order to use non-loaded pairs for the J carrier and leave the C carrier and voice on loaded facilities, filters are placed at the terminal pole to separate the J carrier frequencies from the C and voice frequencies at that point. A limitation on the use of existing cable is that suitable pairs must be selected by crosstalk measurements and balanced at 140 kilocycles to meet the requirements of the system. The paper insulated conductors have the largest attenuation of any of these facilities, and the loaded spiral-four the least. The various entrance arrangements from the open wire to the office equipment are described in more detail elsewhere.² The choice of the facility used in any particular case will depend upon the resultant overall economy.

The large number of non-loaded pairs in the existing 1.6 mile entrance cable at San Antonio indicated that sufficient pairs could be selected which would be satisfactory from the crosstalk standpoint for J carrier operation. Six pairs were subsequently selected and balanced.

At Austin a single toll entrance cable, one mile in length, with two complements, terminates the line from the two directions. Although the two complements are separated by a layer shield, this cable is not suitable from a crosstalk standpoint for operation of the J carrier in and out of the office; therefore, at least one additional cable is required from the central office to the toll line. For this purpose a new non-loaded spiral-four entrance cable was indicated for the type J system with the type C and voice circuits continuing to use the existing cable. The separation of the type J circuits from the non-J circuits on the same pairs is accomplished by filters which are located in a small building at the junction of the toll line and the entrance cables. The use of a single entrance cable for the non-J wire in both directions on the telephone line indicated that it might be necessary in the future to use crosstalk suppression filters at this point. Accordingly, the filter hut was made large enough to include future crosstalk suppression

² Loc. cit.

filters if required as well as the line filters which separate the type J from the non-J circuits.

A repeater station at Temple could have been located in the existing central office or could be located in a separate building in or near the city. In either case a new power plant was needed since the existing plant could not be economically modified to serve the J carrier repeaters. The telephone line is continuous through the city, only those wires used for Temple circuits being terminated in the office through one entrance cable 0.6 mile in length. This cable is not suitable for J operation in both directions, which would require one additional cable if the repeaters were located in the central office. Numerous signal and supply lines in proximity with the telephone line within the city offered interaction crosstalk complications. A separate repeater station near the toll line in or near the city avoids the placing of a long entrance cable, reduces the overall system attenuation, and eliminates the problem of interaction crosstalk from paralleling lines. Other factors including cost showed very little difference between a separate station and placing the repeaters in the central office. An unattended station near the toll line was indicated.

A common entrance cable at Dallas terminates the wire on both the Houston and San Antonio lines, the terminal of the Houston line being 2.9 miles from the central office, and of the San Antonio line one mile further. This cable previously had been placed in three different sections, each section having a different make-up, and there was considerable doubt as to the number of suitable pairs for J operation that could be obtained. The use of either a loaded or non-loaded spiral-four cable would not improve attenuation sufficiently to change the number or materially alter the locations of the repeater stations from those tentatively selected, but would provide some additional margin for sleet conditions. The expense of loading the spiral-four cable, if placed, could not be justified by the improvement in overall attenuation. Using either non-loaded spiral-four or existing non-loaded paper insulated conductors requires filters at the open-wire terminus. With these considerations, it was decided that suitable pairs would be used in the existing cable until exhausted. Subsequent crosstalk selection tests have indicated that twelve pairs, six for each line, are available.

Since there was no suitable central office building at Milford, the repeater station in that vicinity must of necessity be in a new building preferably near the toll line. Commercial power is available only near the town, forcing a tentative location to be selected at the edge of the city.

TABLE I
DISTRIBUTION OF GAIN AND LOSS BY REPEATER SECTIONS

Repeater Section	CABLE		OPEN WIRE			
	Length Miles	Loss db	Length Miles	Wet Weather Loss db	Maximum Tolerable Attenuation in db per Mile at 140 KC Using	
					45 db Repeaters	75 db Repeaters
Dallas-San Antonio System						
Dallas-Milford.....	3.9 mi., 16 ga.	17.50	49.2	16.4	0.560	1.165
Milford-Temple.....	Nominal	Nominal	84.7	27.3	0.532	0.885
Temple-Austin.....	1.1 mi., Spiral-4	2.20	74.4	24.8	0.575	0.980
Austin-San Antonio.....	{ 1.1 mi., Spiral-4 1.8 mi., 16 ga. }	10.30	78.5	26.2	0.635	0.825
Dallas-Houston System						
Dallas-Corsicana.....	2.9 mi., 16 ga.	13.20	52.5	17.5	0.610	1.175
Corsicana-Hearne.....	0.5 mi., 16 ga.	2.25	90.5	30.2	0.484	0.805
Hearne-Hockley.....	Nominal	Nominal	85.6	28.6	0.545	0.875
Hockley-Houston.....	5.6 mi., 16 ga.	25.20	29.1	9.7	1.190	1.700
Dallas-Longview System						
Dallas-Edgewood.....	1.4 mi., 16 ga.	6.30	55.3	21.0	0.984	1.240
Edgewood-Longview.....	0.5 mi., 16 ga.	2.30	69.7	26.5	0.614	1.040

With these selections of entrance cable facilities and tentative repeater station locations, the distribution of gain and line loss by repeater sections is shown in Table I. A satisfactory distribution of line loss has been obtained and an analysis of these data shows that further improvement is impracticable. Therefore, the tentative repeater station locations were adopted.

Figure 4 is a diagram of the major line and equipment parts of the Dallas-San Antonio lead. The J carrier path is shown by heavy solid lines, the C and voice on the same wire with the J by light solid lines, and all other circuits, classed as non-J, by dotted lines. Figures 5 to 8, inclusive, show in more detail the arrangements at the huts and repeater stations. The figures for the Dallas Hut and Temple Repeater Station are typical, and huts and unattended repeater stations not shown differ from these only in minor details. It will be noted that all wire on the toll line is brought through the repeater stations while only that wire on which J carrier is superposed is routed through the huts except at Austin where all wire to the north is brought through the hut to allow the future application of crosstalk suppression filters if required. For both huts and unattended repeater stations, short lengths of loaded spiral-four conductors are used from the six-inch spaced wire at the terminal poles to the equipment in the buildings. A single continuously adjustable load unit is used for each pair and is located with the equipment. Paper insulated pairs under the same cable sheath as the spiral-four conductors are used for the non-J wire.

As previously mentioned, the conditions at Austin were complicated by a single cable for existing circuits and a new cable for J carrier in both directions. Figure 9 is a diagram of the existing and new cables to the filter hut and terminal poles, and Fig. 10 shows the interconnection of circuits and equipment used at the filter hut, terminal poles, and central office.

Terminal and repeater equipment in existing offices is located in space adjacent to other equipment terminating toll circuits, and makes use of the common office equipment and power plant. The relation of the J carrier terminals to the other equipment in the Dallas Toll Office is shown in Fig. 11.

The new repeater stations and the filter huts are arranged for unattended operation. The equipment in the filter huts is such that no adjustment or attention is required other than periodic inspections. In the unattended repeater stations the power supply equipment is automatic in its operation. Although periodic maintenance attention is necessary, it is desirable that any abnormal condition be recognized as soon as practicable and a system of alarms has been provided from each unattended station to an adjacent main repeater or terminal

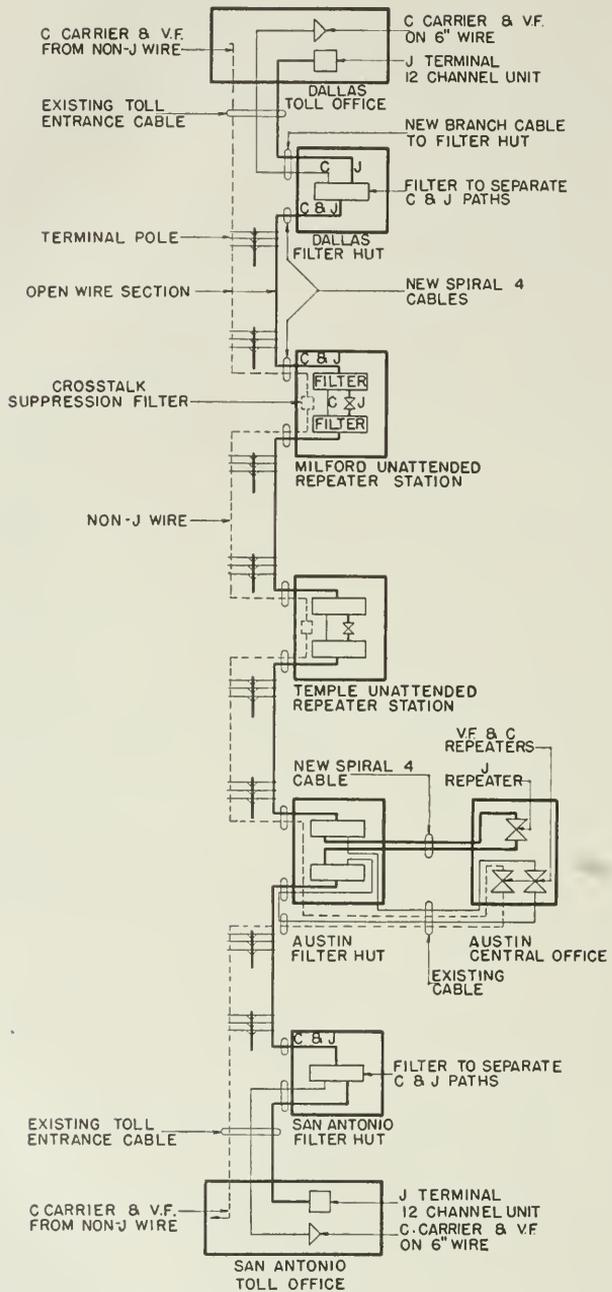


Fig. 4—Arrangement of facilities for a typical J carrier system.

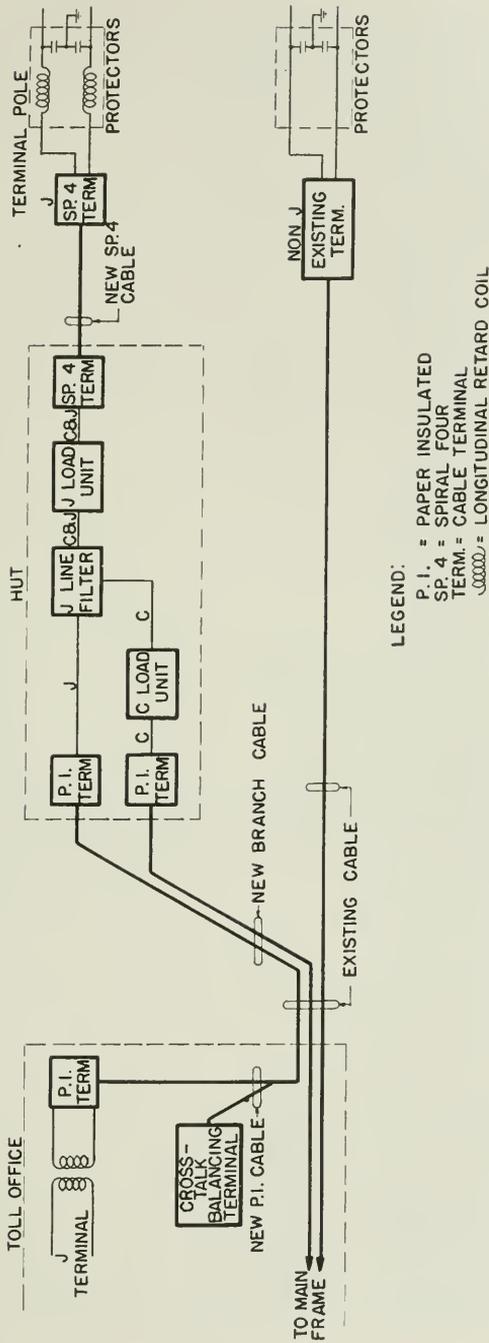


Fig. 5—Circuit connections through Dallas filter hut from open wire to terminal equipment at central office.

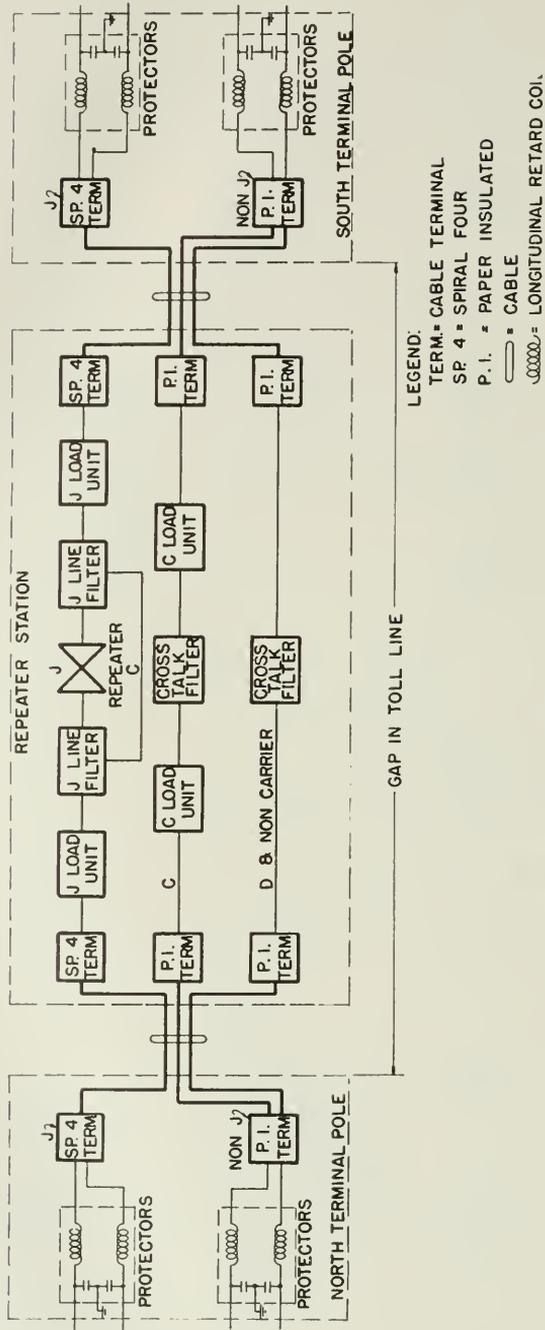


Fig. 6—Circuit connections through Temple unattended repeater station.

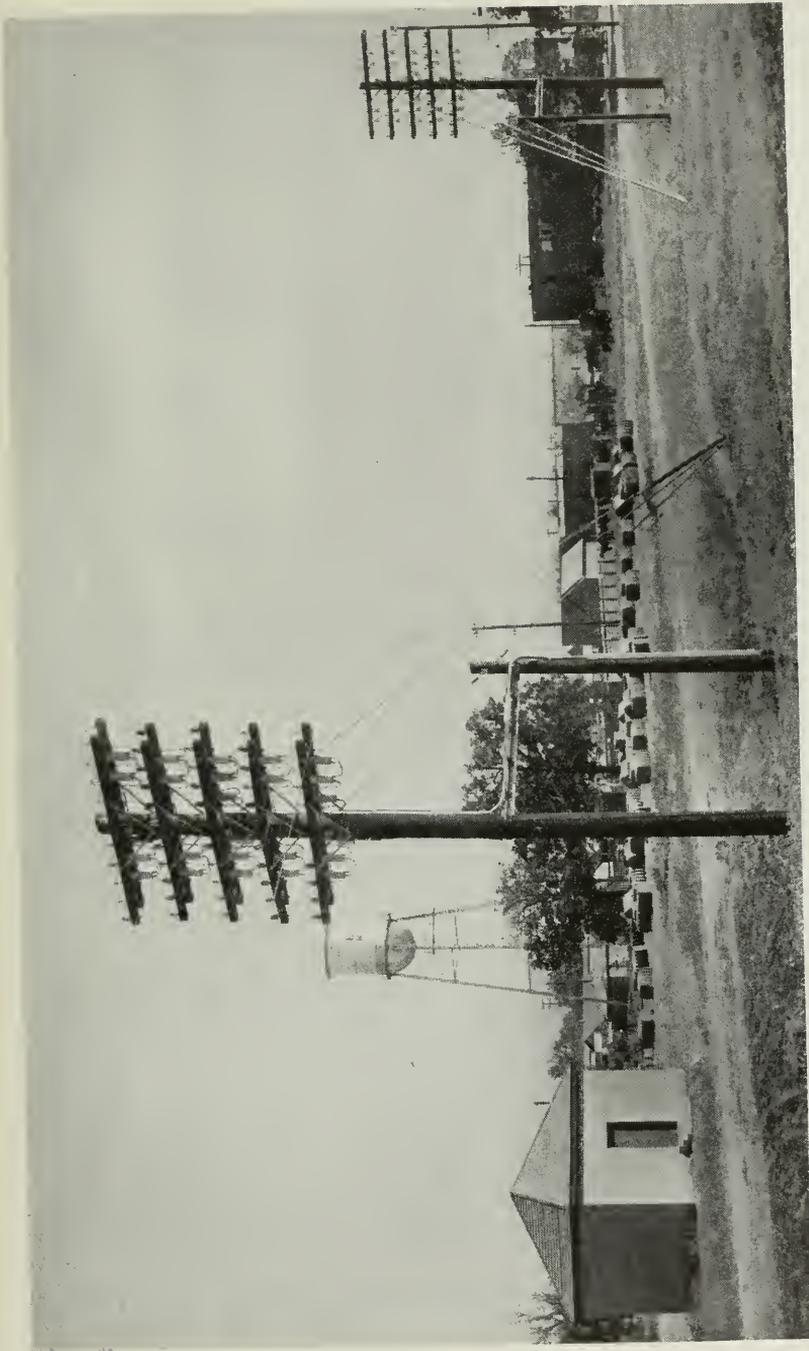


Fig. 7—Arrangement of gap in toll line at unattended repeater station.

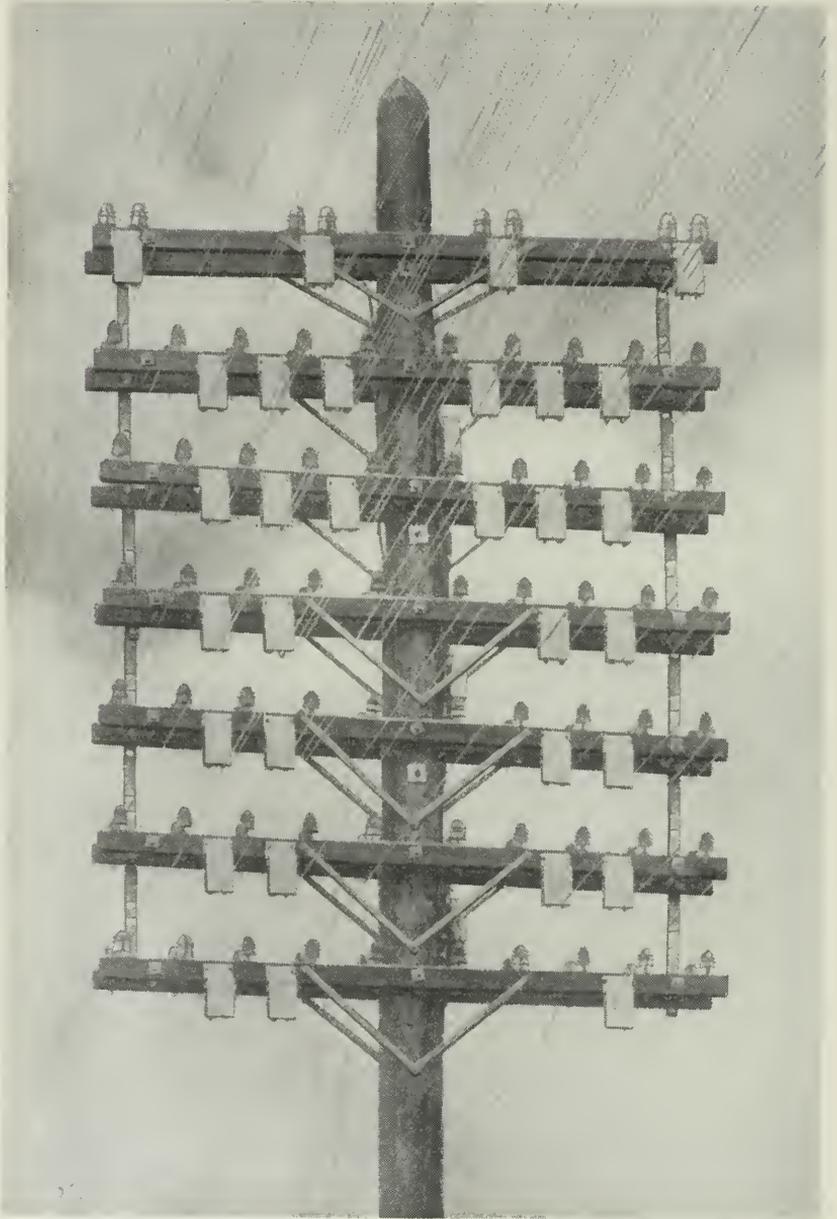


Fig. 8—Longitudinal choke coils and protectors on terminal pole at unattended repeater station.

office. This alarm has been arranged to operate by direct current over one conductor between offices without interfering with existing telephone circuits but at the expense of one DC telegraph path. For fuse failure, rectifier failure, power off, power restored, high-low voltage, high-low temperature, fire, burglary, pilot channel failure, and end of pilot channel control, alarms are sent and identified. A questionable alarm may be rechecked from the attended office.

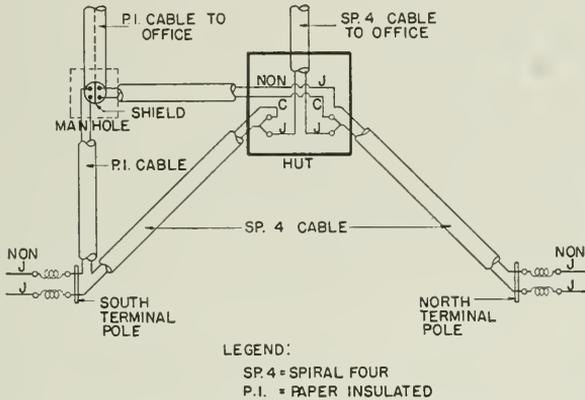


Fig. 9—Cable arrangement at Austin.

SPECIAL PROBLEMS

Some of the problems encountered in connection with the other two systems may be of interest. At Corsicana, a repeater point on the Dallas-Houston system, a filter hut was used on only one side of the repeater station. The situation which led to this arrangement is that an intermediate cable in the Dallas-Houston line extends 0.2 mile north and 0.5 mile south from the local central office. As it is necessary that the J system operate through this entrance cable and since space was available in the local central office building, repeater equipment similar to that installed in unattended buildings was placed in one room in the office.

The section of cable north of the central office terminates on a corner in a business district with all adjacent property occupied by buildings, making it more economical to use loaded spiral-four cable to this location than to extend the existing cable to an available site and provide the necessary filter hut and equipment. For the longer cable, it was more economical to provide the filter equipment in a hut in order to use existing facilities. Although this cable terminates in a fully developed residential area, a site for a filter hut was obtained adjacent to an alley in the rear of one of the residences facing the street on which the terminal pole is located.

The use of non-loaded paper insulated pairs in existing entrance cables has been mentioned. However, it is in general not practicable for crosstalk reasons to use all the non-loaded pairs which are available in one cable, and the selection of pairs suitable for type J operation is illustrated by a discussion of the methods used on the Dallas cable.

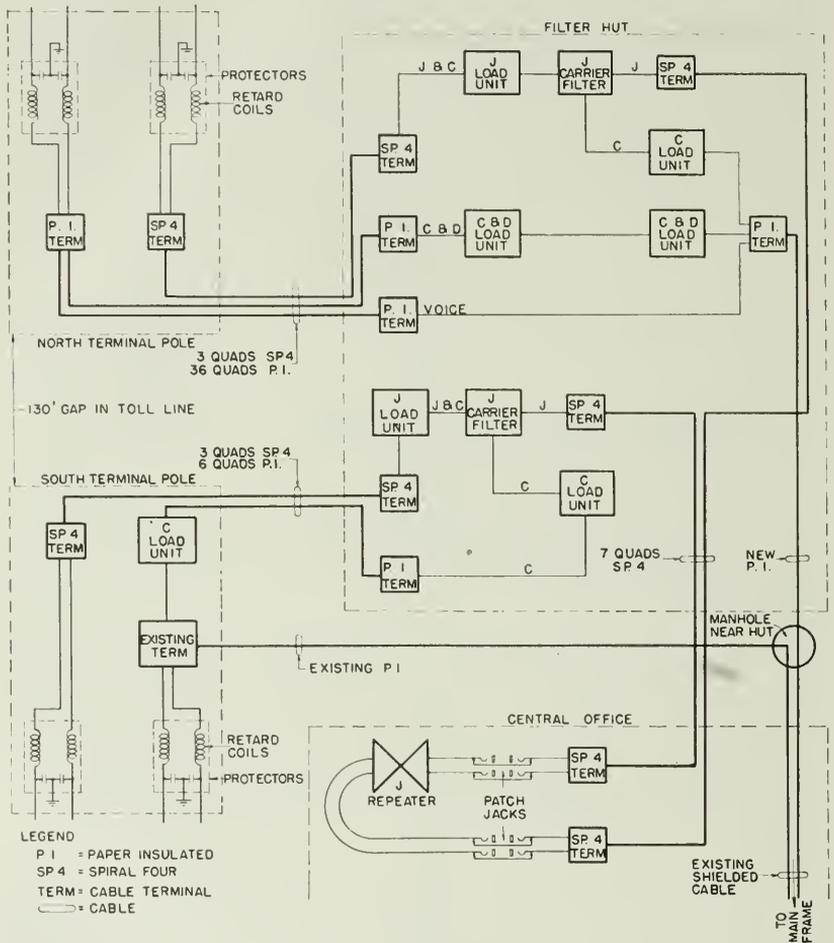


Fig. 10—Circuit connections through Austin.

The Dallas cable is composed of three sections of different make-up. The section nearest the central office, 1.3 miles long, and the intermediate section, 1.6 miles long, each contained 22 idle non-loaded pairs, and the third section, one mile long, had only six. The Houston line

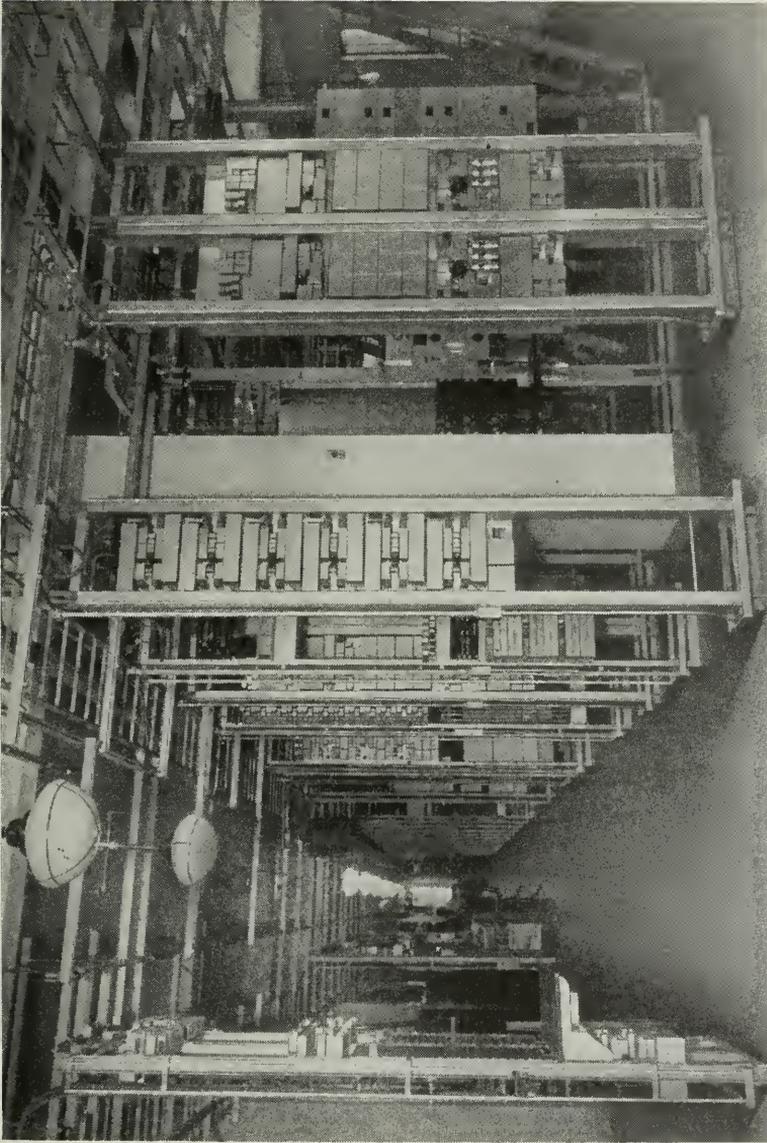


Fig. 11—Toll line terminating and testing equipment at Dallas with J carrier terminals in foreground.

terminates at the end of the second section, the third section extending the cable to the San Antonio line.

Since the number of cable pairs to the San Antonio line was limited to a maximum of six and since the rate of circuit growth over the two lines was expected to be approximately the same, requiring cable relief over the entire distance when the branch to the San Antonio line was exhausted, an objective of six pairs to each line was set up.

Measurements of crosstalk coupling at 140 kilocycles in terms of inductance and capacitance unbalance were made between each pair and all other pairs in each section and the pairs rated in their order of desirability. It is of interest that this required a total of 854 measurements. Those pair combinations whose coupling of the mutual inductance type was high were rated as the least desirable. This was done because capacitance balancing was to be used to obtain crosstalk reduction. The more desirable pairs in the first two sections were connected through to the six pairs in the last section by cut and try method until the overall condition was such that all six pairs were acceptable. By a similar procedure, using the remaining pairs in the first two sections, six pairs to the Houston line were made acceptable. No record is available as to the number of tests made in the cut and try process.

A cable terminal on which balancing condensers were mounted was installed in the central office building and connected to the selected pairs. This terminal contained sixty-six small adjustable wire wound condensers which were connected between each pair and every other pair. The condensers were adjusted to reduce to a minimum the capacitance component of the crosstalk coupling.

BUILDINGS

For the three J carrier systems, four new repeater stations and eight filter huts were needed. The same type of construction was used for all: Concrete foundation with floor slab above grade, double four-inch brick walls with rock wool insulation between but with solid brick at corners and openings, pitched roof with wood framing, fire resistant wall board ceiling, fire resistant composition shingles, and heat insulation above ceiling and below floor slab.

All of the racks for equipment in the unattended repeater stations are arranged in three rows with power, repeater, and line equipment in separate rows within a floor space of 17 feet by 16 feet which will allow the ultimate installation of six repeaters in each building. The entrance cables from the terminal poles enter from iron conduit through the floor and are racked and spliced on the side wall adjacent to the

line bays. The stubs from the cable terminals at the top of the line bays are carried overhead to splices on the wall. A ceiling height of 13 feet is maintained above the equipment but reduced along the pitch of the roof to 11 feet 8 inches at the side walls.

For all huts except that at Austin, three adjacent bays of racks are needed. With these along one side wall of the hut, the opposite side is available for splicing the entrance cable. At Austin an ultimate of



Fig. 12—Unattended repeater station.

nine racks, for filters in both directions of transmission, led to the use of racks along opposite sides of the hut with a splicing pit under the floor made accessible by trap doors in the floor between the lines of racks. In this case, the cable terminals are installed at the bottom of the racks with their stubs dropped directly through the floor slab into the splicing pit. The racks in the hut are seven feet high and a ceiling height of eight feet is used. Figures 12, 13, 14, and 15 are pictures of a typical repeater station, typical filter hut, and the special hut at Austin.

For correct operation of the equipment, temperature limits of 32 to 110 degrees Fahrenheit are desirable. Also, it is necessary that there be no precipitation of moisture on wiring or equipment. To maintain the desired conditions, each of the huts is equipped with a 2 kw. blower type electric heater arranged to operate at low temperature or high relative humidity, but with operation blocked when the temperature

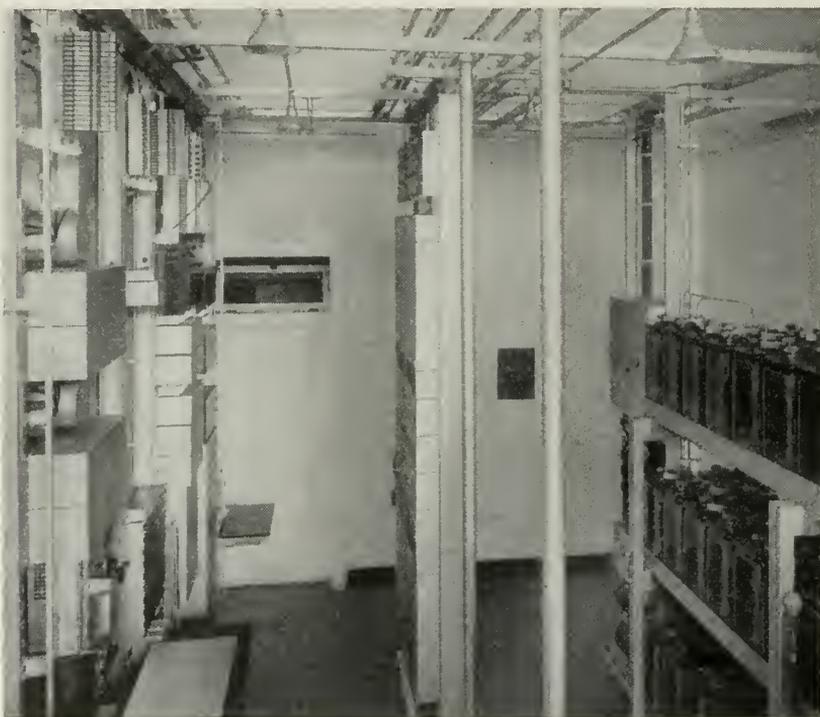


Fig. 13—Equipment in unattended repeater station.

reaches 95 degrees. Each new unattended repeater station is equipped with a 4 kw. heater similarly controlled, and, on account of the heat dissipation of power plant and vacuum tubes, also has forced ventilation which is operative under conditions of high temperature. The system of forced ventilation consists of spun glass intake filter, exhaust fan, electric solenoid controlled shutters at intake and exhaust, and thermostat, and is interconnected with the office alarms to prevent fan operation in case of fire.

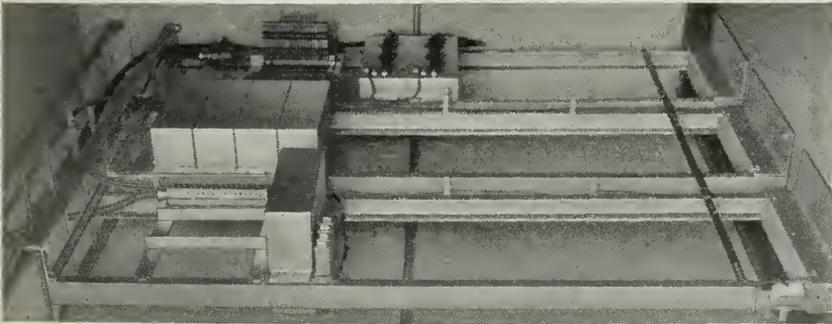
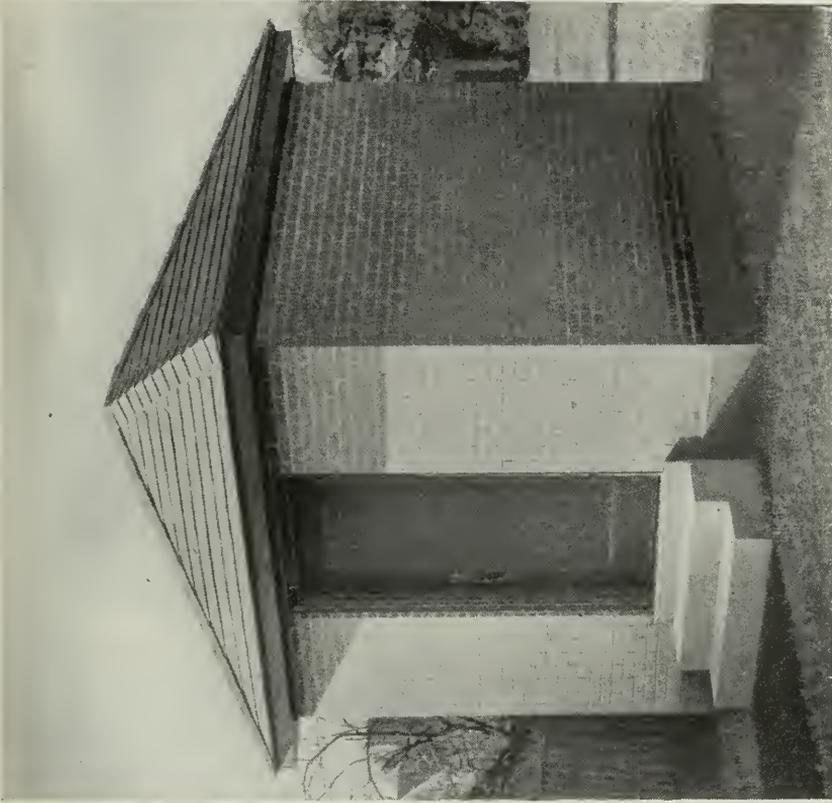


Fig. 14—Filter equipment and typical hut.

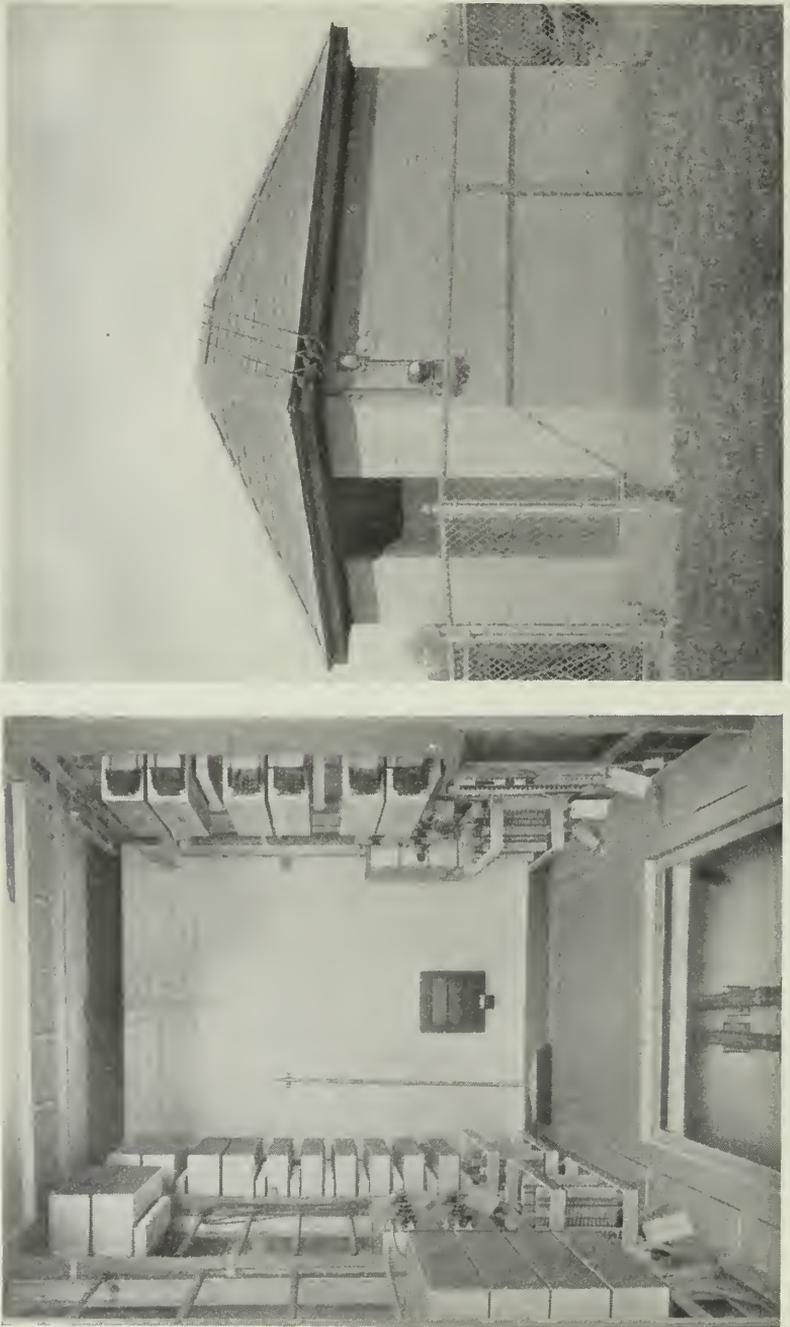


Fig. 15—Filter equipment and hut in Austin.

CONCLUSION

Upon completion of the buildings, equipment installation, and line facility rearrangements, adjustments in the equipment were made to match the lines used. Networks associated with the terminal and intermediate amplifiers were adjusted so that the amplification for any particular frequency would be equal to the attenuation at that frequency in the preceding repeater section; the automatic pilot channel equipment¹ compensates for attenuation changes. In repeater sections containing long toll entrance cables, it was necessary to sacrifice range of automatic pilot channel control to obtain the best equalization. However, satisfactory equalization and range of pilot channel control were obtained in every case.

As mentioned previously, the Dallas-Longview system operates on twelve-inch spaced phantom wire. In Fig. 16 the attenuation

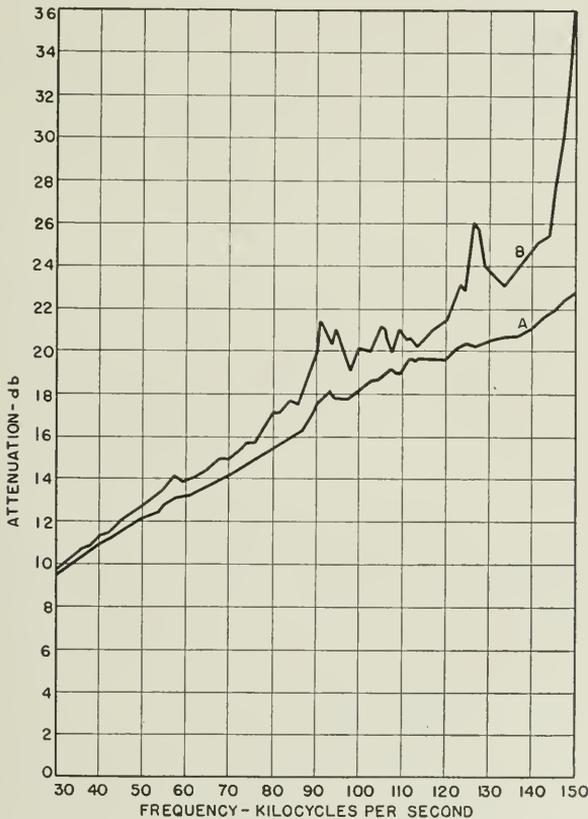


Fig. 16—Attenuation of two twelve-inch spaced phantom pairs of the Edgewood-Longview repeater section.

¹ Loc. cit.

characteristics of two possible pairs are shown. The absorption peaks of pair "A" at 92 and 127 kilocycles are within the frequency range of channels the fourth and twelfth of the J system and would impair the quality of those channels if pair "A" were used. Therefore, pair "B" is used as the regular path for the system. The quality of the channels obtained from these systems is shown by Fig. 17. Curve

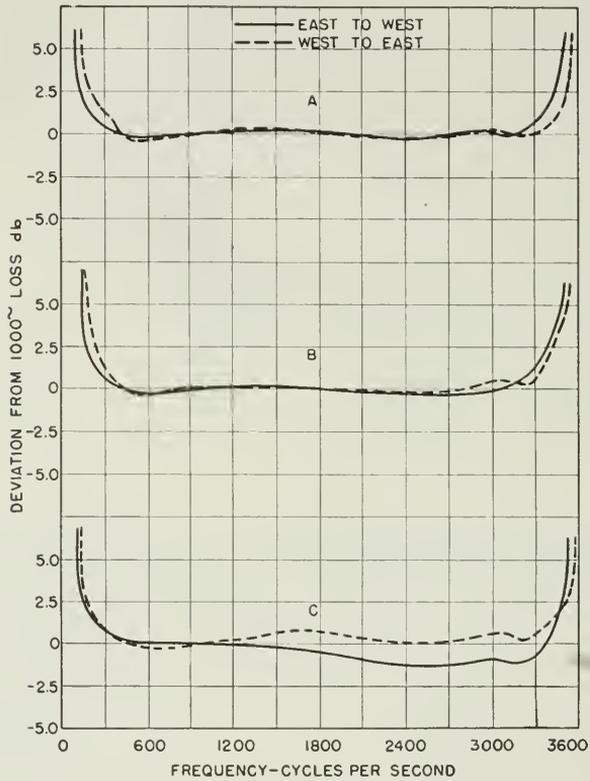


Fig. 17—Quality of derived circuits, "A" for typical channel of systems on six-inch spaced wire, "B" and "C" for the best and poorest channels of system on twelve-inch spaced wire.

"A" is representative of that obtained from a system operating over six-inch spaced wires; "B" and "C" are the best and poorest obtained from the Dallas-Longview system.

The Dallas-San Antonio, Dallas-Longview, and Dallas-Houston systems were placed in commercial service in September, October, and November, 1938, respectively. Experience with these systems is that the circuits obtained compare favorably with those obtained from any other facilities in quality and continuity of service, and that a definite need has been fulfilled in providing an economical method of increasing the capacity of the existing plant.

Line Problems in the Development of the Twelve-Channel Open-Wire Carrier System*

By L. M. ILGENFRITZ, R. N. HUNTER, and A. L. WHITMAN

The development of the type J twelve-channel carrier telephone system for open-wire lines required an increase of nearly 5 to 1 in the transmission frequency range of the lines. In the provision of suitable line facilities a number of new problems were encountered with respect to attenuation, noise and crosstalk. Methods for meeting these problems and the results obtained are described.

INTRODUCTION

A NEW carrier telephone system for open-wire telephone lines has been described recently.¹ This system increases the number of two-way telephone circuits which can be obtained on a single pair of wires from the previous maximum of 4 to a total of 16. This has been achieved by extending the frequency range from a maximum of about 30 kilocycles to more than 140 kilocycles. The exploitation of this new range of frequencies on open wire has involved the solution of a number of interesting problems, among which are these:

(1) Not only does the attenuation of an open-wire line under ordinary weather conditions rise substantially with frequency but extremely large increases in attenuation occur at the higher frequencies when ice forms on the wires.^{2, 3} In spite of these effects a high degree of stability of transmission has been secured on all channels by the provision of automatic control of repeater gain and equalization.

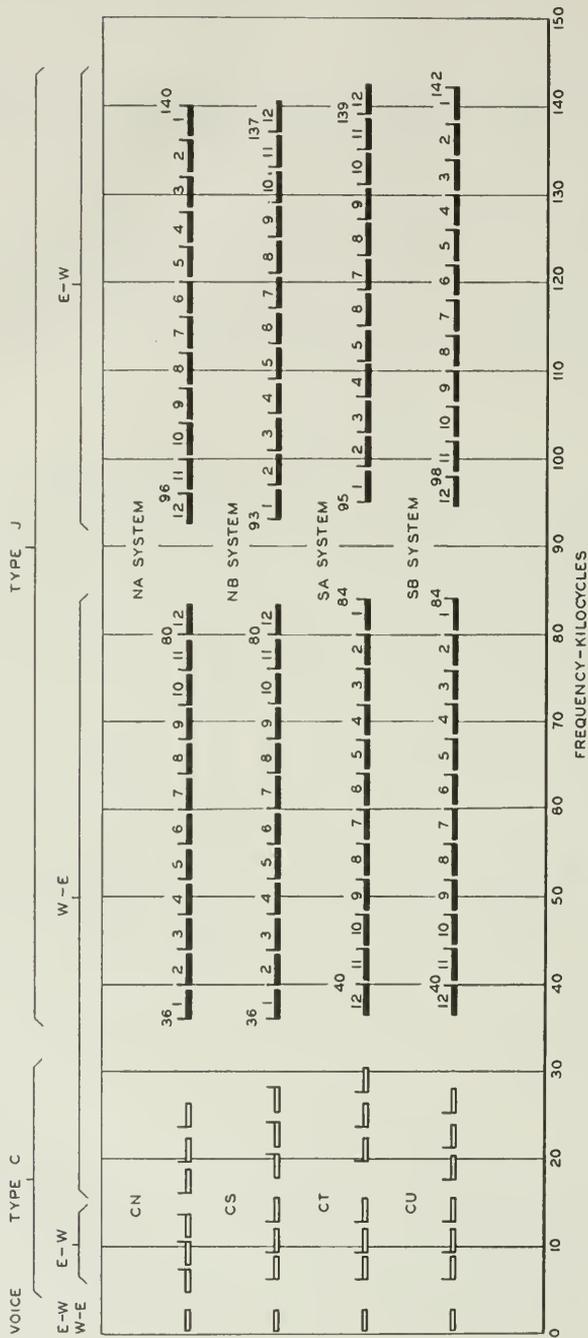
(2) New crosstalk problems created by the extension of the frequency range have been solved by the development of transposition designs with numbers of transpositions not greatly in excess of those employed for the lower frequency systems. Problems have also arisen in controlling the crosstalk around the repeaters and in reducing the effect of impedance departures between the line circuits and the equipment.

FREQUENCY ALLOCATIONS

The type J system operates on circuits on which type C carrier systems were already operating in the frequency range up to about 30 kilocycles. To provide enough frequency separation between the two

* Presented April 18, 1939 before the A. I. E. E., in Houston, Texas.

¹ Reference numbers refer to the list of references appearing at the end of the article.



NOTE: E-W ALSO IMPLIES TRANSMISSION N-S AND W-E IMPLIES S-N

Fig. 1—Frequency allocation.

systems the lower frequency limit of the J system was set at 36 kilocycles; the necessary frequency space for 12 channels in each direction set the upper limit at about 140 kilocycles. This range is split into two parts, one used for transmission in one direction and the other for the opposite direction. Figure 1 illustrates the relation of the frequency bands occupied by the type J and type C systems and the voice-frequency channel. Different "staggered" locations of the frequency bands are to be employed in order to simplify crosstalk problems.

Filters are used for separation of the type J from the type C and lower frequency facilities on the same pair of wires. This separation is done by means of a combination of high and low pass filters which split apart the frequency ranges above and below the band between 30 and 36 kilocycles. To simplify the design of these filters, the low frequency group of the type J system is transmitted in the same direction as the high frequency group of the type C system. This arrangement of transmitting certain frequencies in a particular direction is generally used throughout the telephone plant in order to avoid serious crosstalk difficulties. Accordingly, with few exceptions, west to east transmission or south to north transmission takes place in the same frequency bands throughout the country and similarly, east to west or north to south transmission employs the same frequency bands. These are indicated in Fig. 1.

LINE ATTENUATION

An open-wire pair affords the lowest loss transmission medium of any conductor employed in the telephone plant. It is, however, peculiarly subject to the effect of weather, which may cause large and often rapid changes in the attenuation. In consequence, some form of gain regulation is required.

Even for carrier systems operating up to 30 kilocycles, manual regulation is inadequate for the longer systems and automatic devices have been provided for most systems over 500 miles in length. The attenuation changes caused by changes in resistance of the wire with temperature or by changes in the shunt losses when insulators become wet are much larger at the higher frequencies of the J system, and therefore, an automatic regulating scheme is required. Tests were made on open-wire circuits to determine more precisely the characteristics needed for such a regulator. During sleet storms, when wires are covered with ice, the increases in attenuation are far beyond any caused by rain. Figure 2 shows increases which may be caused by ice as compared with the normal dry and wet weather values.

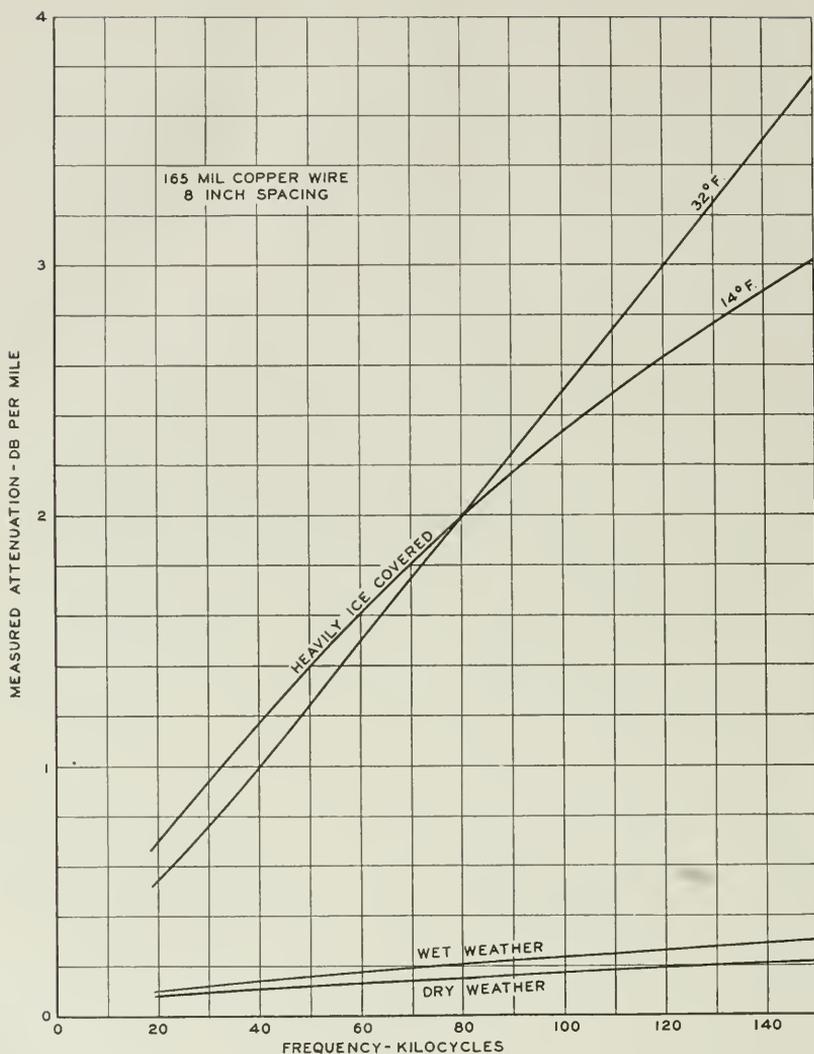


Fig. 2—Attenuation variation with weather.

The deposits on the wire may be actual ice, or in some cases wet snow or frost adhering to the wire. Figure 3 shows an example of such deposits. Theory shows that the increase in attenuation is caused by energy losses in the ice itself and that leakage across the insulators is usually a negligible factor.

An extensive survey of the effects of ice has been carried on at various points throughout the country during the past four years and a

large amount of information has been accumulated. These tests have shown that the shape of the attenuation-frequency characteristic differs considerably for different ice formations and even if the ice deposit remains the same for a time, the attenuation-frequency characteristic may vary with temperature as in Fig. 2. The two upper curves of the figure were measured at different times during the same storm. There

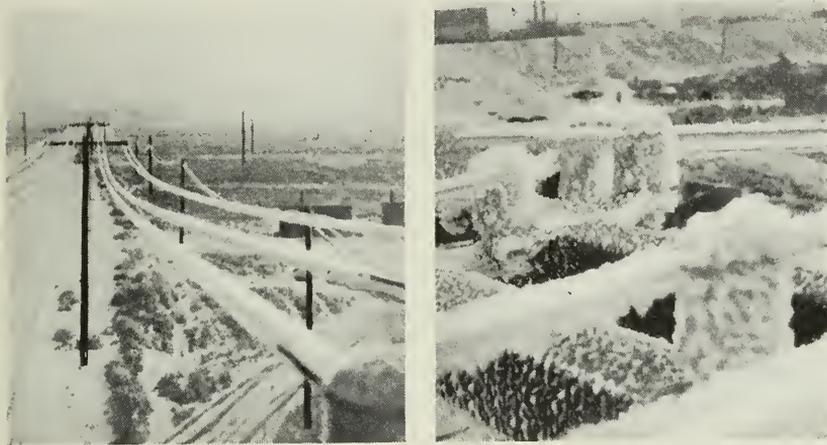


Fig. 3—Ice on wires and insulators near Amarillo, Texas.

was no apparent change in deposit between the two measurements. This change in shape of the characteristic, of course, makes the regulation problem more difficult. In spite of the extreme severity of ice effects in certain regions, it is expected that satisfactory reliability will be obtained on type J systems by placing the repeaters sufficiently close together.

REGULATION PROBLEM

In the first type J systems the regulator, actuated by a single pilot frequency in each direction, compensates for the attenuation changes caused by temperature and wet weather.

The required varieties of attenuation slopes with ice on the wires could not be provided by a simple regulator. Hence provision is to be made in later designs for a regulator with variable slope controlled by two pilot frequencies which is expected to be satisfactory in areas subjected to sleet conditions. The regulating range will also be increased so that a completely automatic control of gain up to about 75 db will be available.

It was found that during periods when ice coated the wires the

circuit noise measured at the end of a repeater section usually decreased as the attenuation increased. This is important because otherwise the extra increase in the repeater gain to take care of the higher attenuation at such times would make the noise excessive. The study of ice conditions throughout the country which has been carried on and is still continuing will be useful in laying out repeater stations along some of the routes which eventually will be candidates for the application of type J systems.

OPEN-WIRE CROSSTALK⁴

The crosstalk problem on open-wire lines is one of the most important. Crosstalk is controlled by transpositions which are introduced into the various pairs in accordance with a predetermined design. The creation of the necessary designs requires consideration both of the complex theory of transpositions and measurements on lines constructed by practical methods.

However, the design of transposition systems is considerably simplified by the use of different frequencies for the two directions of transmission. The only crosstalk between systems which is directly important is that known as far-end crosstalk, which is that between a talker at one end of one circuit and a listener at the opposite or far end of another. Near-end crosstalk, which is that between a talker and a listener at the same or near ends of two circuits, becomes a source of interference between circuits only when portions of it appear as far-end crosstalk because of reflections at points of impedance irregularity in the circuits.

Because of the high cost of a transposition design to keep both near-end and far-end crosstalk down to small values, only small reflections are permitted where open-wire and cable meet, or where circuits are terminated in equipment. A number of the difficulties which had to be overcome to attain small reflections are discussed later in the paper. With this control the transposition designer can concentrate most of his attention on far-end crosstalk, the near-end crosstalk requirements are relaxed, and a cheaper transposition arrangement can be used.

What can happen when reflection occurs may be seen by comparison of the near-end and far-end crosstalk curves in Fig. 4. The similarity in the shapes of the two curves, and particularly the fact that the peaks occur at the same frequencies, show that what appears to be far-end crosstalk is in this case mostly reflected near-end crosstalk. It is for pair combinations such as this one, where the near-end crosstalk is much larger than the far-end, that the closest control of reflection effects is required. With the values of reflection realized in the J system, reflected crosstalk will ordinarily be unimportant.

To obtain satisfactory crosstalk conditions at the higher frequencies some changes in line construction are necessary. To use type J carrier systems on existing open-wire routes, methods were devised for modifying the line construction in as economical a manner as possible.

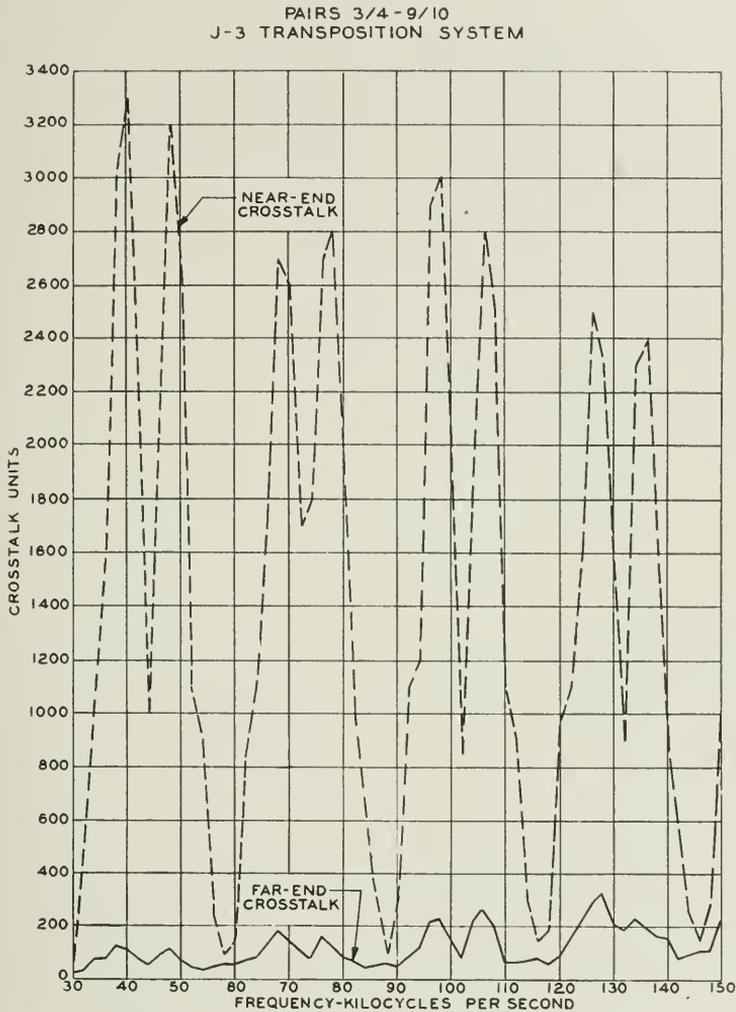


Fig. 4—Near-end and far-end crosstalk—J-3 transposition system.

For new lines, such as the new part of the Fourth Transcontinental line,⁵ advantage was taken of the greater degree of freedom in structural design which was possible.

Figure 5 shows three types of open-wire pole head configuration

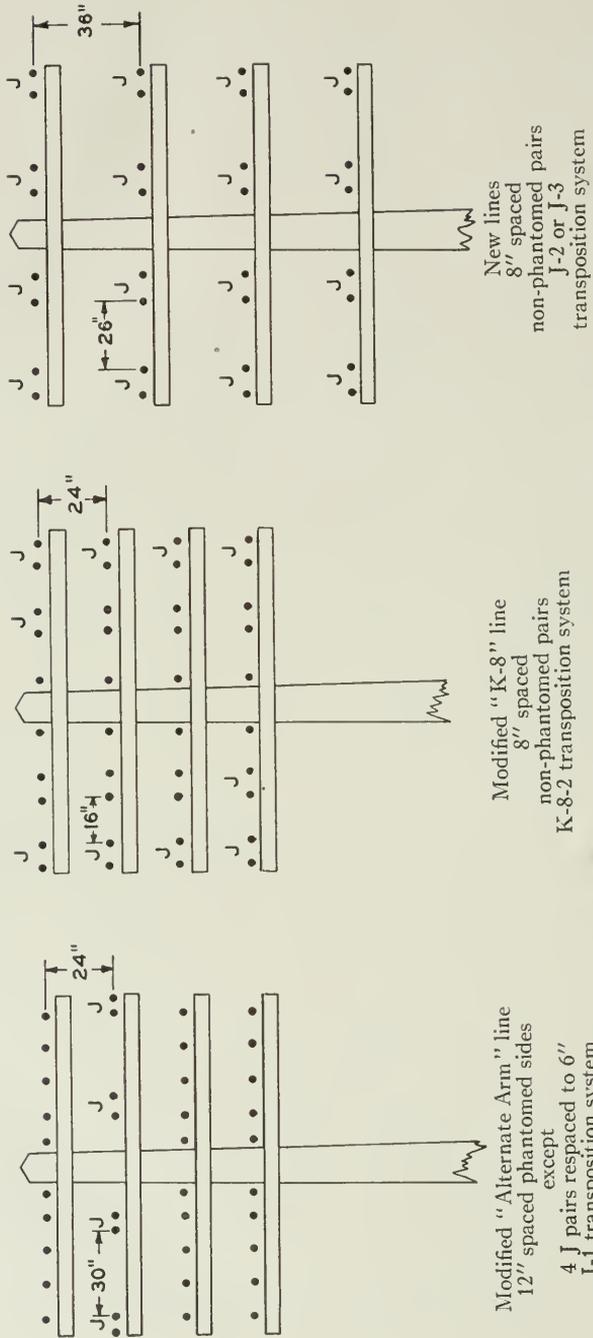


Fig. 5—Three types of open-wire pole head configuration.

suitable for J system operation. The left-hand diagram shows a method of reconstructing part of one of the older types of open-wire lines built with 12-inch spacing between wires of the pairs and with the "Alternate Arm" transposition system which was developed for the use of type C systems on the side circuits of the horizontal phantom groups on alternate arms. This method is a flexible one in that one or more phantom groups may be converted at a time, as on the second crossarm shown. For such an application not only was removal of the phantoms and retransposition necessary, but the spacing of the two wires of each pair was reduced to 6 inches. This general method of construction was used for the Dallas-Houston and Dallas-San Antonio lines,⁶ except that the 6-inch pairs were constructed with new wire on a new crossarm rather than by respacing 12-inch pairs.

Another common type of open-wire pole head configuration, the middle diagram of Fig. 5, is that made up of 8-inch spaced non-phantomed pairs transposed in accordance with the K-8 transposition system on an eight-span base. Through design studies supplemented with field experiments it was found that such a line could be converted for J systems much more cheaply than an Alternate Arm line. If J systems are restricted to the pairs on the outer ends of the crossarms, with two inner pairs, about one or two transposition changes in each pair per mile are enough. This scheme was followed in reconstructing the line between Charlotte, North Carolina, and West Palm Beach, Florida.

For new lines yet to be built, a greater degree of latitude in structural design is naturally possible. The right-hand diagram of Fig. 5 shows an open-wire pole head configuration designed to allow J systems to be operated on all of the pairs. The unique feature of this configuration is that, while 8-inch spacing is preserved between the wires of the various pairs, the adjacent non-pole pairs on a crossarm are separated by twenty-six inches and the crossarms by thirty-six inches. The reduction in coupling made possible by this increased spacing keeps the crosstalk for any combination of pairs down to a suitable value with transposition arrangements not necessarily more complicated than those employed for the other configurations. This type of construction was used for the new parts of the Fourth Transcontinental line.

Fig. 6 shows a comparison of the number of transpositions used in a typical section of open-wire line for various types of circuits from voice frequency phantom circuits to non-phantomed circuits intended for J system operation. From the original arrangement where there was one transposition point in every ten spans, about

$\frac{1}{4}$ mile, the number of transpositions for J carrier operation has been increased so that for the J-3 design, which was used for the new wires on the Fourth Transcontinental line, there are four transpositions in each eight-span interval and every pole is a potential transposition point.

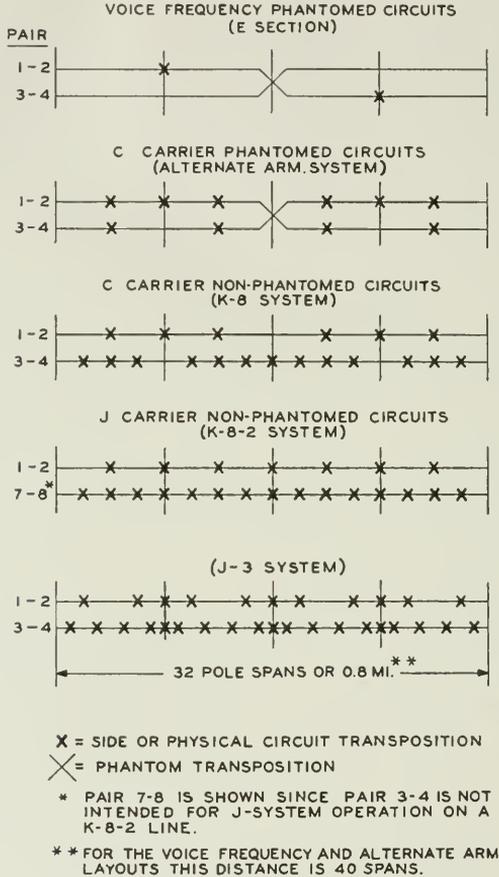


Fig. 6—Illustrative transposition arrangements.

It may be seen from Fig. 6, however, that the number of transpositions required in pairs for J carrier operation is not necessarily larger than the number employed in systems intended for C carrier operation with a top frequency of 30 kilocycles. The superiority of the J system transposition arrangements as compared with those designed for C system operation results from the choice of specific arrangements which best limit the systematic effects for frequencies in the J system range.

Typical far-end crosstalk measured between 8-inch spaced pairs 11/12 and 19/20 on a new J-3 line and on a reconstructed K-8-2 line is shown by Fig. 7. The superiority of the new line with its fewer wires, greater

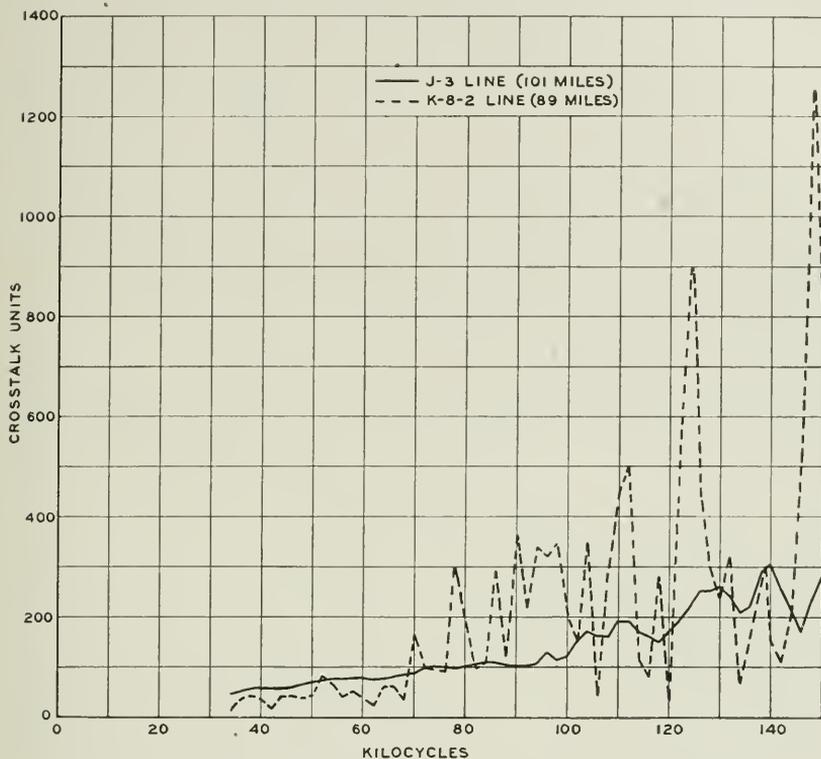


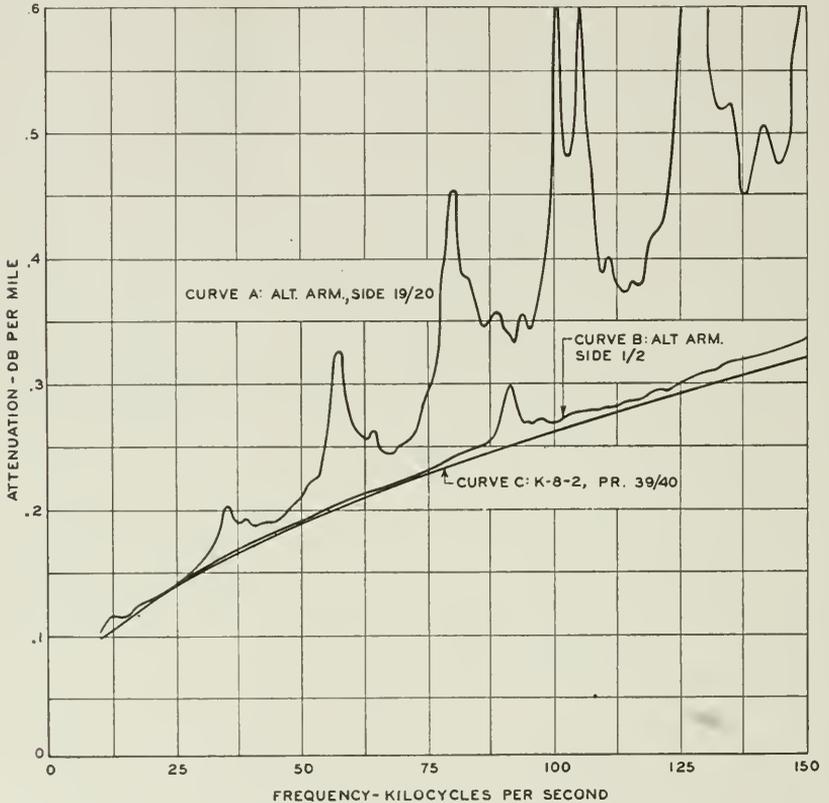
Fig. 7—Far-end crosstalk between 8-inch spaced pairs 11/12 and 19/20.

pair separations, better transposition system and smaller irregularities is evident.

ABSORPTION EFFECTS

The attenuation of an open-wire pair may be quite unsatisfactory if there are what are known as absorption effects, caused by induction into surrounding circuits such that energy is absorbed in particular frequency bands and the attenuation of the pair increased. These effects, which depend on the transposition arrangements in the circuits, may cause objectionable transmission distortion at critical frequencies unless the transpositions are planned to avoid them. The same arrangements necessary to control crosstalk between J systems will automatically eliminate absorption effects with one exception. If

only part of the pairs on a line are designated and transposed for J systems and the remaining pairs are not so transposed, absorption in a J pair can be caused by a nearby non-J pair. Consequently, consideration of the crosstalk relations at J frequencies between all of the pairs



Curves A and B: Side 19/20 and 1/2 respectively, on Alternate Arm line, 104 mil, 12-inch, 56.7 miles, 90° F. at Mascoutah, Ill.

Curve A is transposed for voice frequencies.

Curve B is transposed for carrier operation up to 30 kc.

Curve C is pair 39/40 on K-8-2 line, 104 mil, 8-inch, 68 miles, 50° F. and CS insulation between Denmark, S. C. and Rincon, Ga. Transposed for carrier operation up to 140 kc.

Fig. 8—Attenuation of open-wire pairs of different types.

on the line cannot be avoided even though some of them will not be used for J systems.

Figure 8 illustrates the effect of absorption on three different pairs. Curves A and B show the absorption measured over the type J frequency range on a line of the Alternate Arm type. Curve A was

obtained on a side circuit transposed for operation at frequencies only up to about 10 kilocycles. The absorption at frequencies above this becomes very large. Curve *B* shows the absorption present on one of the C carrier side circuits on the same line transposed for operation up to 30 kilocycles. Curve *C* shows how absorption disappears on a non-phantomed pair specially transposed for type J operation. If this pair were measured at much higher frequencies, similar absorption "bumps" would be found, perhaps at frequencies of 200–300 kilocycles or higher.

Since absorption effects depend on the systematic addition of crosstalk currents along a line, a continuous succession of identical transposition sections tends toward greater absorption while a random succession of different kinds of transposition sections of different lengths will reduce it. The Dallas-Longview J system is operating on an Alternate Arm side circuit, transposed for C carrier operation and without any modifications to adapt it for the higher frequencies. Because of the fortunately irregular succession of different transposition sections found here, it was possible to select, after tests, a pair with no serious absorption.

CONSTRUCTION IRREGULARITIES

With the new transposition designs, the systematic crosstalk resulting from the transposition arrangements has been reduced in nearly every case so far that the remaining crosstalk is controlled principally by construction irregularities. An important source of irregularity is the difference in sags of the various wires in each span of the line, particularly sag differences between the two wires of each pair. Another potentially important source of irregularity is the variation in the spacings between successive transposition poles. It is relatively easy to make this factor unimportant as compared with sag differences.

The large amount by which the crosstalk can be reduced by careful methods of construction coupled with the highly developed systematic transposition patterns is illustrated by the fact that between certain pairs the crosstalk in a 75-mile repeater section is reduced to a value which would be produced by a capacitance unbalance between them of less than 2 mmf, which is about the same in magnitude as the capacitance between wires of a foot of the open-wire pair. This large crosstalk reduction is in spite of the fact that at 140 kilocycles the phase change along an open-wire circuit is about 7° in a single span, the shortest distance between any two transpositions, and about 28° for the more common four-span interval.

INTERACTION CROSSTALK AT REPEATER POINTS

Another type of problem was introduced by what is known as interaction crosstalk. This is the crosstalk which occurs from one side to the other of a J repeater station. Figure 9 illustrates two paths

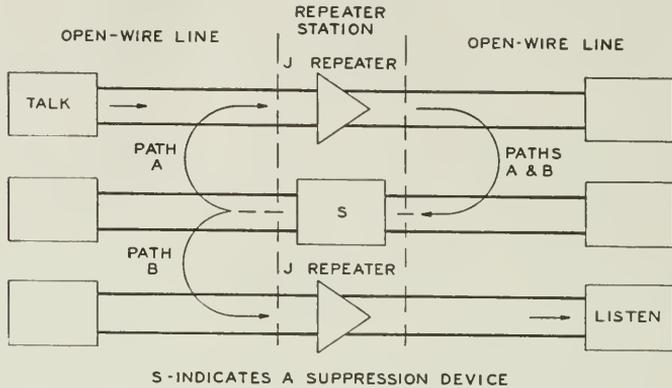


Fig. 9—Interaction crosstalk at a J repeater station.

which it may take. Path A shows the crosstalk from a system to itself which may cause transmission distortion or repeater singing while Path B is the path of crosstalk between different circuits. The essential feature of this interaction crosstalk is that, as Fig. 9 shows, the crosstalk path at a repeater station passes through the J repeater and hence the crosstalk is amplified by the repeater gain.

The new problems of controlling this crosstalk were the result of larger magnitudes of crosstalk at the higher frequencies, the larger repeater gains and the fact that with more repeaters there were more points on a system where it could occur. Magnitudes of interaction crosstalk which had previously been thought of as inconsequential assumed a new importance. For instance, with the gain of about 75 db proposed for the repeater for use in sleet areas, an initial value of unamplified interaction crosstalk as low as 0.25 crosstalk unit would be magnified to 1400 units, which might considerably exceed the far-end crosstalk existing at the same time in one repeater section.

Several new methods for reducing this interaction crosstalk were devised. In the first place, in order to prevent direct coupling between the wires of the open-wire line on the two sides of the station, it was found necessary to cut a gap in the line. With the wires entirely removed for a distance usually of about eighty feet, the line is brought into the station from the two terminal poles by means of the lead-in cables.

It was also seen to be necessary to block the paths provided by the wires of the telephone line itself. For this purpose, crosstalk suppression filters were designed and built to be installed in all of the non-J circuits on the line. These give losses of the order of 70 db at 140 kilocycles not only in the metallic transmission circuits but also in other circuits, made up of various combinations of the line wires, which may conduct crosstalk currents through the stations.

In addition to the crosstalk suppression filters and in order to provide an extra margin of safety against interaction crosstalk currents which might find their way through the repeater station by stray paths, longitudinal choke coils have been connected at the pole heads between the open wires and the lead-in cables. These coils do not disturb ordinary transmission but add high impedance in the longitudinal circuits.

These measures for controlling interaction crosstalk have been found to be adequate so far as the telephone line is concerned. At an occasional J repeater station, however, located on a right-of-way occupied by several pole lines, there is found another pole line paralleling the telephone line with a separation sometimes as little as 2 to 5 feet between the nearest wires of the two lines. Such wires provide other interaction crosstalk paths past the repeater station and impair the effectiveness of suppression measures installed in the line on which the J system is operated. The by-passing effects of such a foreign line can be controlled by crosstalk suppression devices similar to those used in the telephone line wires.

Figure 10 shows a comparison of the interaction crosstalk measured at a J repeater station before any suppression measures were installed, the other wires of the line being continuous at the station location, with the corresponding interaction crosstalk when the line was run through the suppression devices in the station. The values shown would be amplified by the gain of the J repeater on the disturbed circuit before they reached the listener. The effect of the by-passing foreign line is illustrated by the difference between the middle and bottom curves, the bottom curve showing the measured crosstalk when the by-passing line was cut to simulate the effect of suppression measures in it.

STAGGERED SYSTEMS

It would not be possible with the open-wire line configurations now in use to design transposition arrangements that would permit the operation of identical J systems on all pairs. For this reason four types of J systems with different channel carrier frequency allocations

will be provided in the future. The frequency assignments for these systems are shown in Fig. 1.

The "staggering" advantage, or effective crosstalk reduction between systems, is effected because (1) the inversion or displacement of channels in the different systems with respect to each other makes the

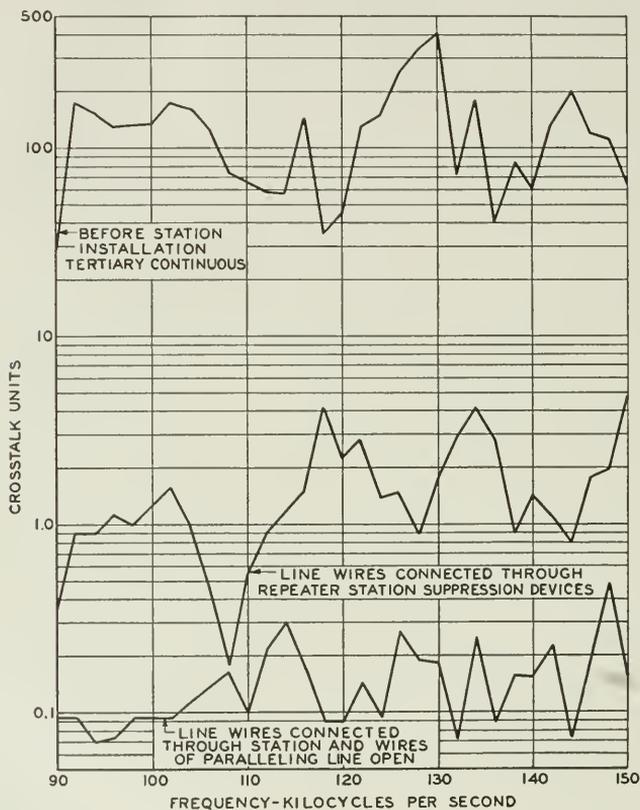


Fig. 10—Unamplified interaction crosstalk between two J circuits at an auxiliary repeater station.

crosstalk unintelligible and (2) the reduction of the overlap between channels results in less energy being transferred between them by crosstalk. The net benefits of "staggering" obtained by the allocations shown in Fig. 1 range from about 6 to 16 db.

The most effective pair assignments for the four types of J systems can best be obtained from actual crosstalk data on the particular sections of line involved. The "staggering" advantages obtained are sufficient so that the highest remaining crosstalk will usually occur between the like J systems operating on non-adjacent pairs.

NOISE

Observed external sources of noise in J systems are atmospheric static, dust storms, radio stations, power line carrier and power supply systems.

Of these possible sources the more important will usually be atmospheric static which will be greatest during the summer months. In regions where dust storms occur, their effects are expected to exceed that of atmospheric static but will be more likely to occur during the winter and early spring.

The following table shows values of noise at 140 kilocycles, caused by atmospheric static, found at the open-wire line terminals of one repeater section; the values are those which it is expected will be exceeded during one per cent of the summer season extending from May to September. If the repeater spacings shown were used, the total static noise in the top channel at the end of a circuit with 20 repeaters would be 20 db above reference noise at the - 9 db level. However, other factors such as ice may require the use of shorter spacings.

Transposition System	Wire Spacing (Inches)	Noise (db) *	Repeater Spacing in Miles—128-Mil Wire
Alternate Arm	12	+ 10	67
K-8-2	8	+ 5	82
J-1	6	- 2	103

* Above reference noise, 10^{-12} watt at 1000 cycles.

LINE IMPEDANCE

As mentioned previously in the discussion of crosstalk, it is important that the line impedances be matched closely and large irregularities be avoided. Because of the different wire sizes and pair spacings, a wide range of open-wire line impedances may be encountered. Novel construction arrangements and the development of new lead-in circuits have made it possible to secure a reflection coefficient of about five per cent at the junction between the open-wire pair and the toll entrance and office equipment at the highest transmitted frequency.

The transposition arrangement and wire spacing of a pair affect the smoothness of its impedance because they affect the reactions between circuits which cause absorption effects. The marked improvement which can be obtained by proper design is illustrated by comparison of Curves A and B of Fig. 11. Curve A shows the impedance of a

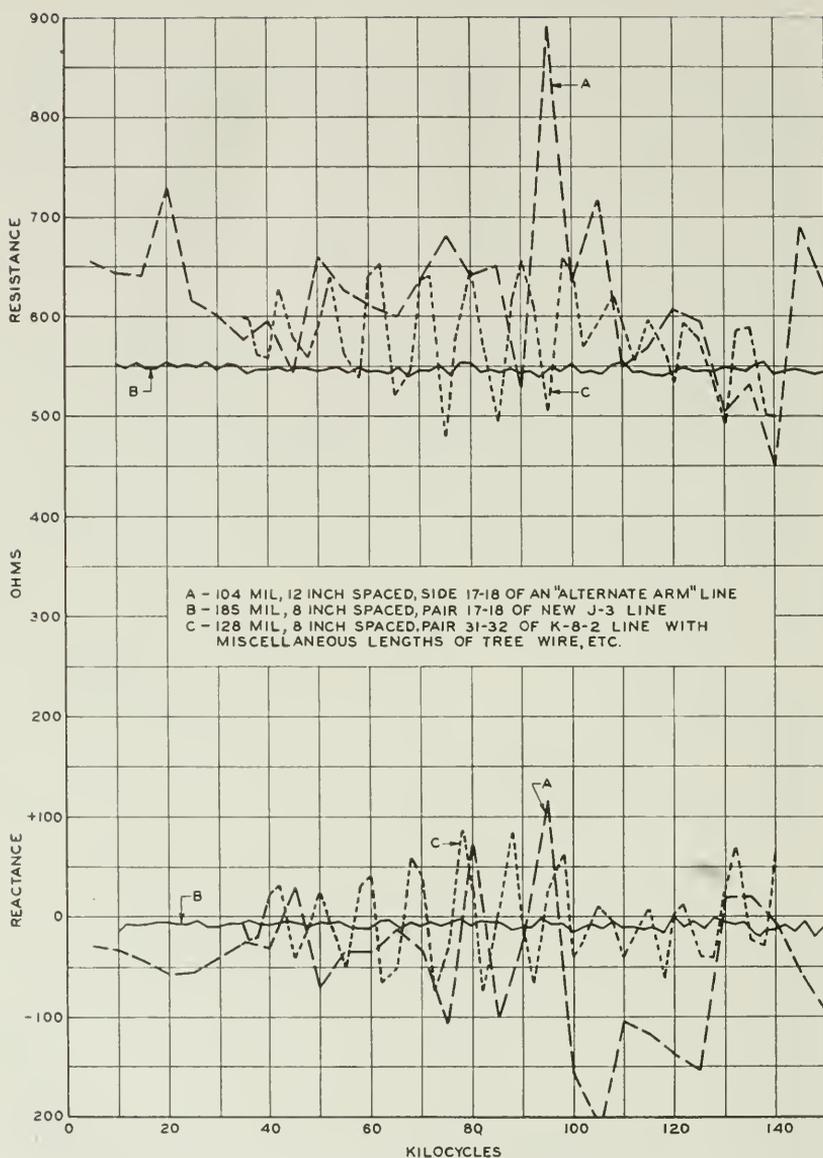


Fig. 11—Impedances of open-wire pairs of different types.

12-inch spaced side circuit on an Alternate Arm line. This particular circuit was one intended for use at frequencies not above 10 kilocycles. In striking contrast Curve B shows the comparatively smooth impedance of an 8-inch spaced non-phantomed pair on a new line transposed in accordance with the J-3 system,

“Tree” wire, a special line wire with abrasion-resistant insulation, has been used on open-wire lines for many years in places where the lines were exposed to tree branches. During line tests in Florida, another use for tree wire was found where the open-wire line, along a causeway or bridge, is subject to fouling by fishing tackle. Curve C of Fig. 11 shows what a half-mile or so of this tree wire, supplemented by

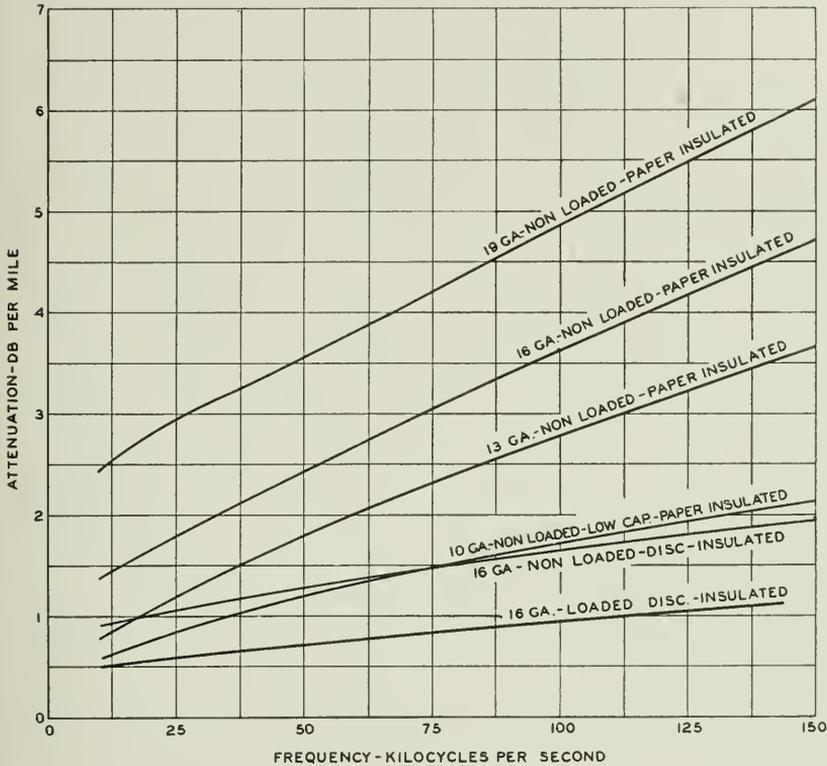


Fig. 12—Attenuation of toll entrance cable pairs.

several sections of 165-mil wire at railroad and power-line crossings, can do to the impedance of a 128-mil pair. To reduce the irregularities a new type of insulated line wire of smaller diameter and with thinner insulation was developed. This wire has about the same impedance characteristic as the line wire.

INTERMEDIATE CABLE TREATMENT

When open-wire lines have to be placed underground to pass through towns or to cross natural barriers such as rivers which cannot

be spanned economically with open wire, cable is used. In the past, the circuits in such cables were frequently loaded to reduce their attenuation and to match the impedance of the open-wire circuits in order to avoid reflection effects and degradation of voice-frequency repeater balance. To load paper-insulated cable pairs for frequencies up to 150 kilocycles would require exceedingly short loading spacing, of the order of 200 feet, which would be expensive and in many cases impractical with existing manhole locations. An alternative, the use of a transformer to match the open wire and cable impedances, was rejected as it was found impractical to design a transformer which would be adequate over the entire frequency range.

To overcome these difficulties, a new low-capacitance type of cable was developed which could be loaded to match the open-wire impedance with coil spacings about the same as those previously used. Loading coils of different sizes were developed to provide for loading to the different impedances of the open-wire circuits.

The new cable employs 16-gauge conductors in a spiral-four arrangement, supported by hard rubber disc spacers about 0.6 inch in diameter. These are surrounded by copper and iron tapes for shielding and strengthening purposes. The units so formed may be assembled either in single units in a lead sheath as for lead-in purposes, or in multiple units, up to a maximum of seven for full-sized cable, within the same lead sheath. For duct runs or submarine cables, the multiple assembly is usually employed, and, in the latter case, with outside armoring and jute protection. If the submarine span is more than about 600 feet, intermediate submarine loading is employed.

As an alternative, it sometimes happens that where a long intermediate cable is involved, an auxiliary type J repeater station can be placed conveniently at one end of this cable. In this case, the filter hut described in the discussion of toll entrance arrangements in the next section may be used at the end of the cable opposite the repeater station and the cable treated as a toll entrance cable for the auxiliary office. A further alternative is to provide filter huts at the two ends of a non-loaded intermediate cable. However, if the cable is short, the new disc-insulated cable with loading is to be preferred.

Previous practice at the ends of open-wire lines has been to use paired bridle wire with weather-proof insulation and usually of smaller gauge than the line wire to connect the open-wire pairs to cable terminals mounted on the pole. Other pairs of bridle wire were connected between the open wires and protectors. Because of the much more severe reflection requirements at the higher frequencies of the type J system, these arrangements were no longer satisfactory. The

characteristic impedance of bridle wire is roughly one-fifth of that of the open-wire circuit and it has been necessary to avoid the use of even several feet of it between the open-wire and the cable terminal or protectors. To accomplish this, separate terminals for each disc-insulated unit are mounted on the crossarm near the open-wire pairs to which they connect. Four insulated wires from each terminal go by the shortest feasible route to the longitudinal choke coils and protectors and thence to the open-wire pairs.

TOLL ENTRANCE ARRANGEMENTS

The new disc-insulated cable used for intermediate cables was also suited for lead-in or toll entrance cables.

When an auxiliary station is established at a point along an open-wire line where there has not previously been an office, it is usually located close to the line so that the lengths of lead-in cable required are comparatively short. Lengths of this cable up to about 175 feet can be loaded to open-wire impedances with adjustable loading units in the repeater station. For longer lead-in cables up to 300 feet, supplementary loading may be mounted directly on the pole at the cable terminals.

When an auxiliary repeater station is not close to the open-wire line, or at main repeater stations which are frequently in towns and separated from the open-wire line by greater lengths of toll entrance cable, it is still possible to use the loaded disc-insulated cable. Because of the cost of this cable and its loading, however, it has sometimes been found more economical to build a hut near the open-wire terminal pole and to separate the type J from the type C and lower frequency facilities at that point by means of filters. The connection from the open-wire line to the hut is provided by what is usually a short length of loaded disc-insulated cable. From that point, the type J frequencies are led into the toll office over non-loaded paper-insulated pairs while the C and lower frequency facilities are brought in over the existing pairs, usually loaded. By thus limiting the frequencies transmitted over the non-loaded cable pairs to the J range, it becomes practical to design transformers for suitable impedance matching.

The line filter sets located in the hut are designed for a nominal impedance of 560 ohms which is a compromise for the range of impedances normally found with different wire sizes and spacings. An accurate match with the line is obtained with a building-out network which is adjusted at the time of installation to fit the particular open-wire pair involved. On the office side of this line filter set a transformer provides for stepping down the impedance from 560 ohms to the

impedance of the toll entrance cable, which is usually about 125 ohms. Adjustment of this impedance over the necessary range to match impedances of particular cable pairs is provided by means of taps on the transformer. At the office another transformer similarly tapped is employed to match the toll entrance cable pair impedance to that of the office wiring.

Fig. 12 shows the losses of the commonly used 19-, 16- and 13-gauge paper-insulated toll entrance cable, a new 10-gauge low capacity cable, and the new disc-insulated cable. Because of the high losses of the smaller gauge pairs, it is sometimes economical to place new 10-gauge cable to save repeater costs.

For the office wiring of the J system a rubber-covered shielded pair is used to provide the desired flexibility and freedom from capacitance variation due to humidity changes. Its impedance at 140 kilocycles is approximately 125 ohms. The repeater and terminal high frequency impedances are designed to match this impedance very closely.

Fig. 13 illustrates the arrangement of the toll entrance equipment involved in matching the line impedance to that of the equipment with a minimum of reflection. The terminal is illustrated to the left. The high frequency line passes to the line filter set which is here shown as located in a filter hut. There it is joined by the type C and lower frequency circuits and passes through the lead-in cable and protective arrangements on the terminal pole.

Proceeding toward the right in the figure, the arrangement at an auxiliary repeater station is shown. In this case the type J frequencies are amplified in the repeater, but the type C and lower frequencies are by-passed through filters which suppress longitudinal and metallic transmission above 30 kilocycles. At the right is shown a combined type J and type C main repeater office.

Satisfactory crosstalk between pairs in entrance and intermediate cables carrying J systems is effected through special selection methods and the application of balancing condensers.

REFLECTION COEFFICIENTS

The success of the various measures taken to insure good impedance matching is shown by the curves of Fig. 14, which are of reflection coefficients measured at an auxiliary repeater station. Curve A, the solid line, gives the coefficient between the open-wire pair and the lead-in cable at the terminal pole. The smaller variations are due partly to irregularities of the open-wire line and, at the lower frequencies, partly to the test terminations at the distant end. The contribution of the cable loading and office equipment is indicated by the dash-line curve

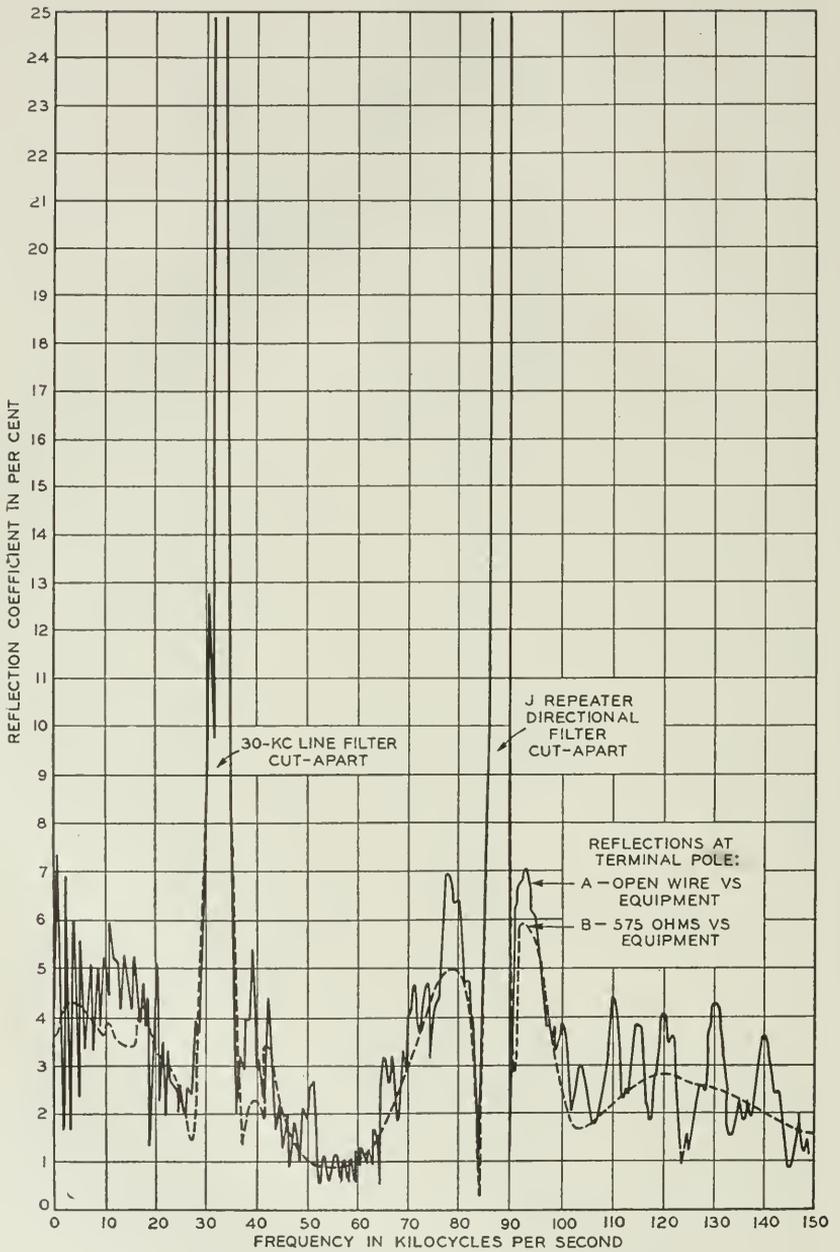


Fig. 14—Reflection coefficient at junction between open-wire and toll entrance equipment.

B , which was obtained with the open-wire line replaced by its nominal impedance, a 575-ohm resistance. The reflection between the open-wire and toll entrance and repeater equipment is well under 5 per cent over nearly all of the transmitted range.

CONCLUSION

The successful transmission of frequencies up to 140 kilocycles over open-wire pairs as compared with earlier operation up to 30 kilocycles has involved modification of the construction of the open-wire lines, new transposition designs, new toll entrance arrangements, including new types of cable, the improvement of impedance matches in various parts of the circuits, closer repeater spacings and, where ice is encountered, provision for much greater gain margins with more flexible regulation.

The first system was placed in commercial service in September 1938. By the first part of this year about 60,000 channel miles were in service over type J systems.

REFERENCES

1. "A Twelve-Channel Carrier Telephone System for Open-Wire Lines," B. W. Kendall and H. A. Affel, A. I. E. E. Winter Convention, January 1939. *Bell System Technical Journal*, January 1939.
2. "Open-Wire Line Losses," L. T. Wilson, *Bell Laboratories Record*, Vol. XVI, November 1937, Pages 95-98.
3. "High Frequency Attenuation on Open-Wire Lines," H. E. Curtis, *Bell Laboratories Record*, Vol. XVII, December 1938, Pages 121-124.
4. "Open-Wire Crosstalk," A. G. Chapman, *Bell System Technical Journal*, Vol. 13, January and April 1934, Pages 19-58, 195-238.
5. "Transcontinental Telephone Lines," J. J. Pilliod, *Electrical Engineering*, Vol. 57, October 1938, Pages 418-419 and 423. *Bell System Technical Journal*, January 1939.
6. "Some Applications of the Type J Carrier System," L. C. Starbird and J. D. Mathis, A. I. E. E., Southwestern District Convention, April 1939. This issue of the *Bell System Technical Journal*.

Abstracts of Technical Articles from Bell System Sources

*Exploration of Pressure Field Around the Human Head During Speech.*¹ H. K. DUNN and D. W. FARNSWORTH. A single speaker in a seated position repeated a fifteen-second sample of connected speech, while r.m.s. pressure measurements were made in thirteen frequency bands, and at seventy-six positions, in different directions and distances. The results are applicable to intelligibility and microphone placement problems. They show, in general, the greater variation with direction at higher frequencies. Directivity due to the size of the mouth opening appeared to enter above 5600 cycles per second, the axis at these frequencies being about 45° below the horizontal, in front.

Frequencies below 1000 cycles per second were found strongest directly downward from the lips, or nearly so. The power radiated in different directions has been calculated, and a summation gives a spectrum of the total speech power emitted by the mouth. It is proposed that similar spectra for other speakers may be obtained from pressure measurements at a single point, using the relations discovered for this speaker. The necessity for protecting a microphone used close to the mouth, from the puffs of air accompanying the speech, is demonstrated and explained.

*A Tubular Directional Microphone.*² W. P. MASON and R. N. MARSHALL. A tubular directional microphone is described which consists of a pressure type microphone coupled to an acoustic impedance element composed of a large number of tubes whose lengths vary by equal increments. The function of this variation in length is two-fold. First, the multiple resonances of the individual tubes occur at intervals so close together that the net effect of the bundle is that of an acoustic resistance over a fairly wide frequency range and so does not impair the high quality of the attached microphone. Second, high directivity is secured, because for sound incidence other than normal each tube introduces a different path length with phase cancellation resulting in a composition chamber between the microphone and the ends of the tubes. The theory of operation is summar-

¹ *Jour. Acous. Soc. Amer.*, January 1939.

² *Jour. Acous. Soc. Amer.*, January 1939.

ized and data are presented to show the performance of the instrument which is in fair agreement with the theory.

*Peak Field Strength of Atmospherics Due to Local Thunderstorms at 150 Megacycles.*³ J. P. SCHAFER and W. M. GOODALL. Atmospherics in the 150-megacycle frequency range were investigated with a broadband receiver and cathode-ray-tube scanning technique. The results are of general interest in connection with the problems of atmospheric noise interference on various types of ultra-short-wave radio-communication channels. Some of the conclusions are:

(1) The peak intensity of disturbances varies 20 decibels between different storms at the same distance. (2) The inverse distance relation is a good approximation for the calculation of the variation of peak disturbance with distance, for any distance and height of receiving antenna likely to be used in a commercial system. (3) The use of high instead of low receiving antennas increases the signal-to-disturbance ratio almost directly with height for storms within 10 miles. (4) The durations of some of the narrower peaks in any particular lightning discharge are at least as short as a few microseconds. (5) The maximum peak field strength of disturbances for a storm one mile distant is 85 decibels and for a storm ten miles distant is 65 decibels above 1 microvolt per meter at a frequency of 150 megacycles with a band width of 1.5 megacycles.

The technique of observations provided a visual indication of the noise interference which might be expected with television signals. It appears that with signal field strengths, such as might reasonably be expected, atmospherics due to thunderstorms will be noticeable for ultra-short-wave television transmission at times when storms are in progress near the point of reception.

*Metal Horns as Directive Receivers of Ultra-Short Waves.*⁴ G. C. SOUTHWORTH and A. P. KING. The paper describes some experiments made to determine the directive properties of metal pipes and horns when used as receivers of electromagnetic waves. The experiments were of two kinds. One consisted of measurements of received power, with and without the horn in place, and the other of the determination of the directional patterns of the horns in two perpendicular planes. The results indicate that electromagnetic horns of this kind provide a simple and convenient way of obtaining effective power ratios of a hundred or more (20 decibels). The effects of varying the several horn parameters are investigated. It is shown that there is an

³ *Proc. I. R. E.*, March 1939.

⁴ *Proc. I. R. E.*, February 1939.

optimum angle of flare. The possibility of forming arrays of pipes or horns is mentioned.

*Hindered Molecular Rotation and the Dielectric Behavior of Condensed Phases.*⁵ ADDISON H. WHITE. The polarizability of a liquid or a collection of randomly oriented single crystals in which polar molecules are unable to move except to rotate from one to the other of two equilibrium orientations separated by an angle β and of potential energies whose difference is E , is

$$\alpha = (\mu^2/6kT)(1 - \cos \beta)/\cosh^2 E/2kT,$$

where μ is dipole moment. This model accounts for the reduction of α in solids and liquids from the value $\mu^2/3kT$ observed in gases, and at the same time provides for anomalous dispersion in terms of discontinuous molecular processes.

⁵ *Jour. Chemical Physics*, January 1939.

Contributors to this Issue

CLIFFORD N. ANDERSON, Ph.B., University of Wisconsin, 1919; M.S., 1920. Supervising principal of schools, Amery, Wisconsin, 1913-17. Ensign Aircraft Radio, U.S.N.R.F., 1917-18. Instructor, Engineering Physics, University of Wisconsin, 1919-20; Standardizing Laboratory, General Electric Company, Lynn, Massachusetts, 1920-21; Fellow to Norway, American Scandinavian Foundation, 1921-22; Department of Development and Research, American Telephone and Telegraph Company, 1922-34; Bell Telephone Laboratories, Inc., 1934 to date. Mr. Anderson's work with the Bell System has been largely in connection with radio-telephony between boats and shore stations.

A. E. BOWEN, Ph.B., Yale University, 1921. Graduate School, Yale University, 1921-24. American Telephone and Telegraph Company, Department of Development and Research, 1924-34. Bell Telephone Laboratories, 1934-. With the American Telephone and Telegraph Company, Mr. Bowen's work was concerned principally with the inductive coordination of power and communication systems. Since 1934 he has been engaged in work in the ultra-high-frequency field, particularly on hollow wave guides.

R. S. CARUTHERS, B.S., University of Maryland, 1926; E.E., 1930. General Electric Company, 1926-28; M.S., Massachusetts Institute of Technology, 1928. U. S. Bureau of Standards, 1928-29. Bell Telephone Laboratories, 1929-. Mr. Caruthers has been engaged in the development of carrier systems.

R. N. HUNTER, B.S., Worcester Polytechnic Institute, 1915. Test Department, General Electric Company, 1915-16. Research Assistant at Massachusetts Institute of Technology, 1916-18. American Telephone and Telegraph Company, Engineering Department, 1918-19; Department of Development and Research, 1919-34. Bell Telephone Laboratories, 1934-. Mr. Hunter's work has been largely on problems of crosstalk reduction in open-wire and in shielded conductor circuits.

L. M. ILGENFRITZ, B.S. in Electrical Engineering, University of Michigan, 1920. American Telephone and Telegraph Company, Department of Development and Research, 1920-34; Bell Telephone

Laboratories, 1934-. Mr. Ilgenfritz has been engaged in the development of carrier systems.

A. G. JOHNSON, B.S. in Ceramic Engineering, Iowa State College, 1924. Western Electric Company, Development Engineering Branch, 1924-. Prior to October 1935, Mr. Johnson was engaged in development work on ceramic, rubber, phenol fibre, and plastic molding manufacture. At the present time, he is Engineer of Raw Materials at the Hawthorne Plant, Chicago, Illinois.

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 high-frequency circuits and vacuum tubes. In his present capacity as Circuit Research Engineer he directs a group in the study of amplifying problems.

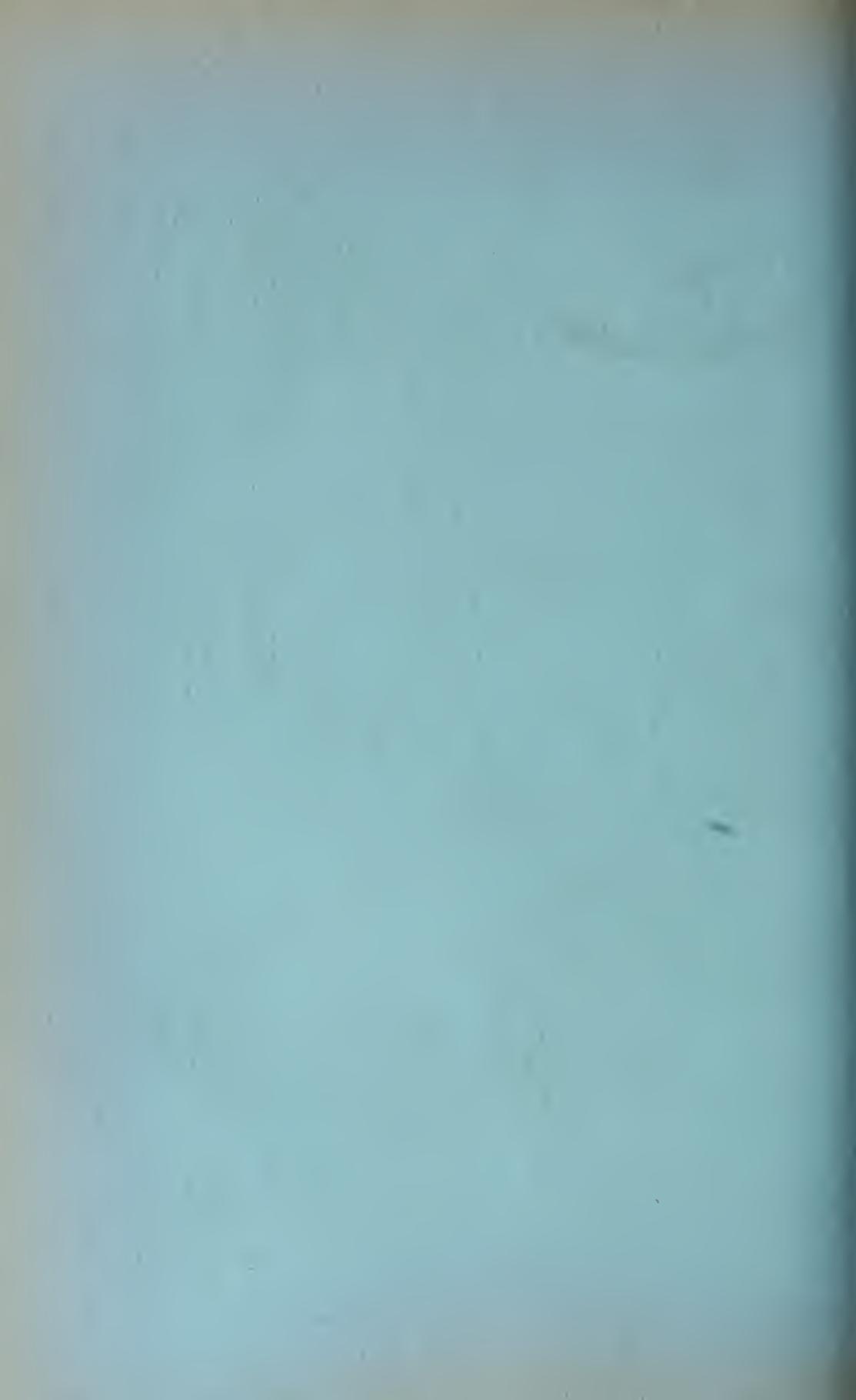
J. D. MATHIS, B.A., University of Texas, 1924; M.A., University of Texas, 1925; Tutor in Physics at University of Texas, 1924-25. Telephone equipment maintenance, 1920-24. Southwestern Bell Telephone Company, equipment engineering, 1925-. Since 1932 Mr. Mathis has been engaged principally in the engineering of central office equipment for telephone repeaters and carrier systems in Texas.

JOHN RIORDAN, B.S., Sheffield Scientific School, Yale University, 1923. American Telephone and Telegraph Company, Department of Development and Research, 1926-34; Bell Telephone Laboratories, 1934-. Mr. Riordan's work has been mainly on problems associated with inductive effects of electrified railways.

L. I. SHAW, B.S. in Ceramics, Alfred University, 1907; M.S., Syracuse University, 1908; Ph.D., University of Wisconsin, 1911; Instructor, Northwestern University, 1911-17; Assistant Chief Chemist, U. S. Bureau of Mines, 1919-24. Western Electric Company, Hawthorne Plant, Chicago, Illinois, 1924-. As Development Engineer, Dr. Shaw's work has been largely in the fields of ceramics, chemistry, hazards, and raw materials.

L. C. STARBIRD, B.E.E., University of Arkansas, 1921; Instructor, University of Arkansas, 1921-25. Southwestern Bell Telephone Company, 1925-; Equipment Engineer, 1926-32, Transmission and Protection Engineer, Texas Area, 1932-. Mr. Starbird's work has been largely in the application of carrier and repeater equipment.

A. L. WHITMAN, Harvard University, A.B., 1918; B.S. in Electrical Engineering, 1920. Harvard University Sheldon Fellowship for traveling study in Europe, 1920-21. American Telephone and Telegraph Company, Department of Development and Research, Inductive Interference and Noise Prevention Group, 1921-34. Member of Technical Staff, Transmission Development Department of Bell Telephone Laboratories, 1934-. Mr. Whitman is now engaged in field studies of noise and crosstalk as related to carrier-telephone transmission systems utilizing a broad band of frequencies.



THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS
OF ELECTRICAL COMMUNICATION

Frequency-Modulation: Theory of the Feedback Receiving
Circuit—*John R. Carson* 395

The Application of Negative Feedback to Frequency-Modulation
Systems—*J. G. Chaffee* 404

Survey of Magnetic Materials and Applications in the Tele-
phone System—*V. E. Legg* 438

Impedance Properties of Electron Streams—*Liss C. Peterson* 465

Plastic Materials in Telephone Use
—*J. R. Townsend and W. J. Clarke* 482

The Dielectric Properties of Insulating Materials
—*E. J. Murphy and S. O. Morgan* 502

Abstracts of Technical Papers 538

Contributors to this Issue 544

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

F. B. Jewett	H. P. Charlesworth	W. H. Harrison
A. F. Dixon	O. E. Buckley	O. B. Blackwell
S. Bracken	M. J. Kelly	G. Ireland
	W. Wilson	
R. W. King, <i>Editor</i>	J. O. Perrine, <i>Associate Editor</i>	

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, 1939
American Telephone and Telegraph Company

The Bell System Technical Journal

Vol. XVIII

July, 1939

No. 3

Frequency-Modulation: Theory of the Feedback Receiving Circuit

By JOHN R. CARSON

THIS paper may be regarded both as a continuation of a prior one by the writer and Thornton C. Fry¹ and as a companion of that by J. G. Chaffee² the inventor of the circuit under consideration. For an understanding of the present, an acquaintance with the prior paper¹ is absolutely necessary, since the fundamental analysis and the formulas there developed are too lengthy to be repeated here. References to that paper will be designated by (Ref.).

As the name implies, in the feedback circuit part of the incoming signal, after passing through a band-pass filter, a frequency detector³ and a demodulator, is fed back through a variable frequency oscillator. The output of the variable frequency oscillator is connected to one branch of a modulator on the other branch of which the incoming high-frequency wave is impressed. While this method of feedback differs in some respects from that of the well known feedback amplifier, it is a fair inference that some if not all of the very important advantages of the feedback amplifier may also be present in the circuit under discussion. This inference is verified by the mathematical analysis of this paper.

After a brief development of the elementary theory and formulas of the feedback circuit as a receiver of frequency-modulated waves, the greater part of the paper is devoted to deriving formulas for the signal-to-noise power ratio—a criterion of fundamental importance in estimating the merits of the system. These are then compared with the corre-

¹ "Variable Frequency Electric Circuit Theory," *this Journal*, October 1937.

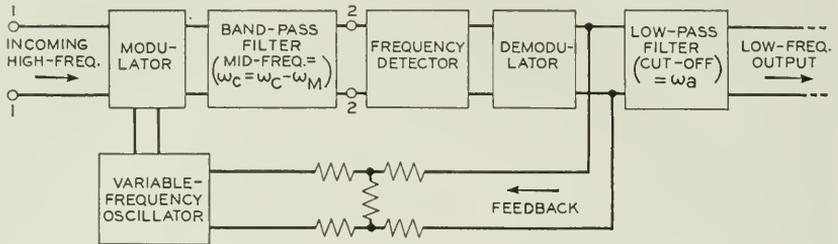
² "The Application of Negative Feedback to Frequency-Modulation Systems," *I. R. E. Proceedings*, May 1939; this issue of the *Bell Sys. Tech. Journal*.

³ The function of the "frequency detector" is to detect or render explicit the variation of the "instantaneous frequency" of the frequency-modulated wave. A more precise term, therefore, would be "frequency variation detector," but for brevity the term used in the text is preferable.

sponding ratios formulated in (Ref.) for straight reception and also reception with amplitude limitation. In this way, as regards reduction of noise and "fading," the feedback circuit is found to have advantages comparable with those attainable by amplitude limitation.⁴

I

The receiving system operates as follows (see sketch):



Feedback receiving circuit.

The incoming frequency-modulated wave at terminals 1, 1 is taken as

$$E \exp \left(i\omega_c t + i\lambda \int^t s dt \right), \quad (1)$$

where E is the wave amplitude, ω_c the carrier frequency, λ the modulation index and $s = s(t)$ is the low-frequency signal which it is desired to recover.

This wave is impressed at terminals 1, 1 on one pair of terminals of a "product" modulator; on the other pair of terminals of the modulator there is impressed the output of a local variable-frequency oscillator:

$$M \exp \left(i\omega_M t + i\mu \int^t \sigma dt \right). \quad (2)$$

Here ω_M is the "carrier" frequency of the oscillator, μ (a positive real quantity) is the index of modulation of the oscillator and $\sigma = \sigma(t)$ is the low-frequency current fed back to the oscillator.

The output wave of the modulator is then equal to

$$c_1 EM \exp \left(i(\omega_c - \omega_M)t + i \int^t (\lambda s - \mu \sigma) dt \right) + c_1 EM \exp \left(i(\omega_c + \omega_M)t + i \int^t (\lambda s + \mu \sigma) dt \right). \quad (3)$$

⁴ Armstrong, *Proc. I. R. E.*, May 1936, also see (Ref.).

The second term of (3) is suppressed by the band-pass filter.⁵ Then writing $\omega_C - \omega_M = \omega_c$, it follows that the effective output wave is

$$c_1 EM \exp \left(i\omega_c t + i \int^t (\lambda s - \mu \sigma) dt \right). \quad (4)$$

ω_c is the intermediate carrier frequency and is always $< \omega_C$, the transmitted carrier frequency. The constant c_1 is a parameter depending on the characteristics of the modulator.

The wave (4) is transmitted through the band-pass filter, and the wave arriving at terminals 2, 2 is then

$$c_1 c_2 EM \exp \left(i\omega_c t + i \int^t (\lambda s - \mu \sigma) dt \right). \quad (5)$$

The parameter c_2 (taken as a constant) depends on the transmission characteristics from the modulator to terminals 2, 2.

Assuming an ideal frequency detector (see Ref.) the output to the terminals of the rectifier (or demodulator) is

$$c_1 c_2 c_3 EM \left(1 + \frac{\lambda s - \mu \sigma}{\omega_1} \right) \exp \left(i\omega_c t + i \int^t (\lambda s - \mu \sigma) dt \right). \quad (6)$$

Here the parameters c_3 and ω_1 depend on the characteristics of the frequency detector.

Finally assuming that

$$\frac{\lambda s - \mu \sigma}{\omega_1} < 1 \quad (7)$$

the low-frequency output of the rectifier⁶ is

$$c_1 c_2 c_3 c_4 EM \left(1 + \frac{\lambda s - \mu \sigma}{\omega_1} \right). \quad (8)$$

If the constant term of (8) is suppressed and a fraction η of the rectified output is fed back to the oscillator we have, finally,

$$\sigma = \frac{m}{\mu} \frac{\lambda s}{1 + m}, \quad (9)$$

⁵ Indeed the principal function of the band-pass filter is to suppress frequencies in the neighborhood of $\omega_C + \omega_M$.

⁶ More generally a demodulator. In the present paper a straight-line rectifier is postulated for mathematical simplicity but the theory applies equally well to other forms of detection or demodulation.

where

$$\begin{aligned} m &= c_1 c_2 c_3 c_4 \frac{\eta \mu E M}{\omega_1}, \\ &= C \frac{\eta \mu E M}{\omega_1}. \end{aligned} \quad (10)$$

The low-frequency current delivered to the receiver through the low-frequency output circuit proper differs from σ as given by (9) by a constant factor only.

From the foregoing we note that

$$\lambda s - \mu \sigma = \frac{\lambda s}{1 + m} \quad (11)$$

and that the "instantaneous frequency" of the intermediate high-frequency wave (4) is

$$\omega_c + \frac{\lambda s}{1 + m}. \quad (12)$$

Hereinafter, without any loss of generality, we suppose that $-1 \leq s(t) \leq 1$. Consequently the intermediate frequency-modulated wave has a frequency variation lying between $\pm \lambda/(1 + m)$, whereas in the incoming frequency-modulated wave, the frequency variation lies between $\pm \lambda$.

We note also from (9) that if the parameter m is large compared with unity, the low-frequency received wave is approximately given by

$$\sigma = \frac{\lambda}{\mu} s. \quad (13)$$

The recovered signal is thus (for large values of m) seen to be independent of the amplitude, E , of the incoming high-frequency wave; therefore, the system is insensitive to "fading."

II

We now take up the problem of calculating the relative low-frequency *noise* and *signal* powers, the ratio of which is of fundamental importance in appraising the merits of the receiving circuit. In this we shall closely follow the methods developed in Section IV (Ref.).

We suppose that at terminals 1, 1 there enters, in addition to the signal, a typical noise element

$$a_n \exp(i(\omega_c + \omega_n)t + i\alpha_n). \quad (14)$$

We write for convenience in the subsequent analysis

$$a_n = A_n E. \tag{15}$$

A_n is then the *relative* amplitude of the noise element, referred to the amplitude E of the high-frequency signal. We shall suppose throughout that A_n is small compared with unity; that is, the noise is small compared with the signal.

We further suppose that at terminals 2, 2 between the band-pass filter and the frequency detector there is introduced a second typical noise element

$$b_n \exp(i(\omega_c + \omega_n)t + i\beta_n), \tag{16}$$

which is entirely independent of the noise element (14). This may be regarded as caused by tube-noise in amplifiers (not shown in sketch).

We write

$$b_n = B_n E \tag{17}$$

so that B_n is the relative amplitude of the noise element, referred to the amplitude E of the incoming signal wave. It also is assumed small compared with unity.

The total input to the frequency detector, neglecting the random phase angles, is then

$$c_1 c_2 E M \left[\exp\left(i \int^t \Omega dt\right) + A_n \exp\left(i \int^t (\Omega + \Omega_n^a) dt\right) + \frac{B_n}{c_1 c_2 M} \exp\left(i \int^t (\Omega + \Omega_n^b) dt\right) \right], \tag{18}$$

where

$$\begin{aligned} \Omega &= \omega_c + \lambda s - \mu \sigma \\ \Omega_n^a &= \omega_n - \lambda s \\ \Omega_n^b &= \omega_n - \lambda s + \mu \sigma. \end{aligned} \tag{19}$$

The output of the frequency detector is then (see Ref.)

$$\begin{aligned} c_1 c_2 c_3 E M \exp\left(i \int^t \Omega dt\right) & \times \left[1 + \frac{1}{\omega_1} (\lambda s - \mu \sigma) \right. \\ & + A_n \left(1 + \frac{1}{\omega_1} (\omega_n - \mu \sigma) \exp\left(i \int^t \Omega_n^a dt\right) \right) \\ & \left. + \frac{B_n}{c_1 c_2 M} \left(1 + \frac{\omega_n}{\omega_1} \right) \exp\left(i \int^t \Omega_n^b dt\right) \right]. \end{aligned} \tag{20}$$

After the output as given by (20) is rectified, the constant term suppressed, and only first powers in A_n and B_n retained, we get finally

$$\sigma = \left(\frac{m/\mu}{1+m} \right) \left[\left(\lambda s + A_n \left(\omega_1 + \omega_n - \frac{m}{1+m} \right) \lambda s \right) \cos \int^t \Omega_n^a dt + \frac{B_n}{c_1 c_2 M} (\omega_1 + \omega_n) \cos \int^t \Omega_n^b dt \right]. \quad (21)$$

Now the right-hand side of (21) corresponds precisely with formula (64) (Ref.) on which the calculation of the relative low-frequency noise and signal powers is based. Consequently following the methods developed in Ref. and assuming A_n and B_n small we get

$$\bar{\sigma}^2 = \left(\frac{m/\mu}{1+m} \right)^2 \left[\lambda^2 \bar{s}^2 + \left(\frac{1}{3} \omega_a^2 + \omega_1^2 + \frac{\lambda^2 \bar{s}^2}{(1+m)^2} \right) \omega_a N_a^2 + \left(\frac{1}{3} \omega_a^2 + \omega_1^2 + \overline{(\lambda s - \mu \sigma)^2} \right) \omega_a N_b^2 / c_1^2 c_2^2 M^2 \right]. \quad (22)$$

The *relative* low-frequency noise and signal powers are then (omitting the common factor $\left(\frac{m/\mu}{1+m} \right)^2$)

$$P_N = \frac{1}{3} \omega_a^3 N_a^2 \left[1 + 3 \left(\frac{\omega_1}{\omega_a} \right)^2 + 3 \frac{(\lambda/\omega_a)^2 \bar{s}^2}{(1+m)^2} \right] + \frac{1}{3} \frac{\omega_a^3 N_b^2}{c_1^2 c_2^2 M^2} \left[1 + 3 \left(\frac{\omega_1}{\omega_a} \right)^2 + 3 \frac{(\lambda s - \mu \sigma)^2}{\omega_a^2} \right], \quad (23)$$

$$P_S = \lambda^2 \bar{s}^2.$$

In these formulas N_a^2 is proportional to the noise power level in the neighborhood of the carrier frequency ω_c ; it enters at the input terminals 1, 1 (see Ref. Appendix 2). N_b^2 is proportional to the noise power level in the neighborhood of the intermediate carrier frequency ω_c ; it enters at terminals 2, 2. ω_a is the highest essential frequency in the low-frequency signal $s(t)$; it is the cut-off frequency of the low-pass filter.

Formula (22) is solvable (see Appendix) but a simple approximate solution, valid when the noise is small compared with the signal, is made possible by observing that under this restriction

$$\lambda s - \mu \sigma = \frac{\lambda s}{1+m} \quad (11)$$

to a good approximation. Introducing this approximation into P_N as given by (23) and writing

$$N^2 = N_a^2 + N_b^2/c_1^2c_2^2M^2, \tag{24}$$

we have

$$P_N = \frac{1}{3} \omega_a^3 N^2 \left(1 + 3 \left(\frac{\omega_1}{\omega_a} \right)^2 + 3 \frac{(\lambda/\omega_a)^2 \bar{s}^2}{(1+m)^2} \right), \tag{25}$$

$$P_S = \lambda^2 \bar{s}^2.$$

Now from the inequality, necessary for rectification,

$$\omega_1 > \frac{\lambda}{1+m}$$

it is seen that as the parameter m is increased, ω_1 may be reduced by the factor $1/(1+m)$. In accordance with this, we replace ω_1 by $\omega_1/(1+m)$ in (25) and get

$$P_N = \frac{1}{3} \omega_a^3 N^2 \left[1 + \frac{3}{(1+m)^2} \left(\left(\frac{\omega_1}{\omega_a} \right)^2 + \left(\frac{\lambda}{\omega_a} \right)^2 \bar{s}^2 \right) \right], \tag{26}$$

$$P_S = \lambda^2 \bar{s}^2.$$

The noise power P_N can be still further reduced by eliminating ω_1 from (26) by a circuit arrangement explained in Ref. Section III; if this is done we get, instead of (26),

$$P_N = \frac{1}{3} \omega_a^3 N^2 \left(1 + 3 \frac{(\lambda/\omega_a)^2 \bar{s}^2}{(1+m)^2} \right), \tag{27}$$

$$P_S = \lambda^2 \bar{s}^2.$$

We have now to compare the relative noise and signal powers of the feedback with (1) straight reception *without* feedback and (2) reception with *amplitude limitation*.

For *straight reception* (without feedback) we have (see equation (68), Ref.), corresponding to (26),

$$P_N = \frac{1}{3} \omega_a^3 N^2 \left(1 + 3 \left(\frac{\omega_1}{\omega_a} \right)^2 + 3 \left(\frac{\lambda}{\omega_a} \right)^2 \bar{s}^2 \right), \tag{28}$$

$$P_S = \lambda^2 \bar{s}^2,$$

and corresponding to (27)

$$P_N = \frac{1}{3} \omega_a^3 N^2 \left(1 + 3 \left(\frac{\lambda}{\omega_a} \right)^2 \bar{s}^2 \right), \tag{29}$$

$$P_S = \lambda^2 \bar{s}^2.$$

When noise reduction is effected by *amplitude limitation* the corresponding relative noise and signal powers (see equation (78), Ref.) are

$$\begin{aligned} P_N &= \frac{1}{3} \omega_a^3 N^2, \\ P_S &= \lambda^2 \bar{s}^2. \end{aligned} \quad (30)$$

If we assume as above that $-1 \leq s \leq 1$ and \bar{s}^2 is of the order of magnitude of $1/2$, then in practical applications $\lambda/\omega_a \gg 1$ and $\omega_1 > \lambda$. On this basis comparison of (26) with (28) and (27) with (29) shows that, when $m \gg 1$, the noise power with *feedback* is very much smaller than *without feedback*, the ratio of the noise powers in the two cases being approximately $1/(1+m)^2$. (This assumes, of course, that N^2 is approximately equal in the two cases.)

Comparing, however, the noise power with *feedback* to that obtainable by *amplitude limitation*, it will be seen that in order to reduce the former to the order of magnitude of the latter it is necessary that

$$\frac{(\lambda/\omega_a)}{(1+m)} < 1. \quad (31)$$

From the preceding it is seen that the performance of the feedback circuit and the reduction in noise-power ratio obtainable depend in a fundamental manner on the parameter m , defined above by the formula

$$m = c_1 c_2 c_3 c_4 \frac{\eta \mu EM}{\omega_1}. \quad (32)$$

If the characteristics of the modulator rectifier and variable-frequency oscillator are stipulated, it is possible to calculate m in terms of these characteristics and the constants and connections of the network. It is experimentally determinable (among other ways) as follows:

Let the feedback circuit be opened between the low-pass filter and the variable-frequency oscillator, and let the filter be closed by an impedance equal to that of the oscillator as seen from the filter. Then $m = 0$ (since there is no low-frequency feedback to the oscillator) but m/μ is finite.

Denoting the value of σ under these circumstances by σ_1 , it follows from (9) that

$$\sigma_1 = \frac{m}{\mu} \lambda s. \quad (33)$$

Consequently dividing σ_1 by σ , as given by (9), we have

$$\begin{aligned} 1 + m &= \sigma_1 / \sigma, \\ m &= \frac{\sigma_1 - \sigma}{\sigma}. \end{aligned} \quad (34)$$

Stated in words, $1 + m$ is the reciprocal of the ratio of the values of σ *without* and *with* the low-frequency feedback into the oscillator. It should be noted that this requires that the band-pass filter transmit the frequency band 2λ centered on ω_c .

APPENDIX

In formula (22), the expression $\overline{(\lambda s - \mu\sigma)^2}$ has been replaced by $\lambda^2\overline{s^2}/(1 + m)^2$, its value when the noise is absent. When noise is present, but small compared with the signal, this should still give a good approximation for $\overline{\sigma^2}$. We now propose to derive an exact solution of (22); to this end we write

$$\lambda s - \mu\sigma = \frac{\lambda s}{1 + m} - n(t) \tag{1a}$$

which is always possible.

Now inspection of (1a) shows that $n(t)$ is the value of $\mu\sigma$ when $s = 0$; consequently $\overline{n^2} = \mu^2\overline{\sigma_0^2}$ where $\overline{\sigma_0^2}$ is given by (34). Furthermore, since s and n are entirely independent, $\overline{sn} = 0$, and

$$\overline{(\lambda s - \mu\sigma)^2} = \frac{\lambda^2\overline{s^2}}{(1 + m)^2} + \mu^2\overline{\sigma_0^2}. \tag{2a}$$

Substitution of (2a) in (22) gives for P_N , instead of (23),

$$P_N = \frac{1}{3} \omega_a^3 N^2 \left(1 + 3 \left(\frac{\omega_1}{\omega_2} \right)^2 + 3 \frac{(\lambda/\omega_a)^2}{(1 + m)^2} \overline{s^2} \right) + \omega_a \mu^2 \sigma_0^2 N b^2 / c_1^2 c_2^2 M^2. \tag{3a}$$

The second term is a second order quantity in the noise power and may therefore be neglected when the noise is small, as is assumed throughout this paper.

The Application of Negative Feedback to Frequency-Modulation Systems*

By J. G. CHAFFEE

Negative feedback can be applied to a frequency-modulation receiver of superheterodyne type by causing a portion of the output voltage to frequency-modulate the local oscillator in such phase as to reduce the output signal. As a consequence of this arrangement the effective frequency modulation of the intermediate wave is diminished by the feedback factor. This reduction is accompanied by a decrease in noise and distortion. Restoration of the original signal level by increasing the degree of modulation at the transmitter brings about a corresponding increase in signal-to-noise ratio provided the disturbance is not too great, while distortion ratios are improved to about the same extent. These effects are treated analytically for the case where the disturbance level is sufficiently low to permit simplifying assumptions to be made. The results are in general agreement with observations made on an experimental laboratory system.

Comparing the feedback system with a frequency-modulation system using amplitude limitation, the ratio of signal level to noise level in the absence of modulation is identical in two systems. During modulation the noise level increases in the feedback system by an amount depending upon the ratio of the effective frequency shift of the intermediate-frequency wave to the signal band width. By keeping this ratio small, the increase in noise during modulation can be made relatively unimportant.

In cases where the disturbance level is high, phenomena have been observed which are very similar to those encountered when amplitude limitation is used.

INTRODUCTION

THIS paper describes a method for improving the performance of receivers designed to receive frequency-modulated waves. In its broader aspects this method can be described as the application of the principle of negative feedback to a superheterodyne frequency-modulation receiver. In its details the application of the feedback principle necessitates the use of a rather unusual circuit arrangement. This circuit differs from that of the simple feedback system in that the voltages fed back are not of the same frequency as those applied

* Presented before New York Section of I. R. E., May 3, 1939. Published in *Proceedings, I. R. E.*, May 1939.

to the input of the receiver, and are caused to influence the response of the system by modifying the performance of the modulator.

In the ordinary feedback amplifier a part of the output voltage is carried back to the input and there combined with the applied voltage. The result is to modify the output and if the gain of the system is thereby reduced the feedback is said to be negative. The many advantages which result from negative feedback have been described by Black¹ and are coming to be more generally appreciated. The present paper deals with a method for adapting this principle to a frequency-modulation receiver and will show an example of its application to an experimental system in the laboratory.

GENERAL DISCUSSION

Method of Applying Feedback

Consider a frequency-modulation receiver in which the incoming wave is combined with the output of a local oscillator in a modulator to produce a wave of intermediate frequency. This is then amplified, converted into an amplitude-modulated wave, and finally detected. The frequency of the intermediate wave is equal to the instantaneous difference in the frequencies of the incoming carrier and the local oscillator. So long as the frequency of this oscillator remains fixed the intermediate wave will be frequency-modulated in exact correspondence with the incoming wave. Suppose now that the local oscillator is frequency-modulated from a source of the same frequency and phase as that applied to the transmitter. As the index of modulation at the local oscillator is increased from zero the extent to which the intermediate wave is modulated will diminish since its instantaneous frequency is equal to the difference in the frequencies of the two sources. It then follows that if these two devices are modulated to the same extent the difference frequency will become constant and the output of the system will be zero. Finally a further increase in modulation of the local oscillator will cause the intermediate wave to be modulated with a 180-degree phase reversal.

This process can be readily analyzed as follows: Assume the oscillator at the transmitter to have been frequency-modulated by the signal wave

$$e = E_1 \cos pt. \quad (1)$$

The voltage delivered to the modulator by the incoming wave will be

$$A \cos (\omega_1 t + x_1 \sin pt + \phi_1) \quad (2)$$

¹ H. S. Black, "Stabilized Feedback Amplifiers," *Elec. Engg.*, vol. 53, pp. 114-120, January 1934.

where $x_1 = \Delta\omega_1 \div p = \rho E_1 \div p$, and $\Delta\omega_1$ is 2π times the maximum frequency shift. The local oscillation impresses the wave

$$B \cos (\omega_2 t + x_2 \sin pt + \phi_2) \quad (3)$$

where

$$x_2 = \Delta\omega_2 \div p.$$

Application of these waves to a square-law modulator will yield a difference frequency wave proportional to

$$AB \cos [(\omega_1 - \omega_2)t + (x_1 - x_2) \sin pt + \phi_1 - \phi_2] \quad (4)$$

for the case where $\omega_1 > \omega_2$, or when the reverse is true

$$AB \cos [(\omega_2 - \omega_1)t - (x_1 - x_2) \sin pt + \phi_2 - \phi_1]. \quad (5)$$

In either case the resultant modulation index of the intermediate wave is the numerical difference of the original indexes, the difference in sign

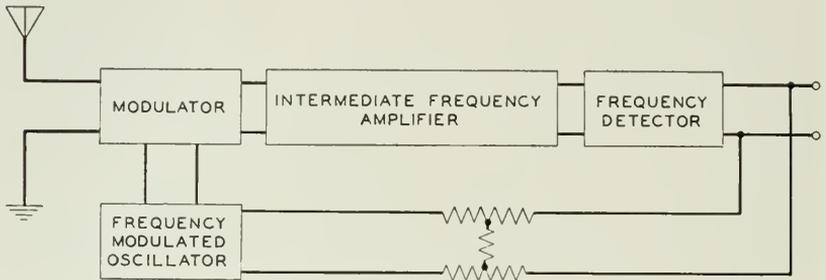


Fig. 1—Basic feedback circuit.

in the two cases signifying that the detected outputs will be of opposite phase. If $x_1 = x_2$ the modulation is reduced to zero, and if $x_2 > x_1$ modulation reappears with a phase reversal. It might be noted that if x_2 were originally made negative, thus causing the two oscillators to be frequency-modulated in opposite phase, the apparent modulation of the incoming wave could be increased indefinitely.

Suppose, now, that instead of frequency-modulating the local oscillator from an independent source, the equivalent is accomplished in a practical way. For this purpose a voltage from the output of the receiver is impressed upon the local oscillator as shown in Fig. 1. The transmitted wave will then have a modulation index $x_1 = \rho_1 E_1 \div p$, while the local oscillator, being acted upon by a portion of the output voltage E_0 , will have an index $x_2 = k\rho_2 E_0 \div p$. If the frequency detector² is assumed to be linear the amplitude of the detected output

² The term *frequency detector* is used in this paper to designate the combination of conversion circuit and amplitude detector. A more extended discussion of modulation and detection is given in Appendix A.

will be proportional to the product of the amplitudes A and B of the incoming and local oscillator waves, the resultant index of the intermediate wave, and the slope factor a_1 . Thus we can write the output voltage amplitude

$$E_0 = \alpha a_1 AB(x_1 - x_2)p = \alpha a_1 AB(\rho_1 E_1 - k\rho_2 E_0). \quad (6)$$

Therefore

$$E_0 = \frac{\alpha a_1 AB \rho_1 E_1}{1 + \alpha k a_1 AB \rho_2}. \quad (7)$$

Setting $\alpha a_1 AB = \mu$ and $k\rho_2 = -\beta$ we obtain the familiar form encountered in the analysis of feedback amplifiers

$$E_0 = \frac{\mu(\rho_1 E_1)}{1 - \mu\beta}. \quad (8)$$

Without feedback the output of the system is merely $\mu(\rho_1 E_1)$. The feedback factor $1 + \alpha a_1 k AB \rho_2$ is a measure of the extent to which the over-all gain of the system has been modified by feedback. If this factor is greater than unity the feedback is negative, while if k is made negative by reversing the feedback connections the effect is regenerative, and instability is encountered when the factor becomes zero.

It will be noted that when $\alpha a_1 k AB \rho_2 \gg 1$, (7) becomes

$$E_0 = \frac{\rho_1 E_1}{k\rho_2}. \quad (9)$$

Thus for large amounts of feedback, the output signal becomes independent of such factors as fading of the incoming wave, variations in the local oscillator voltage, or changes in detector efficiency. Hence automatic gain control is secured. This feature is equivalent to that found in ordinary feedback amplifiers in that for large amounts of feedback the over-all gain becomes independent of variations in the performance of the amplifier proper.

Reduction of Noise

The application of negative feedback in the manner described brings about a reduction in signal level by decreasing the effective modulation of the received wave. It then becomes possible to increase the modulation level at the transmitter to a corresponding degree and thus to restore the output signal to its former value. This process is made possible through the use of frequency rather than amplitude modulation since the permissible degree of modulation is then deter-

mined by the receiver characteristics. It will be shown that feedback also reduces the noise level at the output of the receiver, provided that the disturbance is not too great. Thus when the modulation level is raised to offset the effect of feedback an improvement in signal-to-noise ratio is realized.

The mechanism by which noise is reduced can be described qualitatively as follows: Noise at the output terminals of the receiver is caused to frequency-modulate the intermediate wave in such fashion as to produce, upon detection, a component which tends to cancel that which would exist in the absence of feedback. An analysis of this process for the case where the carrier is large compared with the disturbance responsible for the noise is developed³ in Appendix B. It is assumed that the disturbance can be represented by a continuous spectrum of sinusoidal voltages of equal amplitude but phased at random. Impressed along with the disturbance is the signaling wave (2). Then if N^2 is the mean disturbing power per unit of band width in the vicinity of the carrier frequency, and r_1 is the resistance of the input circuit, it is shown that the output noise power is⁴

$$P_N = \frac{2N^2 r_1}{F^2} \left[a_0^2 + \frac{a_1^2 \Delta\omega^2}{2F^2} + \frac{a_1^2 q_a^2}{3} \right] q_a \quad (10)$$

where a_0 and a_1 are, respectively, the gain and slope factor of the intermediate amplifier and conversion system as defined by (47), and q_a represents the upper limit of frequency response of the output circuit, or the upper limit of audibility as the case may be. F is the feedback factor ($1 - \mu\beta$). The corresponding signal power is

$$P_s = \frac{A^2 a_1^2 \Delta\omega^2}{2F^2}. \quad (11)$$

The reduction in signal level occasioned by feedback can be offset by increasing the frequency shift of the transmitted wave. If it is increased so as to have the value $\Delta\Omega = F\Delta\omega$ then the shift of the inter-

³ An analysis of the effect of feedback upon noise in this system was first developed by J. R. Carson by methods similar to those used in "Variable Frequency Electric Circuit Theory with Application to the Theory of Frequency Modulation," Carson and Fry, *Bell Sys. Tech. Jour.*, vol. 16, pp. 513-540, October 1937. This has been embodied in a paper by Mr. Carson entitled, "Frequency Modulation: Theory of the Feedback Receiving Circuit," published in this issue of the *Bell Sys. Tech. Jour.* Carson's treatment is more general in that an arbitrary signal wave is postulated whereas the analysis given in Appendix B is restricted to a sinusoidal signal wave. The methods used here are more elementary and may therefore appeal to a somewhat wider audience.

⁴ The expressions for signal and noise power used in this section are relative. Factors determining their absolute magnitude are given in the Appendix. In all cases the symbol $\Delta\omega^2$ is to be taken as signifying $(\Delta\omega)^2$.

mediate-frequency wave will be restored to its original value $\Delta\omega$ and the signal level will remain unchanged. Then the noise power becomes

$$P_N = \frac{2N^2r_1}{F^2} \left[a_0^2 + \frac{a_1^2\Delta\Omega^2}{2F^2} + \frac{a_1^2q_a^2}{3} \right] q_a \quad (12)$$

which can be written

$$P_N = \frac{2N^2r_1}{F^2} \left[a_0^2 + \frac{a_1^2\Delta\omega^2}{2} + \frac{a_1^2q_a^2}{3} \right] q_a. \quad (12a)$$

The noise-to-signal power ratio is improved by the factor F^2 , since

$$\frac{P_N}{P_s} = \frac{1}{F^2} \frac{4N^2r_1}{A^2} \left[\frac{a_0^2}{a_1^2\Delta\omega^2} + \frac{1}{2} + \frac{q_a^2}{3\Delta\omega^2} \right] q_a. \quad (13)$$

Of the factors in (12) the first is the result of modifications of the amplitude of the incoming wave by the disturbance. Although subject to reduction by feedback it can be balanced out completely by the use of differentially connected frequency detectors having slope factors a_1 and $-a_1$. The second term is dependent upon the degree of modulation of the intermediate wave. It is usually of less consequence in its effect upon the listener. The remaining term is the result of phase modulation of the signal wave by the disturbance. Under the condition that the output signal is held constant by increasing the transmitted band, all terms which contribute to the noise level in a given case are reduced alike by feedback.

If differential frequency-detection is employed (12a) becomes

$$P_N = \frac{2N^2r_1a_1^2}{F^2} \left[\frac{\Delta\omega^2}{2} + \frac{q_a^2}{3} \right] q_a. \quad (14)$$

During non-signaling periods the first term becomes zero. Hence during periods of modulation the background noise power is increased by the factor

$$1 + \frac{3}{2} \frac{\Delta\Omega^2}{F^2q_a^2} = 1 + \frac{3}{2} \frac{\Delta\omega^2}{q_a^2}. \quad (15)$$

If conditions are such that the maximum shift experienced by the intermediate-frequency wave is numerically equal to q_a , then the noise level will be increased by 4 decibels during periods of full modulation. In the experimental system to be described the ratio of $\Delta\omega$ to q_a was allowed to attain a value of 1.75, resulting in a maximum increase of 7.5 decibels.

In order to secure large noise reduction it is necessary to produce a frequency shift in the transmitted wave much greater than the signal band width. Thus, in common with frequency-modulation systems employing amplitude limiters,⁵ this advantage is secured at the expense of band width. In this connection it is of interest to compare amplitude limitation and feedback systems on the basis of equal width of transmitted band. Hence it will be assumed that in each case the transmitted wave is modulated to the extent of $\pm \Delta\Omega = \pm F\Delta\omega$. In Fig. 2 are shown idealized characteristics of conversion systems which might be used in the two systems. The adjustment shown in Fig. 2(a) is suitable for use with the limiter system. With the feedback system that shown in Fig. 2(b) would be necessary to secure the same percentage of amplitude modulation after conversion. This represents

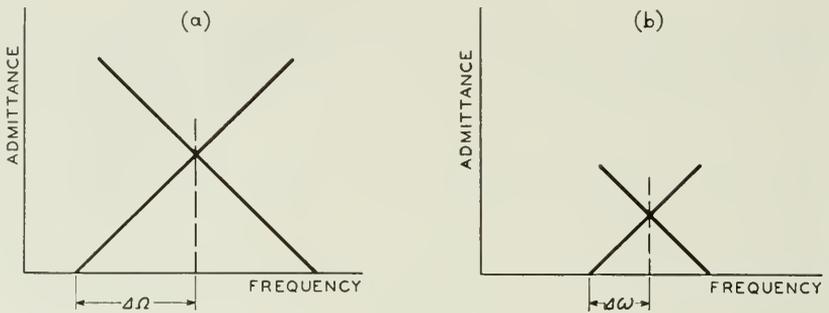


Fig. 2—Idealized conversion-system characteristics for (a) limiter system, (b) feedback system.

the minimum band width which could be provided in the conversion system with feedback, though several considerations make it desirable to use an adjustment lying somewhere between the two shown. The manner of tuning or the slope factors assumed in each case are immaterial to the present comparison provided that, in either system, the linear portion of the characteristic is of sufficient extent to effect proper conversion of the intermediate-frequency wave.

The noise-to-signal power ratio obtainable with the feedback system will be that given by (13) with the first term omitted since it is balanced out by the push-pull arrangement. Thus

$$\frac{P_N}{P_s} = \frac{4N^2r_1}{F^2A^2} \left[\frac{1}{2} + \frac{q_a^2}{3\Delta\omega^2} \right] q_a. \quad (16)$$

⁵ E. H. Armstrong, "A Method of Reducing Disturbances in Radio Signaling by a System of Frequency Modulation," *Proc. I. R. E.*, vol. 24, pp. 689-740, May 1936.

In the system corresponding to Fig. 2(a), the signal power will be

$$\frac{A^2 a_1^2 \Delta \Omega^2}{2}. \quad (17)$$

Equation (10) can be used to determine the noise power level for a frequency-modulation system without amplitude limitation by setting $F = 1$. If balanced detection is used the term in a_0 becomes zero. It has been shown by Carson and Fry³ that the addition of an ideal limiter removes all terms but the third, with either single or balanced detectors. Hence for the limiter system the noise ratio becomes

$$\frac{P_N}{P_s} = \frac{4N^2 r_1}{A^2} \left(\frac{q_a^3}{3\Delta\Omega^2} \right). \quad (18)$$

Since $\Delta\Omega = F\Delta\omega$ this can be put in a form similar to (16)

$$\frac{P_N}{P_s} = \frac{4N^2 r_1}{F^2 A^2} \left(\frac{q_a^2}{3\Delta\omega^2} \right) q_a. \quad (19)$$

Comparing (16) and (19) it is seen that the noise ratio in the feedback system is greater than that for the limiter system by the factor (15). This is a consequence of the increase in noise level which occurs during modulation in the former system. The ratio of noise level during non-signaling periods to signal level is identical in the two systems.

While the noise increment which appears during modulation is usually not of great consequence from a practical standpoint, it can be reduced by increasing the feedback factor beyond the point dictated by the signal band which it is permissible to transmit. In previous discussions it has been assumed that the application of a given amount of negative feedback is to be accompanied by a corresponding increase in modulation level at the transmitter. In this way the modulation of the intermediate frequency wave is kept constant so as to maintain a fixed signal level as the band width of the transmitted wave is increased. Having arrived at a limiting value of band spread the feedback factor can be increased further. Suppose that modulation of the transmitter is to be limited to a value of $\Delta\Omega = F_1\Delta\omega$, but that the feedback applied to the receiver is made to exceed F_1 by a factor which we shall call F_2 . Then the actual feedback factor will be F_1F_2 and we have

$$P_s = \frac{A^2 a_1^2 \Delta \omega^2}{2F_2^2} \quad (20)$$

$$P_N = \frac{2N^2 r_1}{F_1^2 F_2^2} \left[\frac{a_1^2 \Delta \omega^2}{2F_2^2} + \frac{a_1^2 q_a^2}{3} \right] q_a \quad (21)$$

giving

$$\frac{P_N}{P_s} = \frac{4N^2r_1}{A^2F_1^2} \left[\frac{1}{2F_2^2} + \frac{q_a^2}{3\Delta\omega^2} \right] q_a. \quad (22)$$

Thus the additional feedback represented by the factor F_2 is directly effective against the noise increment accompanying modulation. Reduction of this increment brings about a still closer correspondence between the limiter and feedback systems as is seen by setting $F = F_1$ in (19) and comparing with (22).

The above discussion and the analysis given in Appendix B are based upon the assumption that the carrier amplitude is large compared with that of the disturbance. A rigorous analysis, applicable to the case where this ratio is unrestricted, becomes exceedingly involved. However, a rough indication of what is to be expected in the presence of a high level of disturbance can be obtained quite simply from (52) developed in Appendix B. Assuming that modulation is not present this can be put in the simple form

$$\sigma = \frac{1}{F} \frac{Q'(a_0 + a_1\omega_n)}{1 + \frac{Q'}{A'} \left(\frac{F-1}{F} \right) \cos \omega_n t} \cos \omega_n t. \quad (52a)$$

When $Q' \ll A'$ the wave form of the output noise produced by a single element of disturbance is very closely a sinusoid. However, when Q' and A' become comparable in magnitude the output wave becomes badly peaked when $\omega_n t = n\pi$. While the above expression is only a very rough approximation under these conditions, a plot of the wave form so obtained exhibits all of the essential characteristics of the curves given by Crosby in a recent paper⁶ dealing with noise in frequency-modulation systems using amplitude limitation. These curves show a similar peaking of the output-noise wave form when the ratio of carrier to disturbance amplitude is in the vicinity of unity. The description given by Crosby of the manifestations of this phenomenon observed in an experimental system applies rather closely to what has been found in the feedback system. A more detailed account will be found in a later section.

Examination of (52a) shows that the output wave can assume very large and even infinite peak values when Q' and A' are approximately equal. The existence of high peak values of noise implies both a large instantaneous deviation in the frequency of the intermediate wave, and a conversion-circuit characteristic of unlimited extent. The finite

⁶ Murray G. Crosby, "Frequency Modulation Noise Characteristics," *Proc. I. R. E.*, vol. 25, pp. 472-514, April 1937. The curves referred to are given in Fig. 4 of the above paper.

limits of the characteristic of the over-all intermediate-frequency system have the effect of holding the maximum peaks of noise to a value equal to the highest signal peaks obtainable in the absence of the disturbance. Furthermore, the existence of high noise peaks in the presence of modulation can result in the momentary assumption by the instantaneous intermediate frequency of values outside of the region to which the system is normally responsive. Thus the output signal will appear to be chopped by the higher noise peaks, and as a consequence its energy content will be considerably reduced.

The above effects are, of course, present in systems using limiters and have already been discussed in greater detail by Crosby.⁶

Distortion Reduction

One of the chief benefits which can be realized through the use of negative feedback is the reduction of non-linear distortion products generated in the forward branch of the system. While the distortion in properly designed amplifiers is sufficiently low for many purposes, cases frequently arise in which the requirements are much more severe. In an amplifier which is to handle several channels in a high grade multiplex system, the distortion products should be of the order of 60 decibels below the fundamental of the output. This degree of excellence is most readily obtained by using negative feedback.

In radio systems designed for multiplex service it is of equal importance that the distortion level be kept at a correspondingly low level if crosstalk is to be avoided. It is therefore of interest to inquire into the manner in which distortion is modified in the present feedback system.

An analysis of the effect of feedback upon distortion is given in Appendix A. If the transmitter is modulated with a signal wave $S = S(t)$ so that its instantaneous frequency becomes

$$\omega + \rho_1 S \quad (23)$$

then, in the presence of non-linearity in the receiver, the output of the system can be written as a power series in the variable frequency term $\rho_1 S$. Thus for the first three orders we shall have

$$\sigma = \alpha AB [b_1 \rho_1 S + b_2 (\rho_1 S)^2 + b_3 (\rho_1 S)^3]. \quad (24)$$

If feedback is applied without altering the modulation level at the transmitter it is shown that the above series becomes

$$\sigma_F = \alpha AB \left(\frac{b_1}{F} \rho_1 S + \frac{b_2}{F^3} (\rho_1 S)^2 + \frac{1}{F^4} \left[b_3 - \frac{2b_2^2}{b_1} \left(\frac{F-1}{F} \right) \right] (\rho_1 S)^3 \right). \quad (25)$$

When the feedback factor F is large this can be written

$$\sigma_F = \alpha AB \left(\frac{b_1}{F} \rho_1 S + \frac{b_2}{F^3} (\rho_1 S)^2 + \frac{1}{F^4} \left[b_3 - \frac{2b_2^2}{b_1} \right] (\rho_1 S)^3 \right). \quad (26)$$

Upon increasing the modulation by the factor F so as to restore the original level of the fundamental, the output becomes

$$\sigma_{F'} = \alpha AB \left(b_1 \rho_1 S + \frac{1}{F} \left[b_2 (\rho_1 S)^2 + \left(b_3 - \frac{2b_2^2}{b_1} \right) (\rho_1 S)^3 \right] \right). \quad (27)$$

Second order distortion products are reduced with respect to the fundamental level by the feedback factor. Third (and higher) order products are modified to an extent depending upon the relative values of the distortion coefficients and the amount of feedback. If, as can readily be the case when a balanced detecting system is used,

$$b_3 \gg \frac{2b_2^2}{b_1} \quad (28)$$

third order products are reduced in the same manner as those of second order. In any case, by applying sufficient feedback a point will be reached where a given increment in feedback will produce a corresponding reduction in all distortion products.

Equation (25) shows that the greatest improvement in distortion is obtained if the modulation level is not increased when feedback is applied. The large reductions result partly from feedback and in part from the fact that the system is operating at reduced percentage of modulation. Under this condition there is no improvement in background noise ratio, though the noise increment which takes place during modulation is diminished; see (10) and (11). Depression of both noise and distortion, but with greater emphasis upon the reduction of the latter, can be effected by raising the modulation level by an amount somewhat less than the feedback factor. This procedure has already been discussed in connection with (22) which gives the resulting noise-to-signal power ratio. Under similar conditions we have, from (25),

$$\sigma_{F''} = \alpha AB \left(\frac{b_1}{F_2} \rho_1 S + \frac{1}{F_1} \left[\frac{b_2}{F_2^3} (\rho_1 S)^2 + \frac{1}{F_2^4} \left(b_3 - \frac{2b_2^2}{b_1} \right) (\rho_1 S)^3 \right] \right) \quad (29)$$

when the feedback factor is large.

Equations (22) and (29) are most readily interpreted by means of Fig. 3, which illustrates the manner in which the receiver output is modified as the feedback is increased.

It is assumed that a constant signal level is maintained by an increase in modulation level up to a point corresponding to the factor F_1 . Beyond this point the modulation remains fixed while the feedback

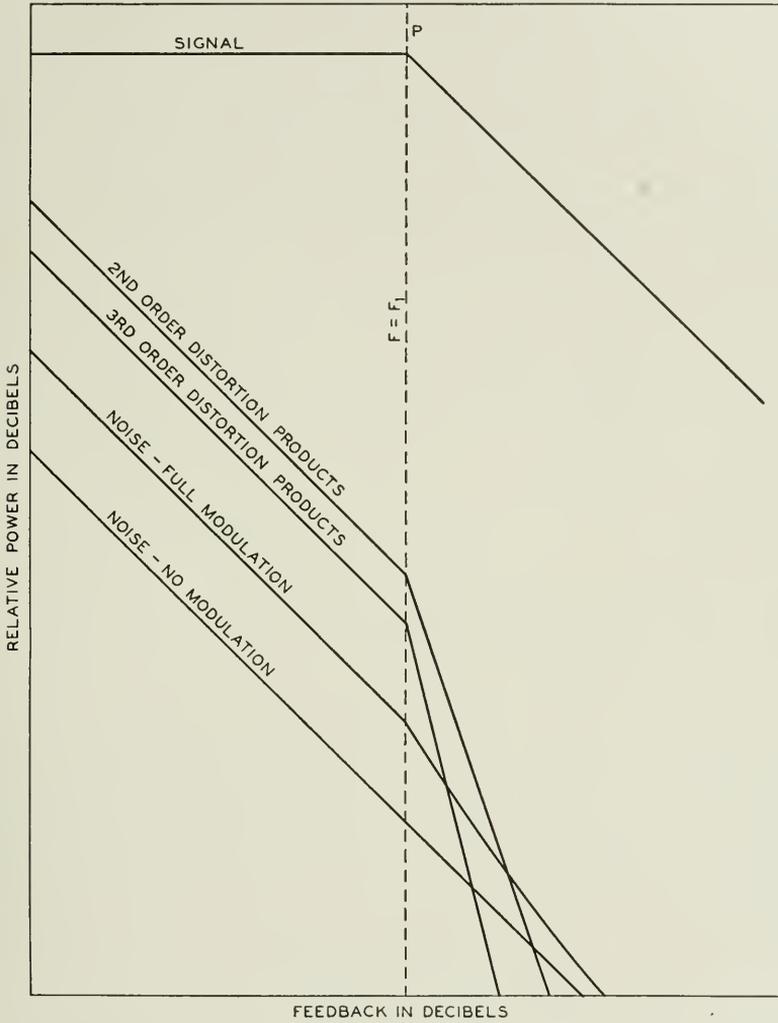


Fig. 3—Theoretical manner in which components of receiver output are modified by feedback. Modulation level at transmitter is assumed to be increased by the feedback factor up to point P , and subsequently to remain fixed.

factor is increased to F_1F_2 . Since signal and noise levels have been expressed in terms of power, distortion levels are similarly expressed.

These levels, to an arbitrary decibel scale, have been plotted against decibels of feedback, given by the expression

$$10 \log F^2 = 20 \log F.$$

Balanced detection and the fulfillment of condition (28) have been assumed.

Over the region in which a constant signal output is maintained by increasing the modulation level, noise and distortion levels decrease in accordance with the feedback. The noise level during modulation continues to exceed the background noise by 4 decibels, assuming an initial frequency shift equal to the highest signal frequency to which the system is responsive.

Beyond the point at which the feedback factor has reached the value F_1 , the modulation level at the transmitter is held constant. A further increase in feedback brings about a corresponding decrease in the effective percentage of modulation for the system, causing the signal level to fall in similar fashion. Distortion products now fall off still more rapidly with respect to the signal, so that an increase in feedback amounting to 1 decibel improves the second and third order distortion ratios by 2 and 3 decibels, respectively.

The ratio of signal to background or non-signaling noise remains fixed in this region in spite of the reduction in effective modulation. This ratio is that which would be obtained in a limiter system in which the same high-frequency band is transmitted. The noise increment, however, is diminished by the additional feedback and is made to approach zero.

By suitable choice of the variables F_1 and F_2 it is possible to proportion the benefits of feedback in the most advantageous manner. Thus if noise is of more consequence than distortion, modulation would be increased to the full extent of the feedback; if distortion is of primary concern, as it might well be in a multiplex system, operation as indicated in Fig. 3 would be preferable.

EXPERIMENTAL RESULTS

Description of Equipment

Experimental confirmation of the principles which have been outlined has been obtained with the aid of a laboratory system shown schematically in Fig. 4. This arrangement provided a transmitter, receiver, and source of disturbance all under local control. The transmitter operated at a carried frequency of 20 megacycles. This was frequency-modulated by means of a circuit basically similar to that

described by Travis.⁷ Tube noise voltage appearing at the output of a high gain radio frequency amplifier supplied the high-frequency disturbance.

At the receiver an oscillator similar to that at the transmitter served to beat down the incoming wave to an intermediate frequency of 438 kilocycles. This was applied to a three-stage amplifier having substantially uniform gain over a band of 50 kilocycles, and thence delivered to a balanced frequency detector. In addition to signal voltage, automatic-frequency-control potentials were derived from the detectors. Both were carried back to the local oscillator, but in order to permit independent control of the amount of feedback their respective paths were kept separate. In this way full frequency control could be had even with signal feedback reduced to zero.

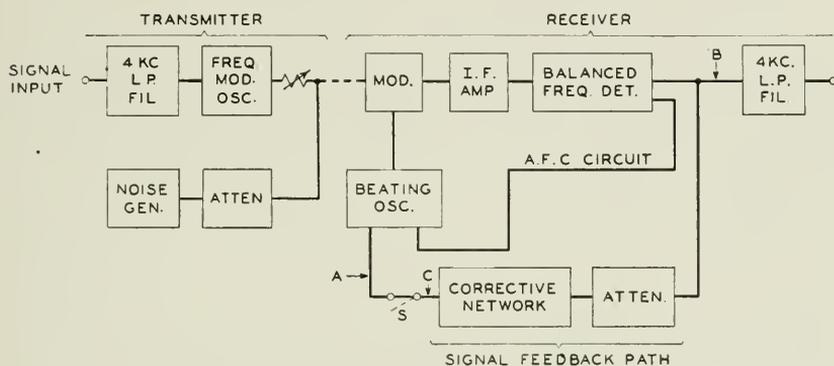


Fig. 4—Schematic of experimental feedback system.

Details of the frequency detector and feedback connections are shown in Fig. 5. The conversion system derives its characteristics from anti-resonant circuits L_1C_1 and L_2C_2 , double-winding coils being used to isolate the rectifier anodes from the plate battery. One circuit is tuned to a frequency 15.4 kilocycles above the intermediate carrier frequency and the other to a corresponding point below, their characteristics intersecting at a point where the gain is approximately one half of the peak value. Detection takes place in linear rectifiers D_1 and D_2 . By means of the arrangement shown, signal potentials are impressed upon the grids of amplifiers A_1 and A_2 , while frequency-control voltage appears across condensers C_3 , C_4 . This voltage becomes zero when the receiver is correctly tuned and appears with proper polarity to

⁷ Charles Travis, "Automatic Frequency Control," *Proc. I. R. E.*, vol. 23, pp. 1125-1141, October 1935.

produce correction of the frequency of the local oscillator in case of slow drifts in the frequency of either oscillator.

The use of a conversion system having peaks separated by an amount considerably exceeding the greatest frequency deviation is the result of a compromise between the readily adjustable and high impedance anti-resonant type of load circuit and others which, though more linear in their characteristics, lead to much lower gain in the conversion stage. While a peak separation of 14 kilocycles would have sufficed in view of the limitations of the transmitter, a considerably greater peak separation without corresponding increase in modulation was used. As a result that portion of the circuit characteristic actually embraced by the modulated intermediate-frequency wave

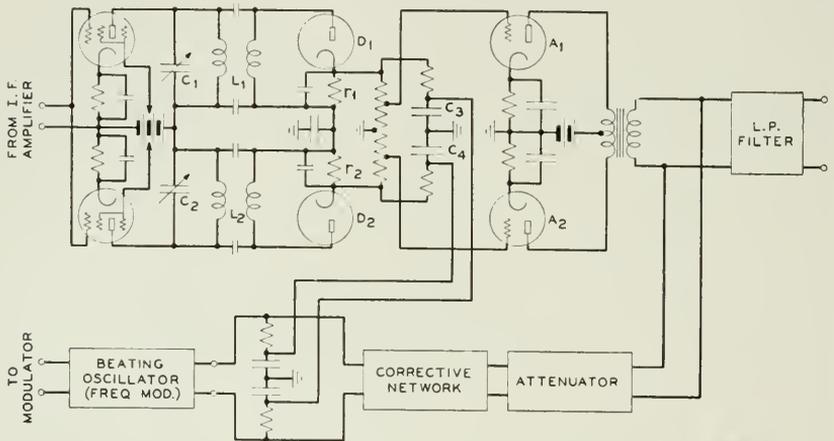


Fig. 5—Details of balanced frequency detector and feedback connections.

presented a much better approximation of a straight line than would have been possible with minimum peak separation. The penalty for adjusting the circuits in this manner is merely a loss in detecting efficiency and not an impairment of the signal-to-noise ratio. This can readily be overcome by additional audio-frequency amplification.

The signal-frequency feedback path includes an attenuator for adjusting the feedback and a corrective network for preventing singing around the feedback loop. Frequency control and feedback paths are finally combined at the modulation terminals of the local oscillator.

The necessity for the inclusion of a corrective network to modify the transmission characteristics of the feedback path is evident from Fig. 6. This shows the measured gain and phase characteristics of the receiver

alone, viewed as a voice-frequency network between points *A* and *B* in Fig. 4. This was obtained by applying signal frequencies to the modulation terminals of the beating oscillator and making observations at point *B* with switch *S* open, proper termination being provided at both sides of the break. The unmodulated transmitter served, in effect, as the beating oscillator during this measurement. At the lower signal frequencies the phase is practically 180 degrees as indicated by (4) with $x_1 = 0$. As the signal frequency is increased the phase

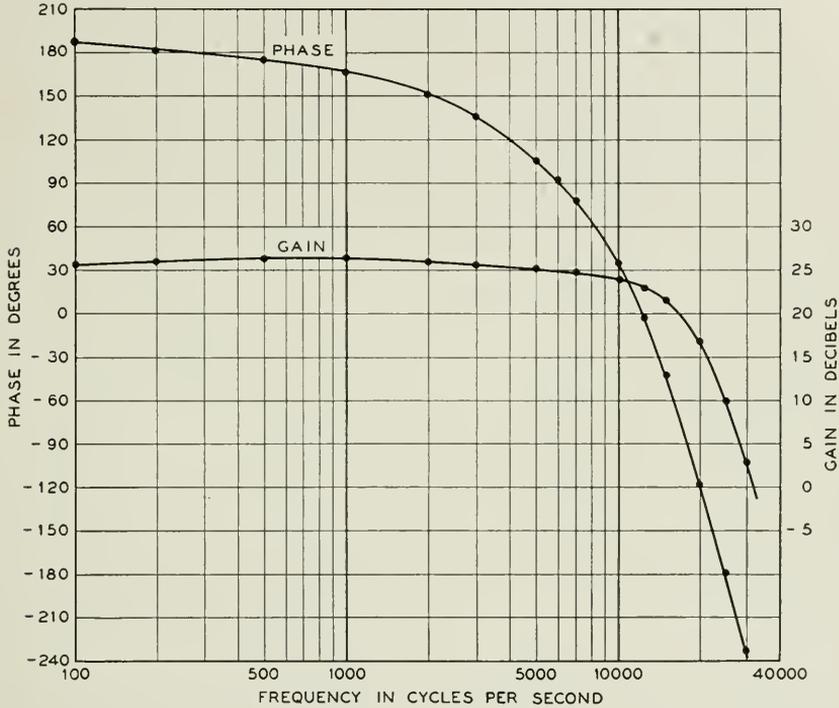


Fig. 6—Phase and gain characteristics of receiver measured between points *A* and *B* of Fig. 4 with switch *S* open. Transmitter in operation but not modulated.

is progressively shifted from this value. Except for that produced by the output transformer, the shift takes place within the intermediate-frequency amplifier and conversion circuits. Its magnitude is a measure of the slope of the phase-frequency characteristic of the intermediate frequency system.

The existence of positive gain at a point of zero phase shows that singing would occur if feedback connections were made directly to the beating oscillator. It was therefore necessary to reduce the gain below

unity at the point of zero phase. This was accomplished by including in the feedback path a network designed by R. L. Dietzold. The gain-frequency characteristic of this network is shown in Fig. 7. The modified loop characteristics as measured between points *A* and *C* with switch *S* open, and with the attenuator set for an 8-decibel loss, are given in Fig. 8. Full feedback is applied only over a band extending to 4 kilocycles, so that the range of frequencies applied to the transmitter and delivered to the listener must be restricted to this figure.

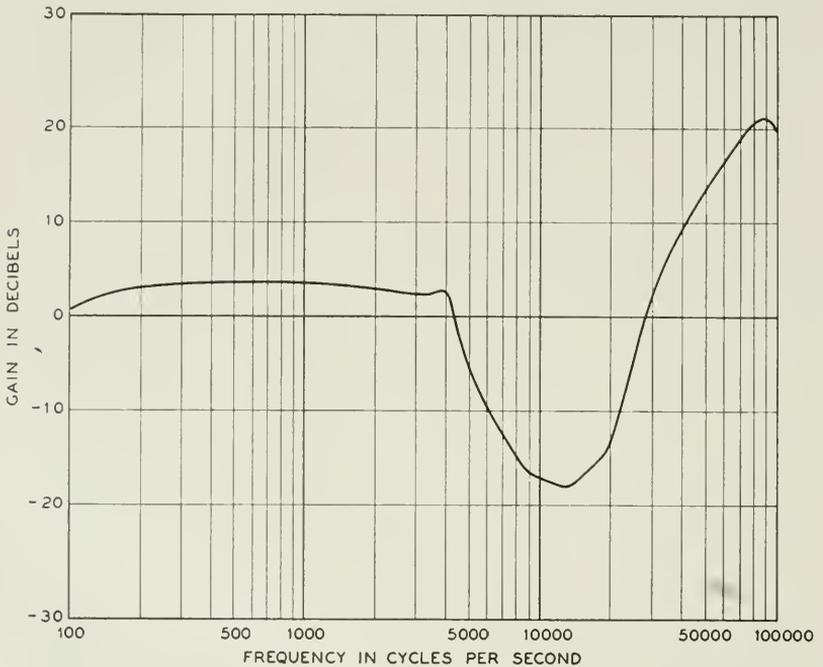


Fig. 7—Gain-frequency characteristic of corrective network inserted in signal feedback path.

The limit of stable feedback which can be realized is indicated by the difference between the loop gain within the useful band and that at the frequency corresponding to zero phase.

Distortion Measurements

The manner in which distortion levels at the output of the receiver were observed to vary with feedback is depicted in Figs. 9 to 12. In each case the modulation level for zero feedback was such as to shift the frequency of the transmitter ± 7 kilocycles at the rate of 1000

cycles per second. Figure 9 shows the effect of increasing the modulation in proportion to the feedback so as to maintain a constant output level for the fundamental. Both second and third harmonics tend to be reduced in proportion to the feedback, the improvement in third-harmonic level being 23.5 decibels for 25-decibel feedback. Failure to realize full reduction of the second harmonic is the result of distortion beginning to manifest itself in one or the other of the modulated oscillators. At the point of 25-decibel feedback the transmitter and

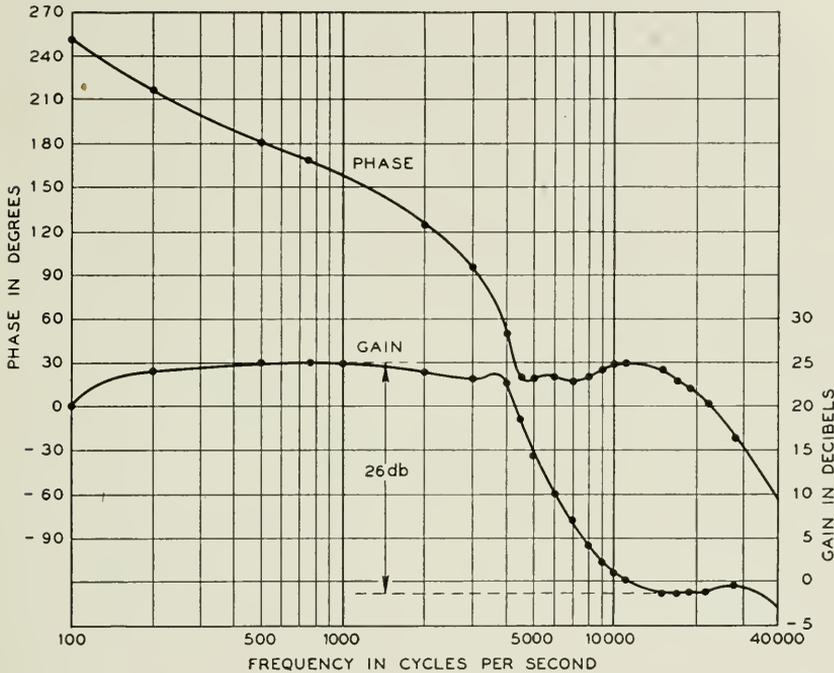


Fig. 8—Gain and phase characteristics of complete feedback loop including corrective network. Measured between points *A* and *C* of Fig. 4 with switch *S* open.

beating oscillator were being modulated to the extent of ± 124.5 and ± 117.5 kilocycles, respectively.

The curves of Fig. 10 were obtained by maintaining a constant fundamental level up to the point of 15-decibel feedback and then allowing the modulation level at the transmitter to remain constant thereafter. The results correspond rather closely with the theoretical curves of Fig. 3 and show the very rapid decrease in distortion which takes place when the modulation level remains unaltered; see (25). A more extreme example of this method of operation is shown in Fig. 11 where

modulation was left at its initial value. Harmonic levels soon reached a point beyond which they could not be measured accurately.

In a practical system the loss in signal level resulting from operation in this manner could easily be overcome by the addition of a low-distortion audio-frequency amplifier at the output of the receiver. This amplifier might well embody negative feedback of the more usual type.

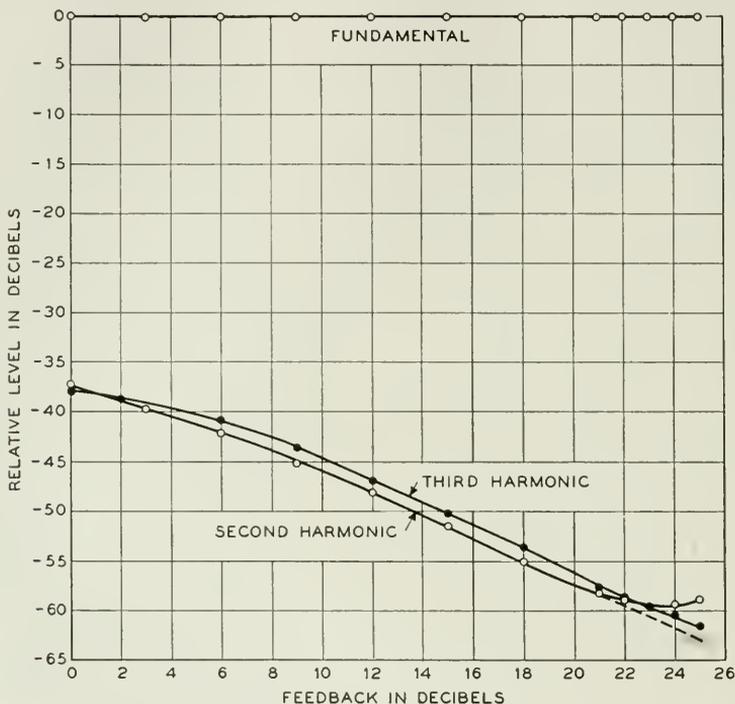


Fig. 9—Effect of feedback upon receiver distortion. Fundamental level kept constant by increasing transmitter modulation in proportion to the feedback. Modulation with no feedback = ± 7 kilocycles at 1000 cycles per second.

A composite of these distortion measurements is given in Fig. 12. Harmonic levels are plotted in decibels below the fundamental and are indicative of the improvements brought about by feedback. If it is assumed that any loss in signal is compensated by additional audio-frequency amplification, the fundamental level would be represented in all cases by the axis of abscissae.

Noise Measurements

In Fig. 13 are given the results of a series of observations of receiver output noise versus amount of feedback for a number of high-frequency

disturbance levels. Measurements were made in the absence of modulation and hence are indicative of the manner in which background noise is modified by feedback. The signal level indicated is that which could be maintained at low noise levels by increasing the modulation in proportion to the feedback, and is not significant for observations falling within or close to the shaded area, as will be explained subsequently.

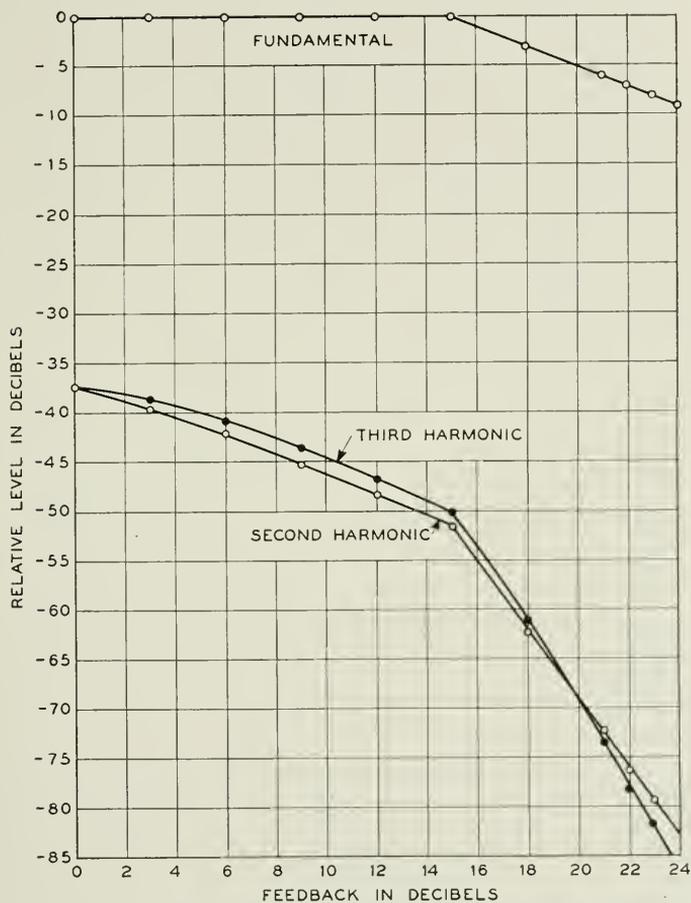


Fig. 10—Effect of feedback upon receiver distortion. Conditions same as indicated for Fig. 9 up to 15-decibel feedback; modulation held constant thereafter.

The lowest noise level shown is that generated within the receiver while the higher levels were produced by disturbances introduced from the noise generator. The relative magnitude of the effective carrier and disturbing voltages at the grids of the amplitude detector is indi-

cated on each curve. This was obtained in the following manner: With the transmitter turned off the noise attenuator was adjusted until the introduced disturbance produced the same value of rectified current as that observed when the carrier alone was applied. This determined the input level from the noise generator which produced equal root-

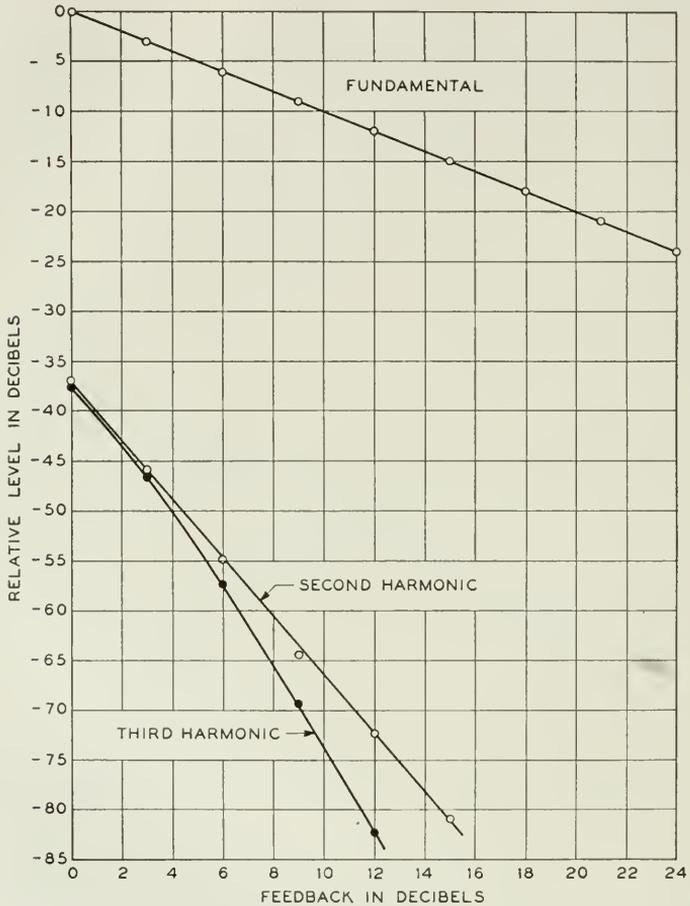


Fig. 11—Effect of feedback upon receiver distortion. Modulation held to a constant value of ± 7 kilocycles.

mean-square values of intermediate-frequency carrier and disturbance. Since at very low inputs from the noise generator the net intermediate-frequency disturbance was determined partly by tube noise generated within the receiver, a curve of output noise without feedback versus

input from the noise generator was obtained. In the region where the effect of receiver tube noise was evident the assumption of a linear relationship between disturbance level and output noise permitted correction of the curve so that equivalent disturbance levels could be related to any setting of the noise attenuator, or to the receiver output noise level without feedback.

The signal-to-noise ratios at the output of the receiver without feedback are considerably higher than the corresponding ratios of carrier and disturbance levels existing at the amplitude detectors. This is the

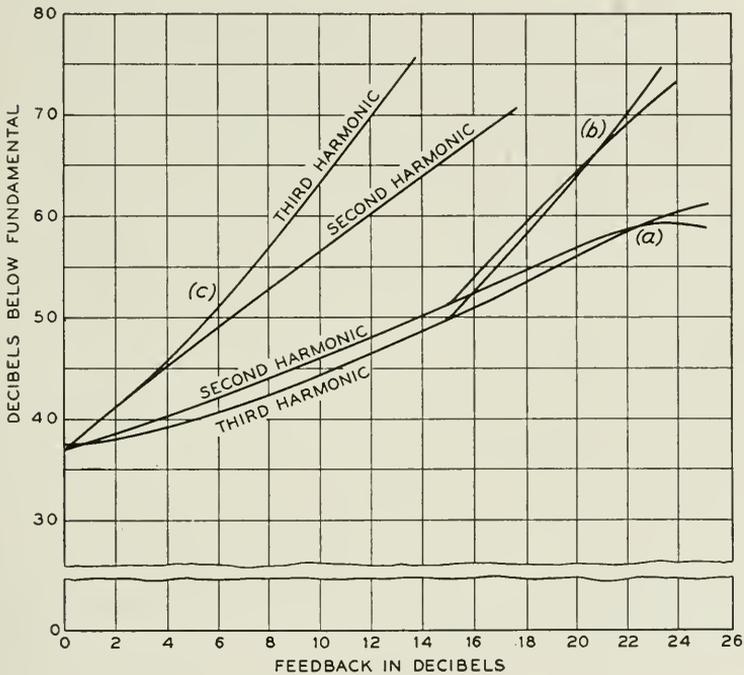


Fig. 12—Composite of data from Figs. 9 to 11 expressing ratios of harmonic levels to fundamental level. Curves (a) from Fig. 9, (b) from Fig. 10, and (c) from Fig. 11.

result of two factors. The intermediate-frequency wave is modulated to the extent of ± 7 kilocycles while the range of frequencies appearing at the output terminals of the receiver is limited to 4 kilocycles. Hence at the output of the balanced detector the noise level in the absence of modulation is 9.6 decibels below that which would be observed at the output of an amplitude-modulation system. Furthermore the admittance characteristic of the complete intermediate-frequency system is such that the effective disturbing voltage delivered to

the amplitude detectors is 11 decibels greater than that admitted by an amplitude modulation system having the minimum intermediate-frequency band width of 8 kilocycles.

Aural observation of the character of the output noise showed that, excluding the shaded area in Fig. 13, the normal characteristics of fluctuation noise are preserved as feedback is applied. Upon crossing

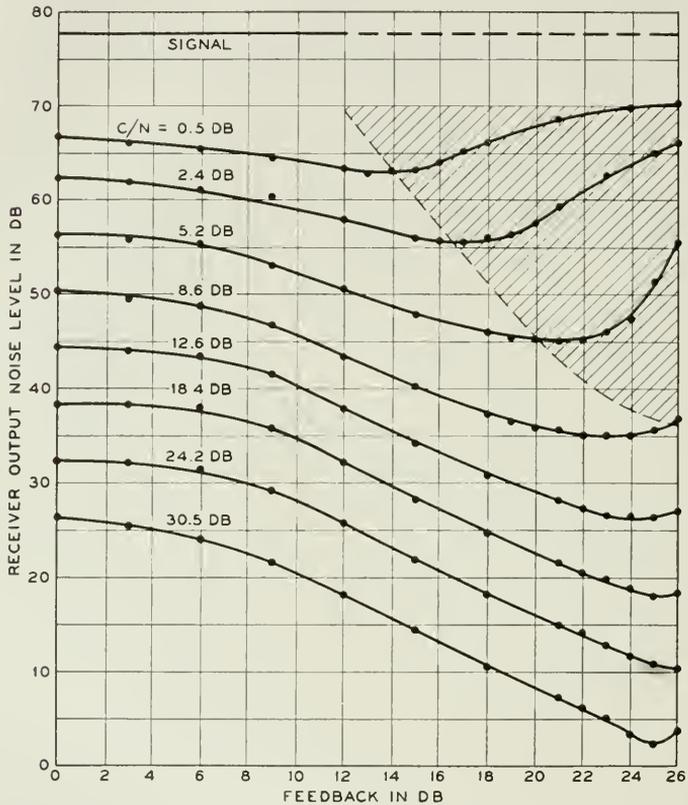


Fig. 13—Effect of feedback upon receiver output noise level for various amounts of high-frequency disturbance. Ratio of root-mean-square values of carrier and disturbance is shown on each curve. Shaded area indicates region of "crackling" in the absence of modulation.

the boundary of the shaded area the noise becomes punctuated with intermittent clicks which increase in rapidity and violence as the feedback factor is raised, giving rise to what can be described as "crackling." After passing through a region of maximum turbulence the noise gradually assumes the nature of a much higher level of fluctuation noise.

The region embracing the appearance of the above phenomenon is also characterized by a marked reduction in signal level. At high modulation levels "cracking" begins at a somewhat lower disturbance level than is necessary to initiate it in the absence of modulation. The initial effect is to impart a roughness to tone modulation. Further increase in disturbance level produces a rapid depression of the signal, so that it soon becomes submerged in noise. The manner in which this depression takes place is shown in Fig. 14. The signal, produced by 1000-cycle modulation was measured by means of a highly selective

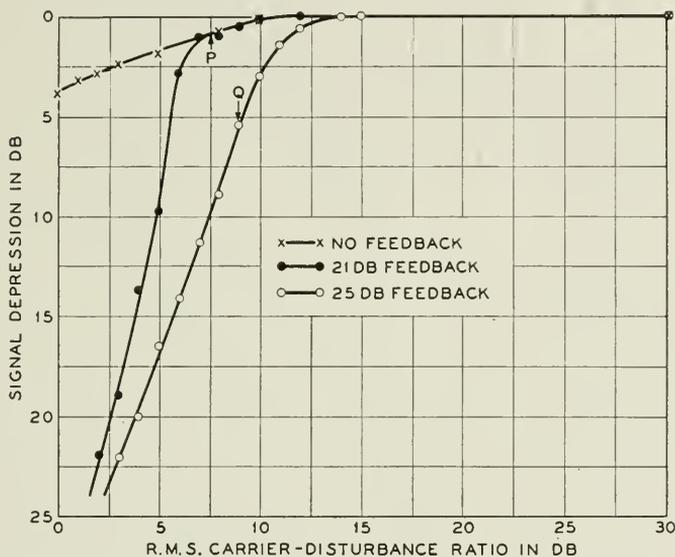


Fig. 14—Depression of output signal by the disturbance. Modulation: ± 7 kilocycles for no feedback, ± 78.5 kilocycles for 21-decibel feedback, and ± 124.5 kilocycles for 25-decibel feedback. Points *P* and *Q* denote incidence of cracking in the absence of modulation, for 21- and 25-decibel feedback, respectively.

analyzer so that observations could be carried well below the general noise level.

The point at which depression of the signal begins coincides with the appearance of roughness in the output tone resulting from the momentary suppression of the signal by the higher noise peaks. A further increase in disturbance level increases the number of peaks per second which rise above the critical value, and the energy content of the signal is rapidly diminished. The point at which faint cracking could first be detected in the absence of modulation is indicated on each curve.

The signal-to-noise ratios obtained with zero and with 25 decibels of

feedback are shown in Fig. 15. These are plotted against the ratio of root-mean-square carrier and disturbance levels at the end of the intermediate-frequency channel. Signal levels were measured in the presence of the disturbance so as to take into account the depression of the signal, while noise levels were observed in the absence of modulation.

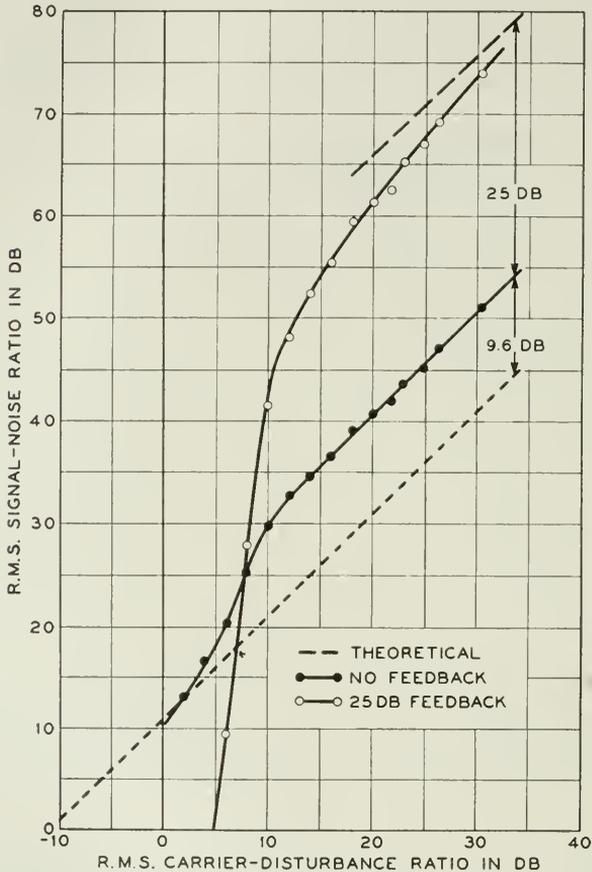


Fig. 15—Output signal-to-noise ratio vs. carrier-disturbance ratio with and without feedback. Modulation ± 124.5 kilocycles for 25-decibel feedback and ± 7 kilocycles for no feedback.

The improvement resulting from the application of feedback is given by the difference between the two curves and approaches the theoretical improvement at the low noise levels. The curve obtained with feedback exhibits a rather sharp break when the ratio of carrier to effective disturbance is in the vicinity of 10 decibels. Experimental

data⁶ published by Crosby indicates that in the case of fluctuation noise the ratio of the maximum peak amplitude to the root-mean-square value is about 13 decibels. The corresponding figure in the case of a sine wave is 3 decibels. Hence equality of carrier peak amplitude and the highest peaks of the disturbance obtains when the ratio of their root-mean-square values is 10 decibels. With feedback this condition appears to define a fairly critical disturbance level above which the output signal-to-noise ratio is very rapidly diminished. Crosby has shown⁶ that with systems employing amplitude limiters a similar condition marks the point beyond which the noise improvement realized at the lower disturbance levels is soon lost. This point he has termed the "threshold of noise improvement."

A less sharply defined break also occurs in the curve expressing noise conditions in the absence of feedback. This is the result of a progressive destruction, at the higher disturbance levels, of the balancing out of amplitude effects in the push-pull detector which is realized when the noise is low.

While direct comparison of the feedback system with an actual amplitude modulation system was not possible with the equipment used, it is thought that a comparison based upon theoretical considerations may be of interest. The procedure is as follows: The noise ratios shown in Fig. 15 for the system without feedback are, for disturbances below the threshold value, 9.6 decibels in excess of those which would be realized in a fully modulated amplitude system. A dotted line, displaced from the linear portion of the measured curve by this amount, is shown. The abscissae of the dotted curve do not represent the true carrier-disturbance ratio which would obtain in the amplitude system for the reason that, ideally, the intermediate-frequency amplifying system would have a band width of but 8 kilocycles. In such a system the signal-to-noise ratio would be equal to the carrier-disturbance ratio except at the very high noise levels. Hence the intercept of the dotted line with the axis of abscissas marks the point of equal carrier and disturbance levels in this system. The difference of 11 decibels between this point and the zero point on the scale as drawn measures the amount by which the disturbance at the rectifiers in the experimental system exceeds that which would be found in the ideal amplitude system.⁸ Consequently, if it is desired to relate the data of Fig. 15 to the disturbance ratio which would exist at the input to the detector in the amplitude system, and hence to the signal-to-noise ratio in that

⁸ Comparison of the areas under idealized curves representing the square of the transmission through the intermediate-frequency systems in the two cases indicates a difference of 10.1 decibels.

system, it is merely necessary to displace the experimental curves to the right by 11 decibels.

Figure 16 shows such a comparison between the feedback system adjusted to give 25 decibels of feedback, and an ideal amplitude modula-

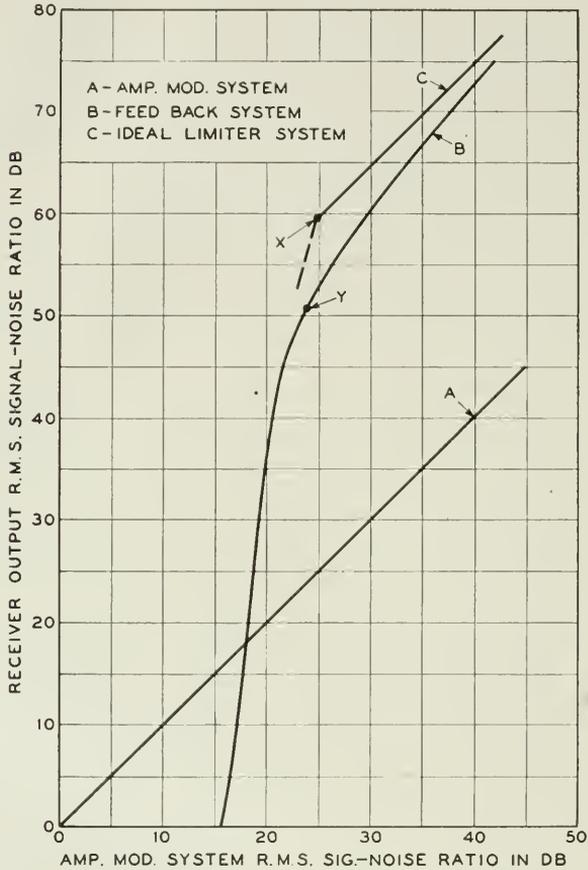


Fig. 16—Theoretical comparison of signal-to-noise ratios obtained at 25-decibel feedback (curve *B*) with amplitude-modulated system (curve *A*) and ideal limiter system with deviation ratio = 31.1 (curve *C*). Point *X* = threshold of noise improvement for limiter system. Point *Y* = point where "crackling" first became evident in the presence of ± 124.5 -kilocycle modulation.

tion system. There is also included a curve showing the theoretical performance which would be approached by a frequency-modulation system using amplitude limitation. Transmitted band width and audio-frequency response equal to that used in the experimental feed-

back system have been assumed. This corresponds to a deviation ratio of $124.5 \text{ kilocycles} \div 4 \text{ kilocycles} = 31.1$, resulting in a theoretical noise deduction of 34.6 decibels at low disturbance levels. The threshold of noise improvement, indicated at the point x , is located at a point where the peak signal-to-noise ratio in the amplitude system is equal to the square root of the deviation ratio.⁶ This factor takes account of the higher disturbance level in the intermediate-frequency channel of the wide-band system. Assuming a factor of 10 decibels between maximum peak and root-mean-square values of fluctuation noise this corresponds to a root-mean-square signal-to-noise ratio of 24.9 decibels in the amplitude system.⁹ In the feedback system the point at which crackling was first observed in the presence of modulation is shown at Y . This coincides very closely with the theoretical threshold of noise improvement in the limiter system.

CONCLUSIONS

It has been shown that the application of negative feedback to a frequency-modulation system affords a means for effecting large reductions in both noise and receiver distortion. The theoretical analyses of these effects, while they have been simplified to an extent which makes them inadequate to cover all conditions which can be encountered in practice, are adequately substantiated by the observed performance of the experimental system within the limitations of the theory.

Substantial benefits are realized only when the amount of feedback is large, and when the disturbance level is not too great. In common with frequency-modulation systems employing amplitude limitation a large reduction in noise must be paid for by increasing the band width of the transmitted wave. While the principles involved in the two systems are quite different, their performance as regards noise modification, both at high and at low levels of disturbance, exhibits striking similarities. The ability to reduce distortion is, on the other hand, a feature found only in the feedback system.

ACKNOWLEDGMENT

The writer wishes to acknowledge his indebtedness to the following of his colleagues: To Dr. H. W. Bode for his investigations of the problem of stability in feedback systems, and to Mr. R. L. Dietzold for the design of the stabilizing network which was used; to Mr. W. R. Bennett whose unpublished work on basic frequency-modulation prob-

⁹ In the absence of more exact information regarding the performance of a system using this large a deviation ratio the portion of this curve below the threshold has been omitted.

lems has been of great value; and to Messrs. E. A. Krauth and O. E. DeLange for assistance in the experimental work. Reference has already been made to the theoretical work of Dr. J. R. Carson.

APPENDIX A

Analysis of Distortion Reduction

Assume that the transmitter is frequency-modulated with a signal wave which we shall represent by the symbol $S = S(t)$. Then the instantaneous frequency of the transmitter will be

$$\omega_1 + \rho_1 S. \quad (30)$$

The instantaneous phase of the transmitted wave is the integral of this expression and the voltage delivered to the input of receiver can be written

$$A \cos \left(\omega_1 t + \rho_1 \int_0^t S dt + \phi_1 \right). \quad (31)$$

Designating the low-frequency voltage delivered at the output of the receiver as $\sigma = \sigma(t)$ the result of feeding back a portion of $k\sigma$ of the output so as to frequency-modulate the local oscillator is the wave

$$B \cos \left(\omega_2 t + \rho_2 \int_0^t k\sigma dt + \phi_2 \right). \quad (32)$$

Application of these two waves to the modulator produces the intermediate-frequency product ¹⁰

$$\alpha AB \cos \left[\omega_0 t + \rho_1 \int_0^t S dt - \rho_2 \int_0^t k\sigma dt + \phi_0 \right] \quad (33)$$

where

$$\begin{aligned} \omega_0 &= \omega_1 - \omega_2 \\ \phi_0 &= \phi_1 - \phi_2. \end{aligned}$$

Terms in the above which involve the integral sign represent phase angles which vary with time. Hence we shall rewrite (33) more compactly

$$\alpha AB \cos [\omega_0 t + \theta(t) + \phi_0]. \quad (34)$$

It has been shown by Carson and Fry³ that the process of detecting a frequency-modulated wave is, in effect, its differentiation. Since the high-frequency wave itself exhibits the integral of the signal wave, see

¹⁰ This expression constitutes a more general form of (4).

(30), it can be reasoned that a differentiation process is necessary for the recovery of the signal itself.

Differentiation of the argument of the cosine term in (31) yields the instantaneous frequency (rate of change of phase with respect to time) of the received wave given by (30). Now it can be shown that with a strictly linear frequency detector, the low-frequency output is proportional to the response of the conversion system at the instantaneous frequency. The recovered signal is, therefore, proportional to the variable part of the instantaneous frequency and hence to the time derivative of the variable phase term in the original wave.

In the case of non-linearity in the characteristic of the frequency detector the output can be expressed, to a sufficiently close degree of approximation, as a power series in the derivative of the phase term $\theta(t)$. Hence the output of the receiver can be written in the form

$$\sigma(t) = \alpha AB \sum_n b_n \left[\frac{d}{dt} \theta(t) \right]^n \tag{35}$$

$$= \alpha AB \sum_n b_n [\rho_1 S - k\rho_2 \sigma]^n \tag{36}$$

where the coefficients b_n are based upon the transfer admittance characteristic of the receiver.

What is now desired is the relationship between σ and S . This can be expressed in the general form

$$\sigma = \alpha AB \sum_n c_n [\rho_1 S]^n. \tag{37}$$

Equations (36) and (37) can now be equated. Replacing σ in the right-hand side of (36) by the series (37) we shall have

$$\begin{aligned} & c_1 \rho_1 S + c_2 (\rho_1 S)^2 + c_3 (\rho_1 S)^3 + \dots \\ &= b_1 [\rho_1 S - k\alpha AB\rho_2 (c_1 \rho_1 S + c_2 (\rho_1 S)^2 + c_3 (\rho_1 S)^3 + \dots)] \\ & \quad + b_2 [\rho_1 S - k\alpha AB\rho_2 (c_1 \rho_1 S + c_2 (\rho_1 S)^2 + c_3 (\rho_1 S)^3 + \dots)]^2 \\ & \quad + b_3 [\rho_1 S - k\alpha AB\rho_2 (c_1 \rho_1 S + c_2 (\rho_1 S)^2 + c_3 (\rho_1 S)^3 + \dots)]^3 \\ & \quad + \dots \end{aligned} \tag{38}$$

After expanding, coefficients of like powers of $\rho_1 S$ can be equated. Then solving for the first three orders of c_n we find

$$c_1 = \frac{b_1}{1 + \alpha b_1 k AB \rho_2} = \frac{b_1}{1 - \mu\beta} \tag{39}$$

$$c_2 = \frac{b_2}{(1 - \mu\beta)^3} \tag{40}$$

$$c_3 = \frac{b_3}{(1 - \mu\beta)^4} - \frac{2\alpha AB k \rho_2 b_2^2}{(1 - \mu\beta)^5} \tag{41}$$

where

$$\mu = \alpha AB b_1 \quad \text{and} \quad \beta = -k\rho_2.$$

Inserting these values in (37) and writing $(1 - \mu\beta) = F$, the receiver output becomes, with feedback,

$$\sigma_F = \alpha AB \left[\frac{b_1}{F} \rho_1 S + \frac{b_2}{F^3} (\rho_1 S)^2 + \left(\frac{b_3}{F^4} - \frac{2b_2^2}{b_1} \cdot \frac{F-1}{F^5} \right) (\rho_1 S)^3 \right]. \quad (42)$$

When $F \gg 1$ this can be written

$$\sigma_F = \alpha AB \left[\frac{b_1}{F} \rho_1 S + \frac{b_2}{F^3} (\rho_1 S)^2 + \frac{1}{F^4} \left(b_3 - \frac{2b_2^2}{b_1} \right) (\rho_1 S)^3 \right]. \quad (43)$$

Without feedback we have

$$\sigma = \alpha AB [b_1 \rho_1 S + b_2 (\rho_1 S)^2 + b_3 (\rho_1 S)^3]. \quad (44)$$

APPENDIX B

Analysis of Noise Reduction

In the following analysis it is assumed that the amplitude of the disturbance producing the noise is sufficiently small compared with that of the incoming signal wave so that the principle of superposition will apply. Hence the manner in which the effect of a single disturbing component is modified by feedback will first be developed. Then the effect of a disturbance consisting of a continuous spectrum is derived by direct summation.

Consider first the case where there are impressed upon the grid of the modulator the incoming wave and the local oscillator voltage¹¹ as defined by (31) and (32), plus a single disturbing component

$$Q \cos [(\omega_1 + \omega_n)t + \phi_n]. \quad (45)$$

Then the intermediate-frequency product will be

$$\alpha AB \cos \left[\omega_0 t + \rho_1 \int_0^t S dt - \rho_2 \int_0^t k \sigma dt \right] + \alpha BQ \cos \left[(\omega_0 + \omega_n)t - \rho_2 \int_0^t k \sigma dt + \phi_n \right]. \quad (46)$$

For simplicity assume that the intermediate-frequency amplifier and conversion circuit have the ideal transfer admittance characteristic

$$Y(\omega) = a_0 + a_1(\omega - \omega_0). \quad (47)$$

¹¹ Arbitrary phase constants will be omitted from these expressions since they do not affect the final result.

Then all derivatives of Y with respect to ω above the first are zero, and the steady-state response is equal to its response at the instantaneous frequency of the applied wave. Hence after conversion we shall have

$$\alpha AB[a_0 + a_1(\rho_1 S - \rho_2 k\sigma)] \cos \left[\omega_0 t + \int_0^t (\rho_1 S - \rho_2 k\sigma) dt \right] + \alpha BQ[a_0 + a_1(\omega_n - \rho_2 k\sigma)] \cos \left[(\omega_0 + \omega_n)t - \int_0^t \rho_2 k\sigma dt + \phi_n \right]. \quad (48)$$

Application of (48) to a linear amplitude detector will yield a low-frequency output proportional to its amplitude. The amplitude factor is readily calculated for the case where $AB \gg BQ$. For if

$$X \cos x + Y \cos y = Z \cos z$$

then

$$Z = \sqrt{X^2 + Y^2 + 2XY \cos(x - y)}$$

and when $X \gg Y$

$$Z \doteq X + Y \cos(x - y). \quad (49)$$

Hence the output of the linear detector will be

$$\gamma \left(\alpha AB[a_0 + a_1(\rho_1 S - \rho_2 k\sigma)] + \alpha BQ[a_0 + a_1(\omega_n - \rho_2 k\sigma)] \times \cos \left[\omega_n t - \rho_1 \int_0^t S dt + \phi_n \right] \right). \quad (50)$$

The term $\alpha\gamma ABa_0$ represents direct current. Assuming that this is not fed back to the local oscillator we can then write

$$\sigma = A'[a_1(\rho_1 S - \rho_2 k\sigma)] + Q'[a_0 + a_1(\omega_n - \rho_2 k\sigma)] \cos \xi \quad (51)$$

where

$$A' = \alpha\gamma AB \\ Q' = \alpha\gamma BQ \\ \xi = \left(\omega_n t - \int_0^t \rho_1 S dt + \phi_n \right).$$

Solving for σ

$$\sigma = \frac{1}{1 + a_1 A' k \rho_2} \left[1 + \frac{a_1 Q' k \rho_2 \cos \xi}{1 + a_1 A' k \rho_2} \right]^{-1} \times [A' a_1 \rho_1 S + Q'(a_0 + a_1 \omega_n) \cos \xi]. \quad (52)$$

If $Q' \ll A'$

$$\sigma \doteq \frac{1}{F} \left[1 - \frac{a_1 Q' k \rho_2 \cos \xi}{F} \right] [A' a_1 \rho_1 S + Q' (a_0 + a_1 \omega_n) \cos \xi] \quad (53)$$

where $F = 1 + a_1 A' k \rho_2 = 1 - \mu \beta$ as before. Finally, neglecting terms in Q'^2 , we have

$$\sigma = \frac{1}{F} \left[A' a_1 \rho_1 S + Q' \left(a_0 + a_1 \omega_n - \frac{A' a_1^2 \rho_1 k \rho_2 S}{F} \right) \cos \xi \right]. \quad (54)$$

The first term is the recovered signal while the remaining terms represent noise. Both signal and noise are modified by feedback. If we let

$$\rho_1 S = \Delta \omega \cos pt \quad (55)$$

then the noise becomes

$$\begin{aligned} \frac{Q'}{F} \left[(a_0 + a_1 \omega_n) - \frac{1}{F} (A' a_1^2 k \rho_2 \Delta \omega \cos pt) \right] \\ \times \cos (\omega_n t - x \sin pt + \phi_n). \end{aligned} \quad (56)$$

By means of the Jacobi expansions this can be put in the form

$$\begin{aligned} \frac{Q'}{F} \sum_{m=-\infty}^{\infty} \left[(a_0 + a_1 \omega_n) - \frac{A' a_1^2 k \rho_2 m p}{F} \right] J_m(x) \\ \times \cos [(\omega_n - mp)t + \phi_n] \end{aligned} \quad (57)$$

where $J_m(x)$ is the Bessel coefficient of the first kind.

Now let it be assumed that the disturbance consists of a very large number of sinusoidal components of like amplitude Q , random phase, and uniformly distributed along the frequency scale. The summation of this series of voltages can be represented by the very general expression

$$f(t) \cos [\omega t + \phi(t)]. \quad (58)$$

So long as $f(t)$, the equivalent amplitude of the high-frequency disturbance, is small compared with the carrier amplitude A , the approximation (49) will be valid and the total output noise can be obtained by summing up the effects of the individual elements which constitute the disturbance.

The effect of a single disturbing element is given by (57). Any term of this expression can be made to have the frequency q if m and ω_n are so chosen that

$$\omega_n = mp \pm q. \quad (59)$$

Then for each value of m in (57) there will be available values of ω_n

to satisfy both of the conditions expressed by (59). The total effect is obtained by summing the output power resulting from each contribution since the original elements have random phases. If r_2 is the resistance of the output circuit the total power of frequency q becomes

$$\frac{Q'^2}{2r_2F^2} \left(\sum_{m=-\infty}^{\infty} \left[a_0 + a_1(mp + q) - \frac{A'a_1^2k\rho_2mp}{F} \right]^2 J_m^2(x) + \sum_{m=-\infty}^{\infty} \left[a_0 + a_1(mp - q) - \frac{A'a_1^2k\rho_2mp}{F} \right]^2 J_m^2(x) \right). \quad (60)$$

This is readily evaluated with the aid of tables appended to an earlier paper.¹² The result is

$$\frac{Q'^2}{r_2F^2} \left[a_0^2 + \frac{a_1^2\Delta\omega^2}{2F^2} + a_1^2q^2 \right]. \quad (61)$$

The amplitude factor Q remains to be defined. If N^2 is the mean disturbing power per unit band width in the vicinity of the carrier frequency and r_1 the resistance of the input circuit, the peak amplitude of any element is defined by the relation

$$N^2d\omega = \frac{Q^2}{2r_1}. \quad (62)$$

Thus the power associated with each element becomes differentially small, and if the value so obtained is entered into (61) there is obtained the output noise power contained in a band extending from q to $q + dq$. Then we shall have

$$dW = \frac{2N^2r_1(\alpha\gamma B)^2}{r_2F^2} \left[a_0^2 + \frac{a_1^2\Delta\omega^2}{2F^2} + a_1^2q^2 \right] dq. \quad (63)$$

The total noise power in a band extending to a limiting frequency q_a is

$$P_n = \int_0^{q_a} dW = \frac{2N^2r_1(\alpha\gamma B)^2}{r_2F^2} \left[a_0^2 + \frac{a_1^2\Delta\omega^2}{2F^2} + \frac{a_1^2q_a^2}{3} \right] q_a. \quad (64)$$

The corresponding signal power is

$$P_s = \frac{(\alpha\gamma AB)^2}{2r_2F^2} a_1^2\Delta\omega^2. \quad (65)$$

¹² J. G. Chaffee, "The Detection of Frequency Modulated Waves," *Proc. I. R. E.*, vol. 23, pp. 517-540, May 1935.

Survey of Magnetic Materials and Applications in the Telephone System

By V. E. LEGG

The great diversity of magnetic characteristics demanded by telephone apparatus, and the large number of available magnetic materials propose intricate problems in the choice of materials and design of apparatus to attain greatest efficiency and economy. The present paper undertakes to evaluate magnetic materials in relation to apparatus needs. After a review of the earlier developments, the materials now available are listed, together with data on technical characteristics and raw materials costs. The advantages of various materials for specific applications are described. The scope of possible further improvements in magnetic materials is surveyed.

HISTORICAL

TWENTY years ago, the telephone system used primarily iron, together with a small amount of silicon iron, for applications requiring soft magnetic materials, and carbon, tungsten or chromium steel for permanent magnet applications. The permalloys¹ were already fairly thoroughly developed by 1920 in what is now the Bell Telephone Laboratories, and 78.5 permalloy² shortly attained commercial recognition for its utility as a continuous loading material for submarine telegraph cables.³ This and other nickel-iron alloys were soon serving in many types of telephone relays, and in various coils where the designs could be profitably modified to adapt them to the new materials. Upon the development of commercial means for embrittling and pulverizing permalloy, this material was soon in extensive use because it offered improved characteristics over the compressed powdered iron core material previously in use. Redesigns of filter and loading coils have introduced such economies that practically all these coils made by the Western Electric Company have until recently employed compressed powdered permalloy cores.⁴

A desire to reduce the losses in a-c. apparatus arising from eddy currents in magnetic parts led to the development of permalloys of higher

¹ H. D. Arnold & G. W. Elmen, *Jour. Frank. Inst.* 195, 621 (1923).

² The approximate chemical compositions of the various materials herein discussed are given in Tables I and II.

³ O. E. Buckley, *Jour. A. I. E. E.* 44, 821 (1925).

⁴ W. J. Shackelton & I. G. Barber, *Trans. A. I. E. E.* 47, 429 (1928).

resistivity, containing several per cent chromium or molybdenum.⁵ The problems of embrittlement and pulverization of molybdenum permalloy were also successfully solved. This material has been recently adopted for manufacture of filter and loading coil cores,⁶ in which material of higher resistivity is especially advantageous.

Attempts to decrease the losses due to hysteresis led to the discovery of the nickel-iron-cobalt alloys—the perminalvars. A molybdenum-perminvar was perfected for use in the continuous loading of submarine telephone cable.⁷

The large economic advantages promised by improvements in soft magnetic materials confined much of the earlier work to this field. However, within the last 20 years new permanent magnet materials have been discovered here and abroad which offered radical improvements in this direction. Such materials have been introduced in telephone apparatus wherever found advantageous.

CHARACTERISTICS OF AVAILABLE MATERIALS

The number of different magnetic materials in use is quite large on account of the various combinations of properties required for special applications and on account of a multiplicity of trade names. For the present purpose, an abbreviated listing is given of typical materials covering the whole range of magnetic properties, particularly those of interest to the telephone system. A compilation of representative data is given in Tables I and II.

The fundamental property which distinguishes a ferromagnetic material is that when it is subjected to a magnetic field it develops magnetic flux considerably larger than similarly attained in air. The magnetizing forces of interest in telephone apparatus range from less than 10^{-3} to upwards of 10^3 oersteds, and the flux densities from less than 1 to 30,000 gauss or more. The relation of flux density B to magnetizing force H for important materials as first magnetized is given on a logarithmic scale in Fig. 1. The ratio of B to H is the permeability, which can be read on the diagonal scale⁸ in the figure. It is evident that the initial permeability μ_0 and the maximum permeability μ_m vary over a wide range from the hard magnet steels to the softest magnetic alloys. For commercial materials, 4-79 Mo-permalloy gives the largest initial permeability—around 22,000, and 78.5 permalloy gives the largest maximum permeability—about 105,000.

⁵ G. W. Elmen, *Jour. Frank. Inst.* 207, 583 (1929).

⁶ O. E. Buckley, *Jour. Applied Phys.* 8, 40 (1937).

⁷ G. W. Elmen, *Elec. Engg.* 54, 1292 (1935).

⁸ Scale due to Aiken; *Jour. Applied Phys.* 8, 470 (1937).

TABLE I
SOFT MAGNETIC MATERIALS IN SOLID OR SHEET FORM

Other		Per Cent Composition				Raw Cost ¢/lb.	Typical Anneal	Material	μ_0 Initial Permeability	μ_m Maximum Permeability	Saturation $4\pi I_s$ gauss	H_∞ Hysteresis Loss erg/cm ³	B_r Residual gauss	H_c Coercive Force oersteds	ρ Resistivity microhm- centimeter 30	Curie Tem- perature θ	Hysteresis Coefficient $a \times 10^6$	σ Density gm./cm. ³
		Mn	Mo	Ni	Co													
3 C, 2 Si						<1	800° (Centi- grade) 900° Pot 1180° H ₂ +880° H ₂	Cast Iron Magnetic Iron Magnetic Iron H ₂ Purified	— 250 25,000	600 5,500 275,000	— 21,500 21,500	20,000 5,000 13,000 300	5,500 13,000 13,000	4.6 1.0 0.05	10 10 10	770° C. 770	50	7.88 7.88
0.5 Si 4 Si						7* 8*	800° Pot 800° Pot	0.5 Si-Iron (Field) 4 Si-Iron (Trans- former)	250 400	3,700 6,700	21,000 20,000	4,500 3,500	12,800 12,000	0.8 0.5	18 60	760 690	— 120	7.7 7.5
9½ Si, 5½ Al						3	Cast	Sendust	30,000	120,000	10,000	100	5,000	0.05	80	—	—	7.1
0.2 Cu	0.6			99.0 45 50		35 17 18	1000° Pot 1100° Pot Long 1200° H ₂	Nickel 45 Permalloy Hypernik	110 2,700 3,000	600 23,000 70,000	6,100 16,500 16,500	3,000 1,200 220	3,600 8,000 7,000	3.4 0.3 0.04	8 45 35	360 440 500	100 0.4 —	8.85 8.17 8.25
5 Cu 3.8 Cr	0.6 1.0 0.6			78.5 78.5 78.5		28 27 29	1000°+Quench 900° Pot 1000° Pot	78.5 Permalloy Mumetal 3.8-78.5 Cr- Permalloy	9,000 7,000 10,000	105,000 80,000 40,000	10,700 8,500 8,000	200 200 200	6,000 6,000 4,500	0.05 0.05 0.05	16 25 65	580 — 455	0.2 — 0.3	8.60 8.58 8.56
15 Cu	0.6 1.0 12.5	4		79 71 80		32 29 40	1000° Pot 1100° H ₂ 800° H ₂	4-79 Mo- Permalloy 1040 Alloy 12.5-80 Mo- Permalloy	22,000 40,000 9,000	72,000 100,000 →1, over room temperature range	8,500 6,000	200 200	5,000 2,500	0.05 0.014	55 —	460 290 40	0.05 — —	8.72 8.76 8.9
2 V				99 50 49		136 69 73	1000° Pot 900° Pot 800° Pot	Cobalt Permendur 2 V-Permendur	70 800 800	240 5,000 4,500	18,000 21,500 21,000	2,000 12,000 6,000	5,000 14,000 14,000	10 2.0 2.0	9 7 26	1120 1000 980	30 — 1.0	8.9 8.3 8.2
	0.6 0.6 0.6			45 70 45		50 35 57	1000°+425° Bake 1000°+425° Bake 1000°+425° Bake	45-25 Perminvar 7-70 Perminvar 7.5-45-25 Mo- Perminvar	365 850 550	1,800 4,000 3,800	15,500 12,500 10,300	4,000 — 2,600	3,300 2,400 4,300	1.4 0.6 0.6	19 16 80	715 650 540	0.01 0.06 0.1	8.6 8.6 8.66

* Approximate finished cost for 14 mil sheet.

Note: Values for μ_0 , μ_m , H_∞ , B_r , H_c , and a are subject to considerable variations, depending on purity of materials, composition, type of heat treatment, and final condition of mechanical stress.

TABLE II
PERMANENT MAGNET STEELS

Per Cent Composition			Raw Cost ¢/lb.	Typical Heat Treatment		Material	μ_0 Initial Permeability	μ_r Reversible Permeability	Saturation $4\pi I_{\infty}$ gauss	B_r Residual	H_c Coercive Force oersted	Energy Product ($B \cdot H$) _{max.} 0.2×10^6	B for ($B \cdot H$) = Maximum gauss	Resistivity ρ microhm-centimeter	Curie Temperature θ 750° C.	Density σ gm./cm. ³
Other	Cr	Ni		Co	Fe											
0.6 C, 0.8 Mn				98.8	800° C. Water	—	75	—	21,000	10,000	50	0.2	6,900	20	750	7.8
0.6 C, 0.4 Mn	1			98	800° Oil	—	—	—	—	9,800	50	0.2	6,900	—	—	—
1 C, 0.4 Mn	3			96	840° Oil	—	—	31	—	9,700	65	0.3	6,100	38	—	7.7
5 W, 1 C	3.5			94	840° H ₂ O	—	—	—	—	10,800	60	0.3	7,000	30	—	8.0
7 W, 0.5 Mn			36	52	940° Oil	—	7	9.4	19,000	9,500	220	0.9	6,000	27	700	8.3
6.7 Ti, 3.7 Al			18	45	Cast	650°	3	3.8	7,100	7,100	780	2.0	4,100	65	—	7.3
13 Al			29	58	1200° Oil	600°	4	5	11,600	6,000	550	1.4	3,500	60	750	7.1
14 Al			25	60	1200° Oil	600°	5	5	—	6,500	500	1.4	4,400	60	—	7.0
12 Al			20	63	1200° Oil	600°	—	4	—	7,300	430	1.4	4,500	60	—	7.1
10 Al, 6 Cu			17	54.5	1200° Oil	600°	—	4	—	7,200	540	1.6	4,400	—	—	—
17 Mo			12	71	1300° Oil	700°	8	12	17,000	10,500	250	1.1	6,500	45	780	8.4
60 Cu			20	20	1000° Oil	600°	3	3	5,000	3,400	390	0.5	1,800	—	—	—
41 Cu			24	35	1050° Oil	600°	4	4	8,600	5,300	440	1.0	3,400	38	850	8.7
77 Pt			23	\$400	1200° Oil	—	1.1	1.1	—	4,500	2,600	3.8	2,500	50	—	—
2 Fe ₂ O ₃ + 1 Fe ₃ O ₄ + 1 Co ₂ O ₃				25¢	950° Vacuum + Magnetize at 500°	—	1.7	1.7	—	1,800	600	—	—	10 ¹²	350	3.8

⁹ Neumann, Bücher & Reinboth, *Z. f. Metallkunde* 29, 173 (1937).

¹⁰ Dannöhl & Neumann, *Zeit. f. Metallkunde* 30, 217 (1938).

Note: Values for μ_0 , μ_r , B_r , H_c and $(B \cdot H)_{max.}$ are subject to considerable variations, depending on purity of materials, composition, and heat treatment.

TABLE III

Curve Number	Material	Typical Heat Treatment (Temperature in Degrees C.)	Initial Permeability μ_0	Maximum Permeability μ_m	Saturation $4\pi I_{\infty}$	Hysteresis Loss at Saturation W_{∞}	Residual E_r	Coercive Force H_c	Resistivity ρ
					<i>gauss</i>	<i>erg/cm.³</i>	<i>gauss</i>	<i>oerst</i>	<i>micro-ohm-cm.</i>
1	Soft Magnetic Materials								
2	Magnetic Iron	900 Anneal	250	5,500	21,500	5,000	13,000	1.0	10
3	4 Per Cent Silicon Iron	800 Anneal	400	6,700	20,000	3,500	12,000	0.5	60
4	45 Permalloy (Ni, Fe)	1100 Anneal	2,700	23,000	16,500	1,200	8,000	0.3	45
5	78.5 Permalloy (Ni, Fe)	1000 + Air Quench	9,000	105,000	10,700	200	6,000	0.05	16
6	4-79 Mo-Permalloy (Mo, Ni, Fe)	1000 Anneal	22,000	72,000	8,500	200	5,000	0.05	55
7	2 V Permenidur (V, Co, Fe)	800 Anneal	800	4,500	24,000	6,000	14,000	2.0	26
	45-25 Perminvar (Co, Ni, Fe)	1000 + 425 Bake	365	1,800	15,500	4,000	3,300	1.4	19
8	Magnet Steels					Energy Product $(E_r I_{\infty}) \times 10^{-4}$			
9	3 Per Cent Chrome	840 Quench	10	100	—	0.34	9,700	65	38
10	Honda (C, W, Co, Fe)	940 Quench	7	—	19,000	0.9	9,500	220	27
11	Mishima (Al, Ni, Fe)	Quench + 600 Bake	4	16	11,600	1.4	6,000	550	60
12	Remalloy (Mo, Co, Fe) Oxide (Fe, Co, O ₂)	Quench + 700 Bake 950 Vacuum	12 1.7	30 —	— —	1.1 —	10,500 1,800	250 600	45 10 ¹²

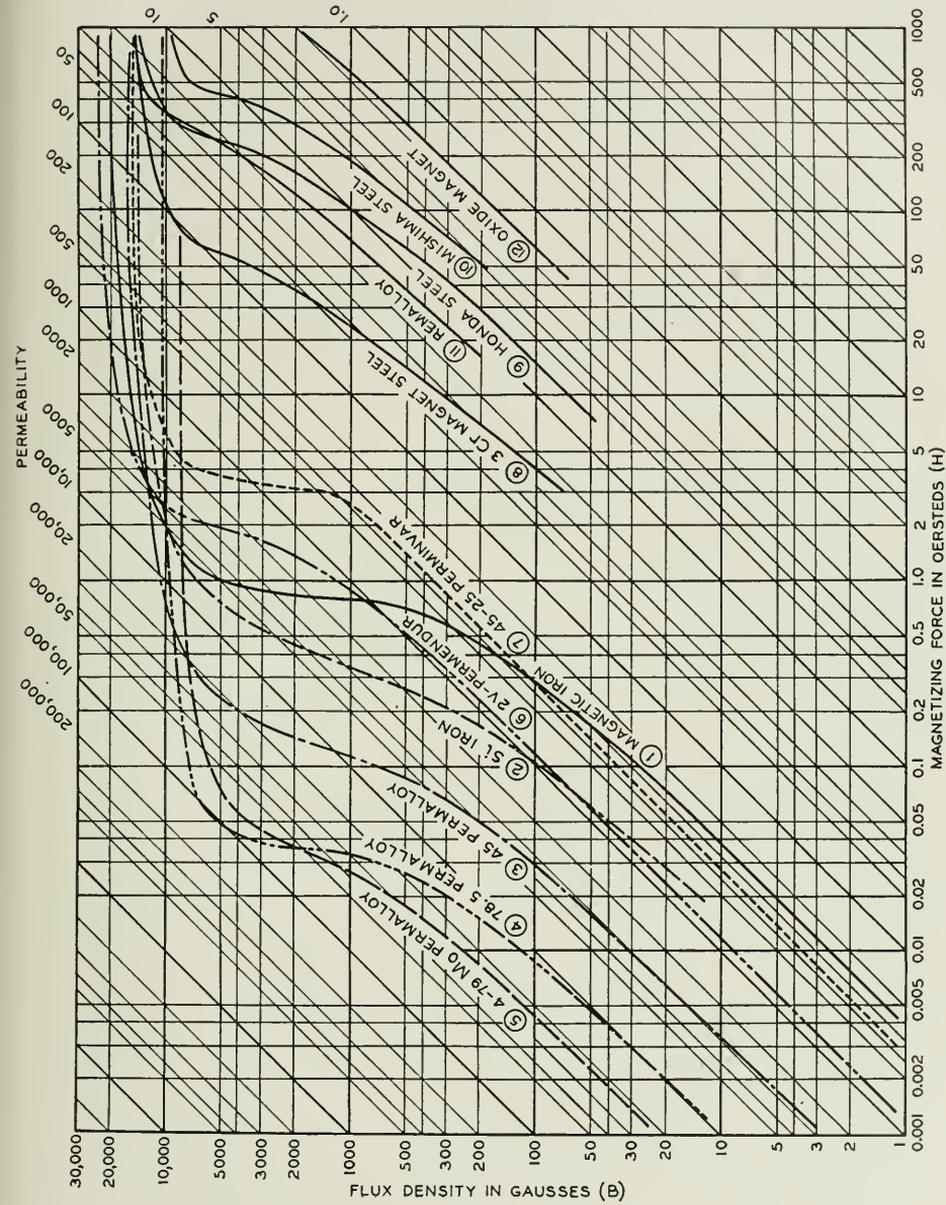


Fig. 1—Magnetization and permeability curves for important magnetic materials (diagonal scale for permeability values).

The saturation flux density for the soft magnetic materials can be read from these curves, and it is listed in the adjacent Table III in the column headed $4\pi I_\infty$. The pre-eminence of 2 V-permendur is noteworthy.

Further important magnetic properties are obtained from the hysteresis loop. This gives the B, H relationship for a material which has already been magnetized up to a peak value H_m . The flux density remaining after the removal of a very large magnetizing force is the residual B_r , and the reverse magnetizing force necessary to bring the flux to zero is the coercive force H_c . The area of the B, H loop when the peak magnetizing force is very large gives the maximum energy W_∞ dissipated by hysteresis in the material when it is carried from positive magnetic saturation to negative, and back again. Table III gives values of W_∞ , B_r and H_c . The low values of these properties for 4-79 Mo-permalloy are noteworthy. Among soft magnetic materials, iron and 2 V-permendur have high residual and coercive force, properties which are occasionally useful.

Permanent magnet materials should have large values of B_r and H_c , although a sacrifice of residual can be more or less compensated by an increase in coercive force. A more fundamental criterion of permanent magnet quality is the peak energy product $(B \cdot H)_{\max}$, obtained from the demagnetizing section of the hysteresis loop.¹¹ Values of this product are given for several magnet steels in Table III. Mishima steel is seen to be foremost in this regard, with remalloy nearly as good.

The last column in Table III gives the resistivity ρ of each material. A high resistivity such as obtained with molybdenum additions to permalloy and permivar is desirable in suppressing eddy current losses for alternating current applications. Eddy current losses can also be suppressed by proper subdivision of the material, but this method becomes costly with fine subdivision.

For apparatus depending upon the tractive force of a magnet, the high flux density properties of materials are most important. Since these do not show up clearly in Fig. 1, an accompanying Fig. 2 has been prepared in which the $(B - H)$ scale is quadratic, and thus proportional to tractive force. The relative merits of various materials for such applications are seen by inspection of these curves. 2 V-permendur is outstanding at high flux densities, iron and 45 permalloy at intermediate, and 4-79 Mo-permalloy at low flux densities.

The use of a purifying hydrogen anneal has been shown by Cioffi¹² to increase the ease of saturating iron and most magnetic alloys. The

¹¹ S. Evershed, *J. I. E. E.*, London, 58, 780 (1920); 63, 725 (1925).

¹² P. P. Cioffi, *Phys. Rev.* 39, 363 (1932).

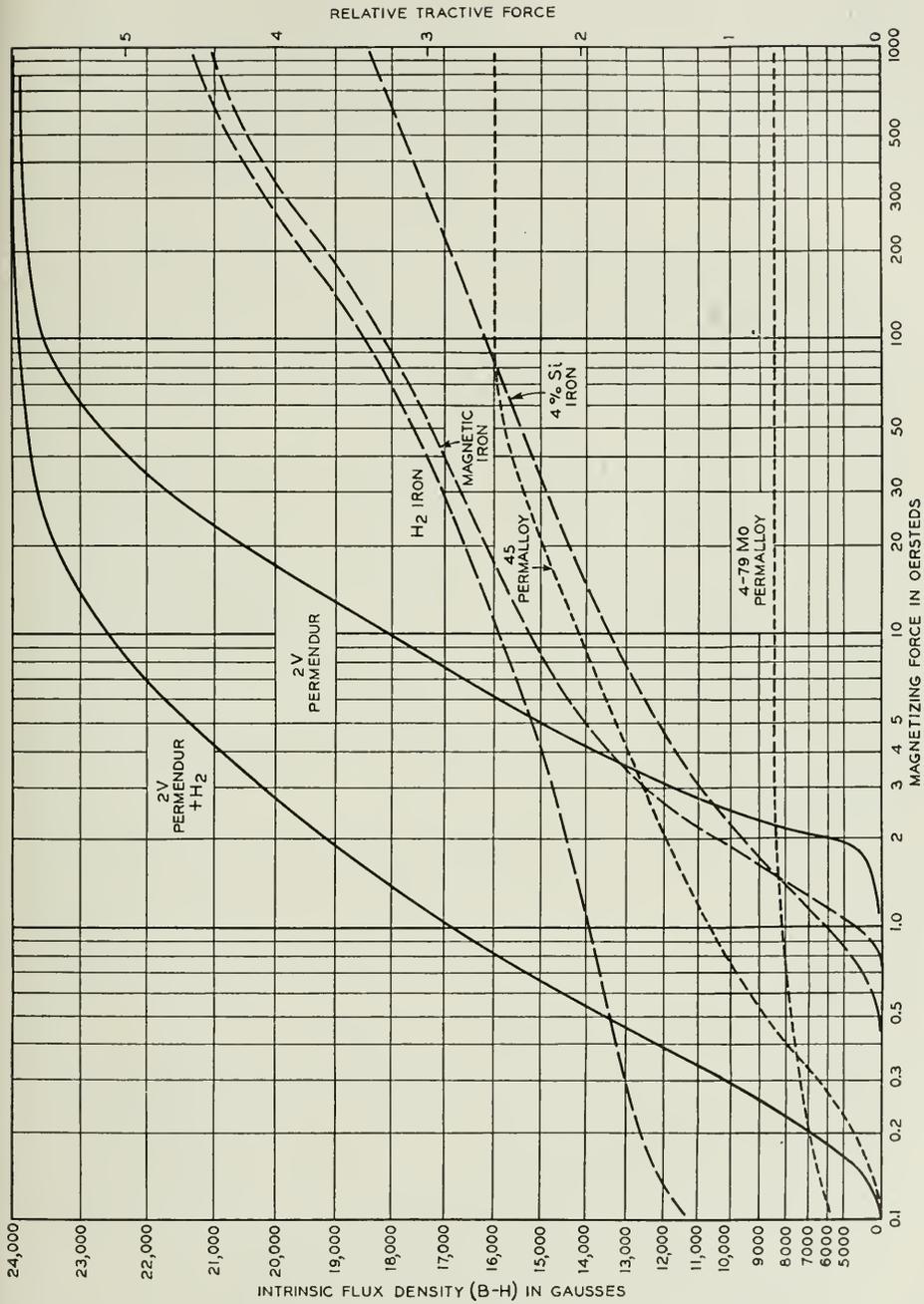


Fig. 2—Magnetization curves plotted to show the relative tractive force for important magnetic materials.

improvements for iron and 2 V-permendur are shown in Fig. 2. Improvements in 4-79 Mo-permalloy and in 45 permalloy are similarly possible.⁷ A hydrogen purified alloy having very nearly the composition of the latter material is produced commercially under the name of Hipernik.¹³ A material similar to hydrogen purified 4-79 Mo-permalloy has been produced under the trade name "1040" Alloy.¹⁴

ADAPTATIONS FOR A-C. USES

For a-c. applications, eddy current, hysteresis and residual losses are generally important, in addition to the permeability. The materials for these applications have to be laminated or used in powdered form to limit eddy current losses. Such modifications generally affect the magnetic qualities, often adversely, through the introduction of impurities or stresses, or through the concentration of magnetic flux in parts of the material adjacent to air-gaps.

Many a-c. applications involve flux densities so low that the permeability does not rise to more than perhaps 10 per cent above μ_0 . Similarly, many applications involve operation at frequencies low enough to insure that the a-c. permeability is approximately equal to the permeability obtained by d-c. methods. (At higher frequencies, eddy current shielding reduces the a-c. permeability below the d-c. value, as discussed in Sec. 4b.) In these ordinary cases, the a-c. losses in a material working at a flux density B_m and frequency f , are conveniently described in terms of the increased resistance R per unit inductance L of the winding encircling the material.¹⁵ These resistance increments are:

1. Hysteresis, $R_h/L = aB_m\mu f$, where a is the hysteresis loop constant. At these low flux densities, the permeability coefficient

$$\lambda = (\mu - \mu_0)/\mu_0 B_m.$$

It is related to the hysteresis coefficient to a good approximation by the equation $\lambda = 3a\mu/8$.

2. Residual, $R_r/L = c\mu f$, where c is the residual loss constant.

3. Eddy current, $R_e/L = e\mu f^2$, where e is the eddy current loss constant. It is proportional to the square of the laminar thickness, or particle diameter, and inversely proportional to the resistivity of the material.

Table IV gives easily attainable values of the constants of importance in a-c. applications for several typical materials. Data on

¹³ T. D. Yensen, *Jour. Frank. Inst.* 199, 333 (1925).

¹⁴ H. Neumann, *Arch. f. tech. Messen* 4, 42T, 168 (1934).

¹⁵ V. E. Legg, *B. S. T. J.* 15, 39 (1936).

other materials are also listed in Table I. Manufacturing tolerance limits are frequently somewhat less favorable than here indicated. In the section on laminated cores in Table IV, the eddy current coefficient e has been computed for several common sheet thicknesses. Values for other thicknesses can be calculated as proportional to the square of the thickness ratios.

TABLE IV
ALTERNATING CURRENT LOSS DATA ON SHEET AND POWDERED MATERIALS

Material	Size	Permeability μ_0	Resistivity $\rho \times 10^6$	Hysteresis Coeff. $a \times 10^6$	Residual Coeff. $c \times 10^6$	Eddy Current Coeff. $e \times 10^9$
Laminated Cores						
45 Permalloy	14 mil sheet	2,700	45	0.4	8	1,160
45 Permalloy	3 mil sheet	2,700	45	0.4	14	53
78.5 Permalloy	6 mil sheet	9,000	16	0.2	0	600
4-79 Mo-Permalloy	6 mil sheet	22,000	55	0.05	0	173
3.8-78.5 Cr-Permalloy	6 mil sheet	10,000	65	0.3	0	146
45-25 Perminvar	6 mil sheet	365	19	0.01	0	505
7.5-45-25 Mo-Perminvar	6 mil sheet	550	80	0.1	0.5	120
2 V-Permendur	14 mil sheet	800	26	1.0	—	2,000
4% Silicon Iron	14 mil sheet	400	60	120	75	870
Magnetic Iron	14 mil sheet	250	10	50	—	5,200
Loading Coil Cores						
Isoperm ¹⁶	1.7 mil sheet	85	45	2.3	20	15
Grade B Iron Powder	80 mesh powder	35	10	50	110	88
Grade C Iron Powder	200 mesh powder	26	10	80	140	31
81 Permalloy Powder	120 mesh powder	75	16	5	37	51
2-81 Mo-Permalloy Powder	120 mesh powder	125	40	1.6	30	17
Sendust ¹⁷	120 mesh powder	65	80	5	100	4.0
Ferrocart ¹⁸	5 micron powder	13	10	5	60	0.8

MECHANICAL SENSITIVITY

Stresses beyond the elastic limit in soft magnetic materials should be avoided if possible, since they lower the permeability, and in general increase the hysteresis loss. A partial exception to this statement is found with the "Isoperms" (essentially 50 per cent Ni, 50 per cent Fe). These alloys, after excessively hard rolling, annealing, and final moderately hard rolling, develop low permeabilities and low hysteresis losses

¹⁶ O. Dahl & J. Pfaffenberger, *Zeit. f. tech. Phys.* 15, 99 (1934).

¹⁷ H. Masumoto, *Tohoku Sci. Rep. Honda Anniv. Vol.*, p. 388 (1936).

¹⁸ M. Kersten, *Elektrotech. Zeit.* 50, 1335 (1937).

\$2.80/lb. increase the cost of an alloy very considerably in comparison with iron or steel at less than 7¢/lb. High priced magnetic alloys can only be justified in general when their extraordinary characteristics permit offsetting apparatus performance or economies. Of course, they may be absolutely necessary for certain types of apparatus.

The cost data for Tables I and II have been calculated from recent prices²¹ for raw materials of quality suitable for magnetic alloys. The cost of raw materials is given in Fig. 3 for selected alloys. The low raw materials costs of iron, silicon-iron, and chromium steel are notable, as well as that of the powder core material "Sendust." 45 permalloy is the cheapest of the high permeability materials, while Mishima steel is the cheapest of the high quality permanent magnet materials.

Comparisons based on raw materials costs are not entirely satisfactory. The cost of alloying and reducing to finished form may overshadow the cost of raw materials, particularly when high purity, exact tolerances, and small rates of production are involved.

APPLICATIONS

Almost all magnetic properties are utilized in some type of telephone apparatus. They are generally linked inseparably with electrical and mechanical properties. The proper design of any apparatus strikes a compromise between the various technical features and cost. The technical features of materials used in present day apparatus are listed below. Acceptable common properties, such as mechanical soundness and workability, are assumed for all materials unless specifically mentioned. Listing is made on the basis of the magnetic effect utilized.

1. *Simple Tractive Force (Relays)*

The force of attraction between two neighboring surfaces of area A , between which the flux density is B , is

$$F = kAB^2.$$

The primary telephone application of this effect is to relays and switches. For greatest tractive force, materials capable of attaining high flux densities are desirable. However, the air gap in the magnetic circuit absorbs such a large proportion of the available magnetomotive force that higher or lower permeabilities in the core material are frequently less important than efficiency of design. A typical relay structure is shown in Fig. 4.

²¹ *Steel*, Oct. 3, 1938.

Reference to Figs. 1 and 2 points to 2 V-permendur as outstanding in the flux density range above 15,000 gauss. This material is excluded for many applications because of its high cost. High temperature hydrogen annealing improves the high flux density behavior of most magnetic materials, as noted above in connection with Fig. 2. Using ordinary methods of annealing, the next best material for high flux operation is the standard magnetic iron, while 45 permalloy is preferable at flux densities below 12,000. For the low magnetizing forces available in sensitive relays, 4-79 Mo-permalloy gives the largest tractive forces.

There are frequently other requirements in addition to tractive force in relay construction. The operation and release characteristics

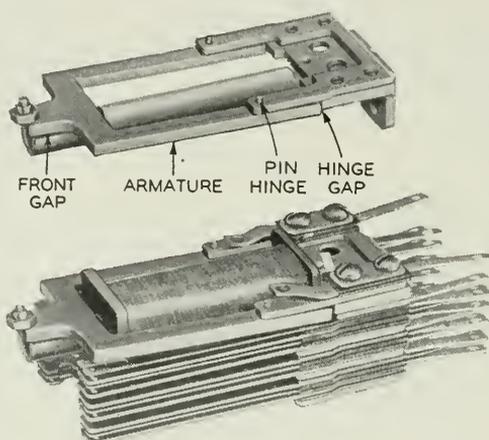


Fig. 4—U-type relay, showing the magnetic circuit

are determined by the resistivity and coercive force of the material. Sensitive, quick release relays require materials like permalloy, while slow release types may utilize the larger coercive force and residual of magnetic iron or cold rolled steel.^{22, 23}

Relays for a-c. applications may have objectionable eddy current losses in their cores. Such losses can be reduced by use of high resistivity material. Thus, 45 permalloy has one-fourth the loss of iron for the same core thickness and flux density.

2. Polarized Tractive Force (Receivers, Ringers, Relays)

If permanent polarizing flux density B_p exists between two neighboring surfaces, and a small additional flux Φ is applied by means of a

²² H. N. Wagar, *Bell Labs. Record* 16, 300 (1938).

²³ F. A. Zupa, *Bell Labs. Record* 16, 310 (1938).

magnetomotive force M , the additional tractive force will be approximately

$$F = 2kB_p\Phi = \frac{2kB_pM}{R_a(1 + n/\mu_r)},$$

where R_a in the latter part of the above equation is the reluctance of the air-gap, n is ratio of the reluctance of the ferromagnetic circuit with iron removed to the reluctance of the air-gap, and μ_r is the reversible permeability, i.e. the permeability measured with very small a-c. magnetizing forces in the presence of the polarizing flux density. It thus appears that materials for such applications should have high saturation values. The apparatus should be designed to obtain a low value of n , and to operate at a flux density to make the above force a maximum.

Figure 5 gives values of μ_r as a function of polarizing or superposed flux density B_p . It should be remarked that the reversible permeability is practically single valued²⁴ when plotted against polarizing flux, in contrast with the "butterfly" loop obtained by plotting against polarizing field strength. The superiority of permendur in Fig. 5 is obvious. Values of the force factor for $n = 100$ and $n = 1000$ have been computed for these materials for an arbitrary value of air-gap reluctance and magnetomotive force. It is evident that the full advantage of permendur is not realized unless the apparatus design is such as to attain a low value of the air reluctance ratio n .

Permanent magnet materials are frequently employed to supply polarizing flux. When they form a part of the circuit for the alternating flux, their reversible permeability becomes important. Reference to Table II shows that 5 per cent W steel has the highest permeability of the common permanent magnet materials. However, its energy product $(B \cdot H)_{\max.}$ is so low that other materials are preferred for applications where space is limited, despite their low reversible permeabilities.

The earliest application of polarized structures in the telephone plant was to the receiver. The receiver of the present day is constructed with remalloy permanent magnets, 45 permalloy pole-pieces, and a permendur diaphragm.²⁵ The magnetic circuit is shown in Fig. 6.

The second application of polarized structures is to the telephone bell or ringer. Here a permanent magnet is used, to supply polarizing flux to the two cores for the coils through the shoe at one end and the

²⁴ R. Gans, *Phys. Zeit.* 12, 1053 (1911).

²⁵ W. C. Jones, *B. S. T. J.* 17, 338 (1938).

armature at the other.²⁶ Non-magnetic stop pins are required on the armature to prevent sticking. This complicated magnetic circuit has been developed to meet the numerous requirements placed on ringers. Polarization is necessary for an a-c. ringer which is to operate without

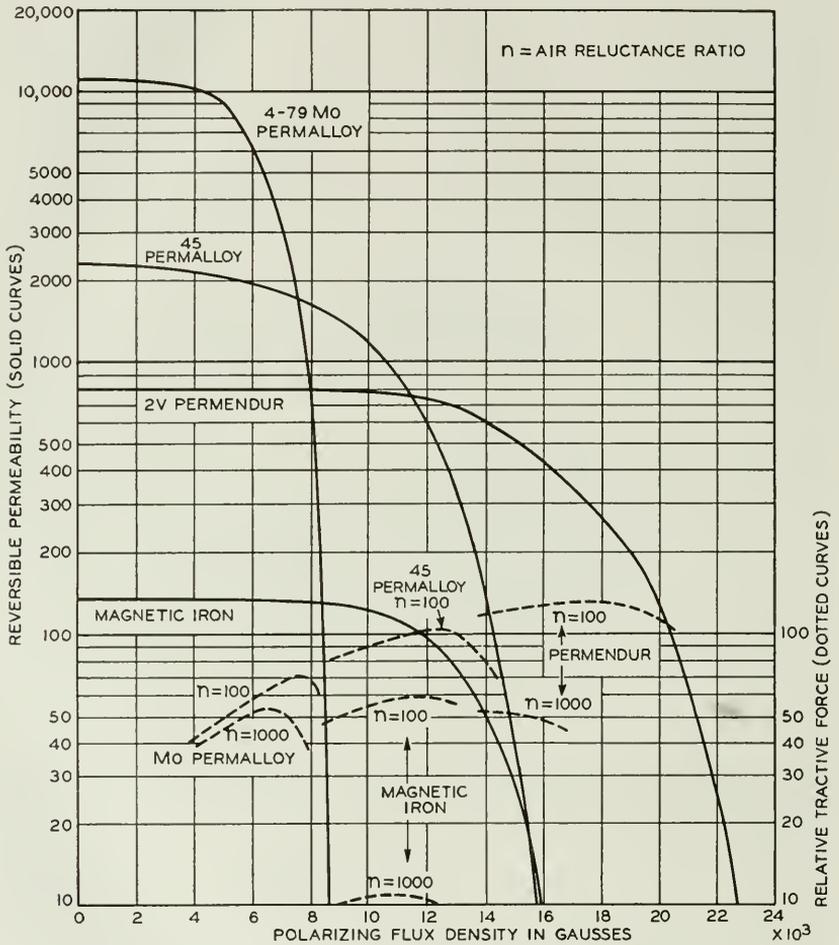


Fig. 5—Reversible permeability and relative polarized tractive force for various magnetic alloys.

interrupter contacts, and it is employed in selective ringing on party lines. A high coil inductance is required to limit the transmission

²⁶ K. B. Miller, "Manual Switching and Substation Equipment" (McGraw-Hill, 1933 ed.), p. 67.

losses due to the shunting effect of ringers across the line, especially in the case of party lines.

The third application of polarized structures is in certain relays for composite ringing, duplex telegraph, and special selecting circuits. Various materials are used in such relays, depending upon the sensi-

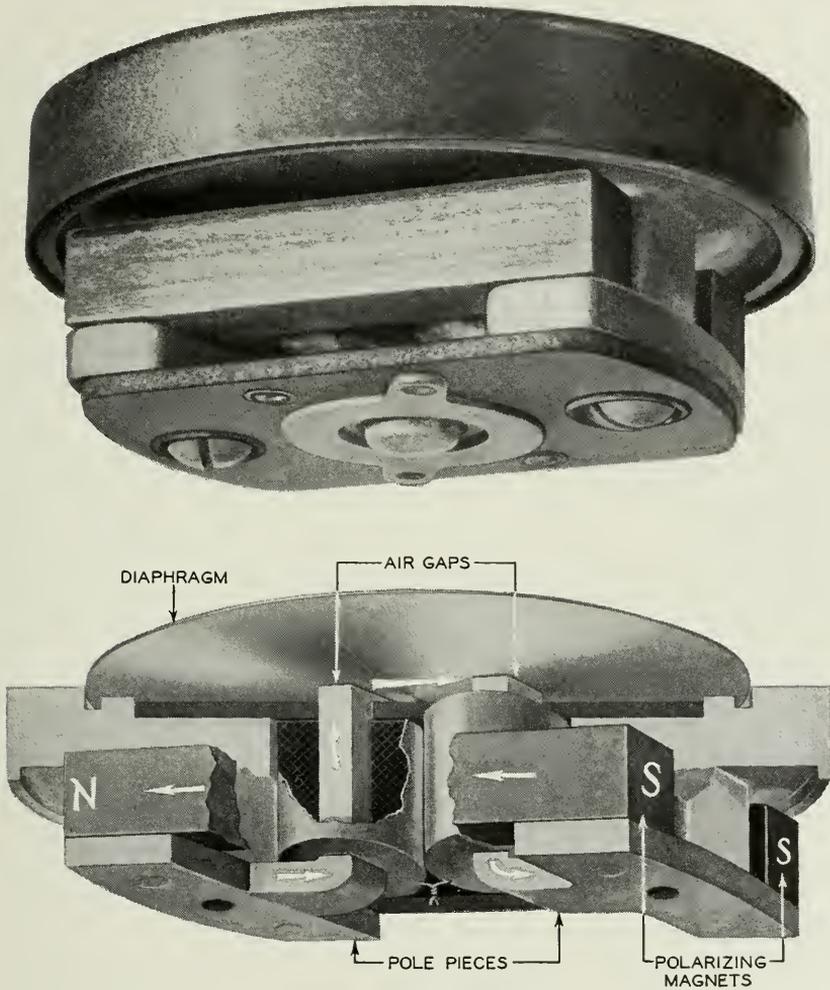


Fig. 6—Magnetic circuit of the new telephone receiver.

tivity required, so that no general statement can be made as to needs in this field. The problems encountered are essentially similar to those met in the receiver and ringer, for which numerous materials are available.

3. Force on Current (Moving Coil Receivers, Light Valves, Motors)

A straight wire of length l carrying a current i in a perpendicular magnetic field of flux density B is pushed at right angles to the field and the length with a force

$$F = Bil = \mu H il,$$

where μ is the permeability and H is the magnetizing force in the nearby material from which the flux is derived. Again, the prime requirement for a useful material is high flux capacity, and high permeability, so that the magnetizing force need not be large. The magnetizing force has been supplied generally in the past by means of direct current in windings built into the apparatus. With the normally available voltages and currents, sufficient magnetizing forces could be obtained only with coils having a large number of turns, and low resistance. This generally involved such large structures that cost considerations compelled the use of iron cores with perhaps pole-pieces made of permendur. Lately many structures are being designed to replace costly electromagnets with permanent magnets made from Mishima type steel.

Moving coil receivers and loud speakers²⁷ are the most important representatives of this type of apparatus. Others are the string oscillograph,²⁸ the light valve,²⁹ phonograph record recorder,³⁰ and various types of power machinery.³¹ Several of these are now constructed with permendur pole-pieces, and cast Mishima steel magnets, or remalloy magnets where hot rolling will assist in producing small, accurately sized parts.

4. Induced Electromotive Force

The electromotive force between the terminals of a coil of N turns linking flux φ is

$$-e = N \frac{d\varphi}{dt}.$$

The arrangement of coils and interlinking flux differs considerably in the various types of apparatus employing this effect.

The flux variation is provided by means of mechanical motion of the coil in instruments such as the electromagnetic microphone. It is varied by means of fluctuations in magnetizing current in inductance

²⁷ E. C. Wente & A. L. Thuras, *B. S. T. J.* 10, 565 (1931).

²⁸ A. M. Curtis, *B. S. T. J.* 12, 76 (1933).

²⁹ G. E. Perreault, *Bell Labs. Record* 10, 412 (1932).

³⁰ H. A. Frederick, *B. S. T. J.* 8, 159 (1929).

³¹ R. D. deKay, *Bell Labs. Record* 16, 236 (1938).

coils and transformers. In the moving coil microphone, the cylindrical coil moves axially in a slot between an inner magnetized cylinder and an outer magnetic cylinder which receives the radial flux threading the coil. For such a cylindrical coil in a uniform radial field, the e.m.f. is

$-e = lB \frac{dx}{dt}$, where l is the total length of wire composing the coil,

and x is its displacement perpendicular to the radial field. For a stationary coil linking a varying flux as in a transformer, $-e = AN \frac{dB}{dt}$

$= AN\mu \frac{dH}{dt}$, where A and μ are the area and permeability of the core within which the magnetizing force is H . It is evident from these equations that high flux density or permeability are desirable, in order to yield the largest e.m.f. with least material.

4a. *Microphones, Magnetic Tape Recorders, Magnetos*

An application involving a moving coil is the public address microphone,³² where the flux is established by means of a cobalt steel magnet, and is concentrated upon the moving coil by means of permendur pole-pieces.

An inverse application in which the coil is stationary and the magnet moves is the magnetic tape recorder.³³ In this, a steel tape which has been magnetized by speech currents is drawn between permalloy pole-pieces in pick-up coils. High coercive force, high signal-to-noise ratio, mechanical soundness, durability, and cheapness of the tape material are desirable.

The telephone magneto employs the e.m.f. generated by rotating a coil in a magnetic field. It has been constructed with iron armature and pole-pieces, and chrome steel field magnets. Recent designs using modern magnetic materials have indicated the possibility of large economies in volume.³⁴ The magnetic properties of available materials have now reduced the volume required by magnetic parts to a point where the major problem in magneto design is to compress the gears and shafts into correspondingly small space and yet maintain sufficient mechanical strength, durability, and convenience of operation.

4b. *Inductance Coils*

A very important application of induced e.m.f. focuses attention on the inductance of a coil of N turns surrounding a (closed) core of area A ,

³² R. N. Marshall & F. F. Romanow, *B. S. T. J.* 15, 405 (1936).

³³ C. N. Hickman, *B. S. T. J.* 16, 165 (1937).⁹

³⁴ *Ericsson Bulletin No. 12*, 46 (1938).

magnetic path length l , and permeability μ . The inductance is increased through the presence of the core by an amount

$$L = 4\pi N^2 \mu A / l.$$

As noted earlier, when eddy current shielding is negligible such an inductance is accompanied at frequency f by hysteresis, residual, and eddy current resistances to give a total as follows:

$$R/L = \mu f (aB_m + c + ef).$$

At frequencies high enough to introduce eddy current shielding, the effective inductance due to a core of laminar thickness t and resistivity ρ is reduced below the ordinary inductance L_0 by the ratio

$$L/L_0 = \frac{1 \sinh \theta + \sin \theta}{\theta \cosh \theta + \cos \theta},$$

where $\theta = 2\pi t \sqrt{\mu_0 f / \rho}$. Figure 7 shows this ratio and the ratio $R_c / \omega L_0$ as functions of θ . As a practical example, the permeability of 6 mil (0.015 cm.) 4-79 Mo-permalloy is reduced to about 75 per cent of its initial value (22,000) at 1 kilocycle, and to about 17 per cent at 10 kc.

The telephone loading coil adds inductance to the telephone line, but it must not add excessive resistance. Furthermore, its inductance must be extremely stable with the lapse of time, and under severe operating conditions, such as occasional current surges induced from lightning discharges. Iron wire cores for loading coils were supplanted over twenty years ago by compressed iron powder cores, and these in turn gave way to permalloy powder cores. The latest improvement is the introduction of 2-81 Mo-permalloy powder cores.⁷ The reduction in size of cores with these improvements is shown in Fig. 8.

Another method of loading a line is by sheathing the conductor with a continuous layer of magnetic material. This method was used with notable success on submarine telegraph cables by wrapping permalloy tape upon the conductor, and annealing before applying insulation.³ The location of the loading material is shown in Fig. 9. The continuous loading of long submarine telephone cables has been shown to be feasible using thin 7.5-45 Mo-perminvar tape.⁷

Retardation and choke coils have a great variety of applications, running from a tiny coil weighing $3\frac{1}{2}$ ounces³⁵ to a 4600 lb. generator ripple suppressor.³⁶ The contrast is evident in Fig. 10. Retardation

³⁵ D. W. Grant, *Bell Labs. Record* 11, 173 (1933).

³⁶ R. A. Shetzline, *Bell. Labs. Record* 17, 34 (1938).

coils, used as network elements, are generally equipped with compressed powder cores, and the same improvements are indicated as for loading coils noted above. Laminated permalloy cores are often used for low-frequency applications (ringing, telegraph), where high inductance is desired, and a-c. losses are naturally low.

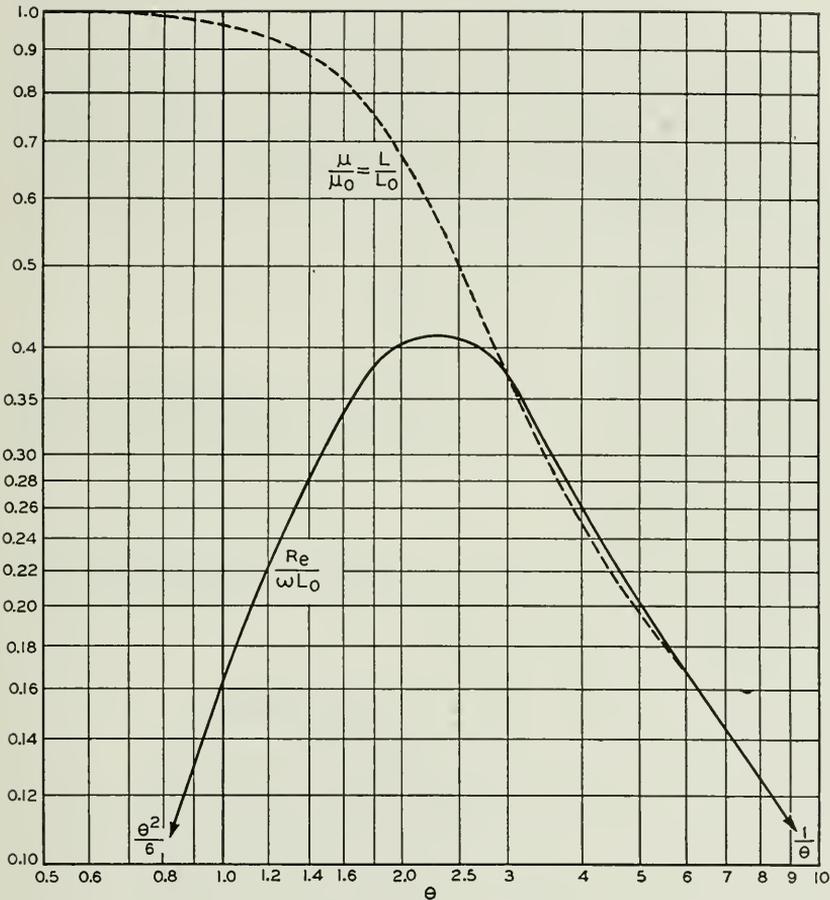


Fig. 7—Effect of eddy current shielding on apparent permeability and on eddy current resistance/reactance ratio of sheet core material.

Coils designed for direct currents must have cores with high a-c. permeability in the presence of superposed field, i.e. they must be made of high permeability magnetic materials having high saturation values. Reference to Fig. 5 shows permendur outstanding in this regard. In practice, the low costs of silicon iron or magnetic iron are

frequently decisive in the selection of these materials instead of the technically superior materials noted above. Air-gaps are often found necessary to reduce the superposed field strength in the magnetic core.

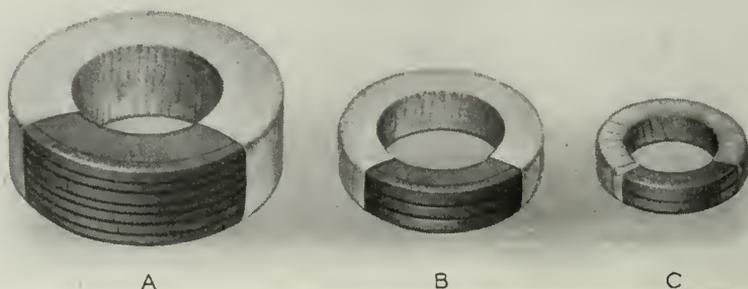


Fig. 8—Relative sizes of compressed powder cores for loading coils; A. Iron, B. 80-permalloy, C. 2-81 Mo-permalloy.

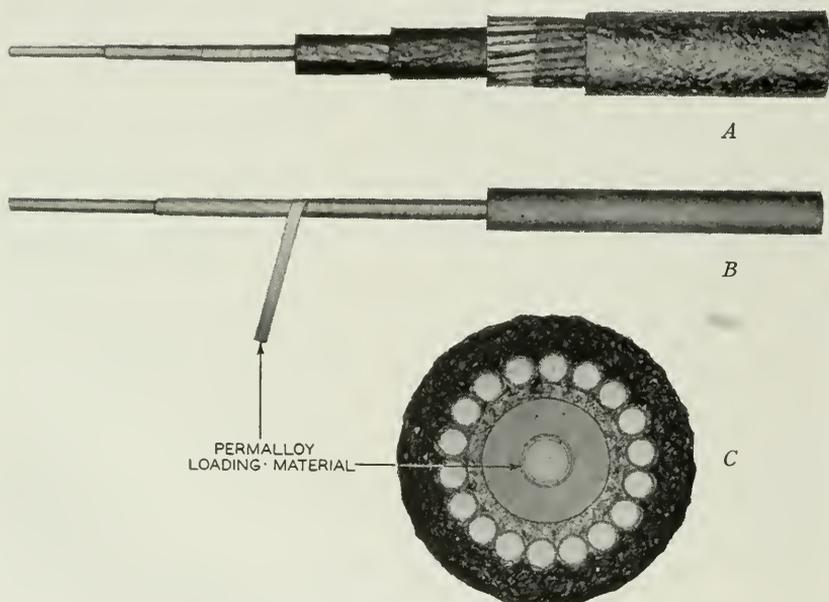


Fig. 9—Permalloy loaded telegraph cable; A. Armored deep sea type, B. Gutta percha insulated core, C. Enlarged cross sectional view.

This serves to retain a fairly large reversible permeability in the core, and, if the air gaps are not too large, it yields a larger net effective permeability than could be obtained without air-gaps.

A further type of inductance is the impulse coil used in harmonic generators.³⁷ The cores of these coils should be saturated over most of the magnetizing cycle, and reverse very quickly and completely just as the magnetizing force passes through zero. This implies use of material having a high permeability, and a high resistivity, such as 4-79 Mo-permalloy.

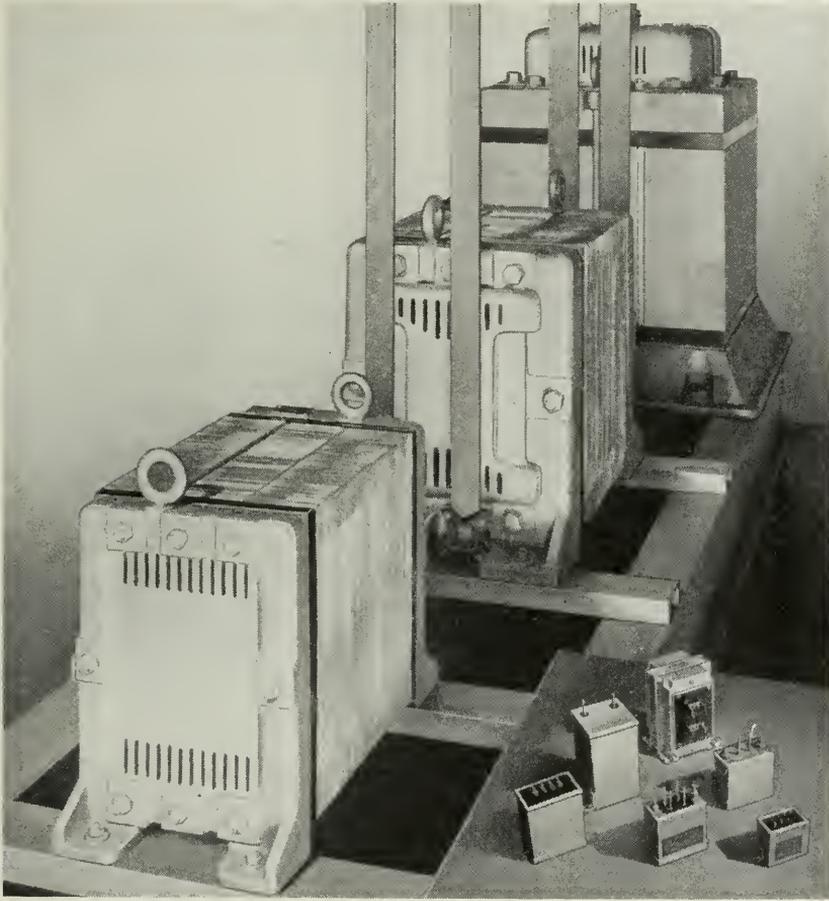


Fig. 10—Large and small coils used in the telephone system.

4c. Transformers

With transformers, the inductance and losses of the individual coils can be analyzed as simple inductances, for which the considerations of

³⁷ E. Peterson, J. M. Manley & L. R. Wrathall, *Elec. Engg.* 56, 995 (1937).

4b apply. In addition, however, the coupling factor between coils on the same core becomes of especial importance. In the usual design the flux linkage common to the primary and secondary is largely contained in the magnetic material, while the leakage flux is controlled by the reluctance of the air path. It is thus evident that a large value of core permeability is required to obtain a high coupling factor. Of course, advantages indicated by high permeability may be lost through improper design.

The earliest transformer employed in the telephone plant was the induction coil. This originally consisted of two windings on a core composed of a bundle of iron wires.³⁸ Later, silicon iron sheet cores were introduced.

Input and output transformers have varied applications for which special types of cores are required. Where superposed current is not involved, space and weight can be economized by use of high permeability materials such as chrome or molybdenum permalloy.³⁹ Eddy current shielding at higher frequencies will offset much of the permeability advantage indicated for these materials unless they are laminated sufficiently. However, thin laminations are costly to prepare and stack, and difficult to insulate and handle without mechanical injury and corresponding reduction of permeability. An intermediate thickness of permalloy sheet is generally chosen, which secures a considerable advantage in permeability over iron, without prohibitive cost.

Where a winding must carry direct current, conditions are similar to those applying to choke coils, and materials with high reversible permeability at high magnetizing forces are required. Frequently, for large magnetizing forces, it is desirable to include air-gaps in the magnetic circuit. Silicon iron is the ordinary material for such application, as it is for power transformers.

4d. *Magnetic Shielding*

A further application of induction effects is in magnetic shielding of apparatus. A magnetic shield consists of a high permeability shell (4-79 Mo-permalloy, or 78.5 permalloy) which shunts flux around the enclosed apparatus. For a-c. shielding, alternate layers of copper and permalloy sheet are very effective in magnetic shunting and eddy current screening of the enclosed space.⁴⁰ High initial permeability,

³⁸ Cf. p. 43 of Reference 26.

³⁹ A. G. Ganz & A. G. Laird, *Elec. Engg.* 54, 1367 (1935).

⁴⁰ W. G. Gustafson, *B. S. T. J.* 17, 416 (1938).

mechanical workability, and low cost are desirable for such applications.

5. Magnetostriction

The relative change in length of a magnetic bar upon magnetization, $\delta l/l$, ranges from negative to positive values, depending on alloy composition, and is roughly proportional to B^2 . If a polarizing field is applied, small additional alternations of field will give accompanying and nearly proportional alternations of length of a bar. These alternations evidence themselves in the electrical constants of a coil enclosing the bar.

This effect can be utilized in oscillators and filters for frequencies which involve the use of mechanically resonating bars of convenient size. Among the high permeability materials, 45 permalloy appears to give the largest magnetostrictional effect.⁴¹ In order to limit eddy current losses, the material must be laminated more or less finely, depending on the frequency.

The inverse magnetostriction effect by which e.m.f. is generated in a coil when the core is vibrated, becomes objectionable as a source of circuit noise in the transformers of high gain amplifiers.³⁹ An alloy with minimum magnetostriction such as 81 permalloy or 4-79 Mo-permalloy is preferred for such cases.

Other effects of magnetostriction are the generation of sound by the cores of coils subject to alternating magnetization, and the appearance of undesired resonance effects in the electrical circuit at frequencies at which the core resonates mechanically.

6. Thermal Variation of Permeability

The initial permeability of ordinary magnetic materials increases more or less slowly with increasing temperature, until a maximum value is reached, above which temperature the permeability declines very rapidly to the non-magnetic, or Curie point. The Curie point of an alloy can be moved down the temperature scale by adding non-magnetic materials, such as Mo, Cr, Cu, etc., to the alloy.⁴²

The positive temperature coefficient of inductance of powder core coils due to thermal change of permeability becomes objectionable in crystal filters, where only very small variations in the resonant frequencies can be tolerated.⁴³ It is made slightly negative to counteract the small positive coefficient of mica condensers by the admixture with

⁴¹ A. Schulze, *Zeit. f. Phys.* 50, 448 (1928).

⁴² G. W. Elmen, *Bell Labs. Record* 10, 2 (1931).

⁴³ C. E. Lane, *B. S. T. J.* 17, 125 (1938).

the regular 2-81 Mo-permalloy powder of a small amount of permalloy powder containing about 12.5 per cent of molybdenum. The latter material has a Curie point just above room temperature.

CONCLUSION: DESIRABLE AND POSSIBLE IMPROVEMENTS

It appears from the above inspection of the magnetic elements of apparatus that there is a general desire for all properties which contribute to magnetic effects to be increased, and for those which cause apparatus energy losses to be decreased. A considerable success has already been achieved in these directions. The obstacles to further improvement are in part difficulties in commercial application of laboratory techniques, and in part ultimate limitations to the properties of materials.

The chart in Fig. 11 shows the recent trend toward the realization of the best possible magnetic properties in new materials or by improved processes. Values are given for the year 1920 (i.e. before commercial application of the permalloys), for 1939 as commercially and as experimentally realized, and for what may be considered as the attainable limit. It is evident that most of the properties which could be improved have been improved in commercial materials by a factor of ten or so within the last twenty years. A further improvement of several properties by a factor as large as ten has been observed by experimental procedures. These improvements have not been utilized in all cases, either because they may not be of great practical value, or because they involve processes which are commercially impracticable, or materials which are very expensive. Thus, the highest value of μ_m has been attained on a single crystal of H_2 purified iron⁴⁴ cut so as to make a hollow parallelogram with sides parallel to the (100) crystal axes, and annealed in H_2 below the α - γ transformation point. The highest permanent magnet quality has been attained with a platinum cobalt alloy⁴⁵ costing some \$400 per pound.

One of the main objectives of commercial magnetics research has been the attainment of higher permeabilities—initial, maximum, reversible, and at high flux densities. The reversible permeability is closely linked with the initial permeability and saturation flux density. The permeability at high flux densities⁴⁶ appears to be susceptible to considerable increases by proper treatment of materials. The initial permeability can be increased by proper technique, but such gains are frequently sacrificed in practice because of unavoidable mechanical

⁴⁴ P. P. Cioffi & O. L. Boothby, *Phys. Rev.* 55, 673 (1939).

⁴⁵ W. Jellinghaus, *Zeit. Tech. Phys.* 17, 33 (1936).

⁴⁶ For example, at $B = 10,000$, or $B = 20,000$.

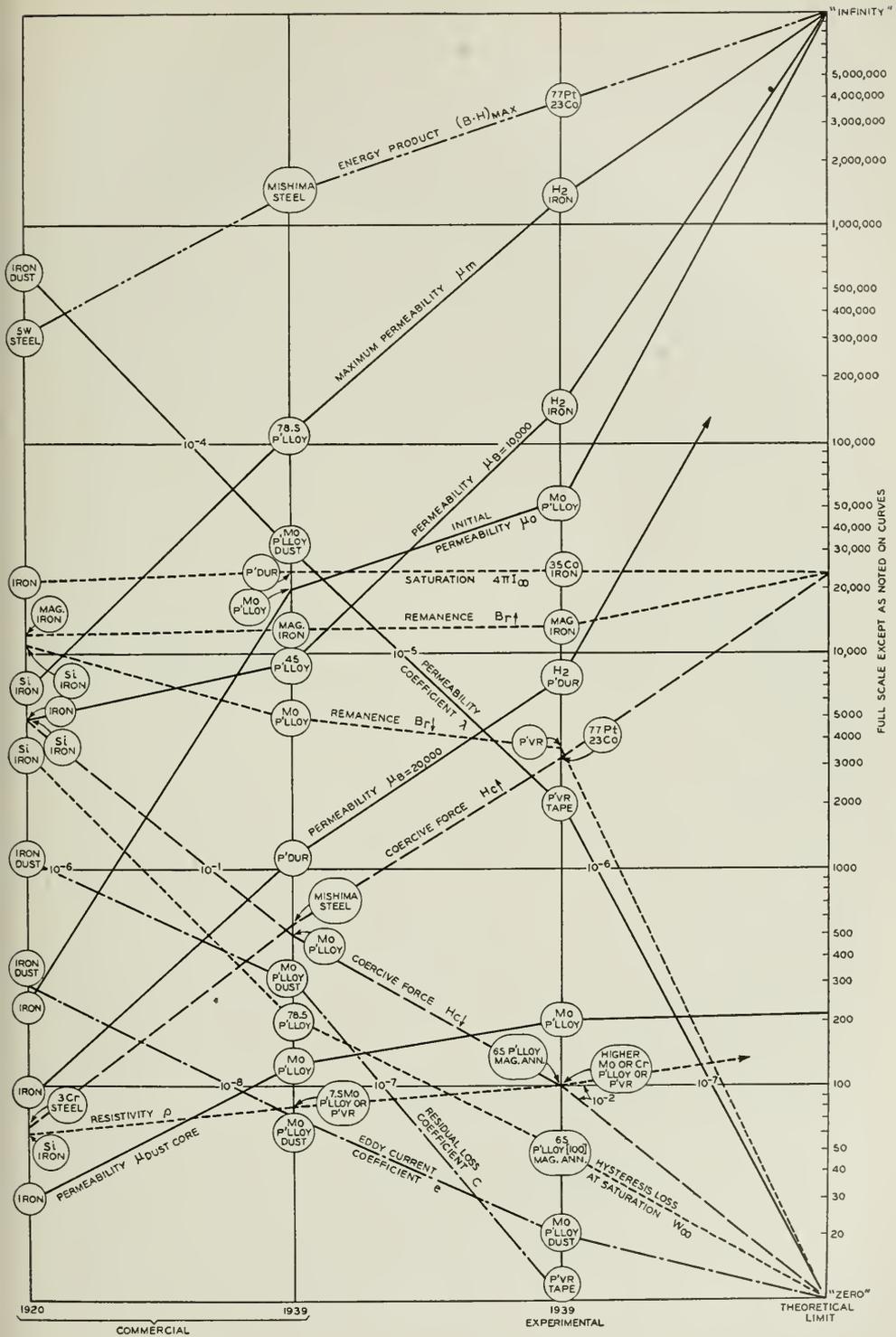


Fig. 11—Improvements in the properties of magnetic materials since 1920, in relation to theoretically possible properties.

stresses, or because of large eddy current shielding. Higher maximum permeabilities than now attainable do not promise great utility. However, the low values of coercive force and hysteresis loss found with materials having high maximum permeability may be sufficiently desirable, regardless of permeability needs.

Another important objective of magnetics research has been to reduce energy losses. Hysteresis loss at low flux densities, as indicated by the loop area coefficient a , should be decreased to cut down harmonic generation and modulation in magnetic core coils. Perminvar has shown desirably reduced losses, but it is sensitive to magnetic and mechanical conditions. Eddy currents are controlled by resistivity and degree of subdivision of the magnetic core. The resistivity of magnetic alloys can be increased to around 100 microhm-cms. by alloying with large enough quantities of chromium, molybdenum, etc., but at a serious sacrifice of magnetic quality for resistivities above about 60. Eddy current suppression by laminating or pulverizing the magnetic material thus offers a greater range of control than resistivity adjustment.

Permanent magnetic materials have also reached a very successful stage, from the magnetic point of view. The greatest handicap of the better materials is extreme hardness, which hampers fabricating processes.

Impedance Properties of Electron Streams

By LISS C. PETERSON

The input impedance of an idealized space charge grid tube is investigated under general conditions of space charge between the accelerating grid and the negatively biased control grid. It is shown that under certain space charge conditions the input capacitance and conductance both may be negative. These impedance properties persist up to frequencies for which the transit angle is quite large. Possibilities of designing electronic negative capacitances are thus opened up. Experimental results are also given; these give a broad confirmation of the theoretical deductions.

PART I

THEORY

IN the early stages of vacuum tube history the theoretical work on d-c. space charge treated mainly potential distributions associated with fairly small initial velocities of the electrons. With the advent of multi-electrode tubes this situation changed for it then became necessary to consider also potential distributions occurring when electrons are injected with large initial velocities. Idealizations were introduced to the extent that consideration was given only to space charge conditions which can exist between two parallel planes at known potentials when electrons with normal velocities corresponding to these potentials are injected into the region through one or both planes.¹

Some time ago it was discovered experimentally that space charge may under certain conditions produce a negative capacitance. The negative capacitances were found during a series of low-frequency measurements of the control-grid-to-ground capacitance of an experimental space-charge-grid tube. In the course of these measurements it was found that with all the electrodes carefully by-passed to ground for a-c. except the negatively polarized control grid, the input capacitance as well as the input conductance was negative in certain domains which depended upon the d-c. operating voltages.

In order to arrive at an understanding of this fact, a-c. phenomena must be considered under the general d-c. space charge conditions

¹ Plato, Kleen, Rothe, *Zeitschrift f. Phys.*, 101, July 1936. C. E. Fay, A. L. Samuel, W. Shockley, *Bell Sys. Tech. Jour.*, January 1938. B. Salzberg, A. V. Haefl, *R.C.A. Review*, January 1938.

referred to. The investigation comprises the calculation of the impedance between two parallel planes in vacuum, at known d-c. potentials, when an electron stream with normal d-c. velocities corresponding to these potentials is injected at right angles to the planes. As a further step, the effect of a negatively polarized grid interposed between the two planes will be considered. This latter arrangement may be taken to correspond to an idealized space charge grid tube.

Before these calculations are presented it is well to set forth as concisely as possible those results of the d-c. space charge analysis¹ which will be of most immediate interest. For this purpose we start with a qualitative review of the work of Fay, Samuel and Shockley.

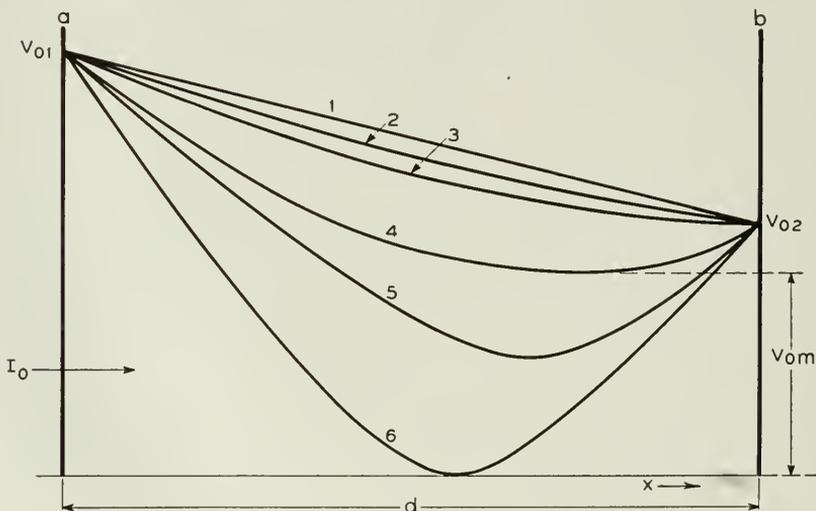


Fig. 1—Potential distributions between two planes of positive potential.

Consider two planes *a* and *b*, Fig. 1, of fixed d-c. potentials V_{01} and V_{02} ($V_{02} \leq V_{01}$) respectively, separated by a distance d . Let a unidirectional and uniform electron stream of I_0 amp./cm.² be injected into the space from the left at right angles to the planes. When the injected current is extremely small the potential distribution does not differ very much from the free space one, represented by curve 1 on Fig. 1. As the injected current is increased slightly, the potential curve starts to sag, curve 2, and a further small increase causes a potential minimum to develop at the electrode of lower potential, curve 3. Still more increase in injected current makes the potential

¹ Loc. cit.

minimum V_{0m} sink and move towards the electrode of higher potential, curve 4. This state of affairs, with a continuously decreasing potential minimum, continues until a critical value of injected current is reached. The potential distribution may now be represented by curve 5. The slightest further increase in injected current causes the potential minimum to sink abruptly to zero: a virtual cathode is formed. This abrupt change will be referred to as a Kipp.

With this qualitative discussion of the various potential distributions in mind a more detailed classification may be made. If both planes are assumed to be at positive potentials we may classify the different potential distributions as follows:

1. Type B
2. Type C
3. Type D

Type B corresponds to virtual cathode operation. This mode of operation will be of no interest in this paper. Type C corresponds to the case when a potential minimum at positive potential is present between the planes and type D to the case when no such potential minimum is present. For the purpose of analysis it is convenient to distinguish between two types of D solution, i.e., D_1 and D_2 . This comes about because the equations for potential distribution between the planes may exhibit a minimum outside the planes. The D_1 solutions correspond to the case when no such minimum exists and the D_2 solutions to the case when such a minimum does exist.

Let us consider Type C distributions in some more detail. For this purpose attention is directed to Fig. 2. Here the ratio $\frac{V_{0m}}{V_{01}}$ is shown as a function of the injected current I_0 with the ratio $\frac{V_{02}}{V_{01}}$ as parameter.

The dotted curve represents a boundary line; for currents smaller than that given by this boundary no potential minimum can exist between the planes. Consider the curve $\alpha\beta\gamma$. At the point α the potential minimum sets in and as more current is injected the potential at the minimum decreases continuously until the point β is reached. Any further increase in current causes the potential minimum to sink abruptly to zero; thus β corresponds to the Kipp point. Now it is seen that within a section of the curve $\alpha\beta$ there are two possible solutions, namely those corresponding to the section $\beta\gamma_1$ of the upper branch and those corresponding to the lower branch $\beta\gamma$. Let us designate by C_1 space charge conditions corresponding to the upper branch $\alpha\beta$ and by C_2 those corresponding to the lower branch $\beta\gamma$.

After this survey of the several possible d-c. space charge conditions we may proceed to the a-c. phenomena involved. At the present time the impedance between the two planes of Fig. 1 can be found only if the electrons move in one direction, i.e. if no virtual cathode is present between the planes, and therefore this assumption is made. It will be further assumed that electrode "a" where the injection takes place is at a-c. ground potential. This insures constant conduction current and electron speed at the plane of injection. The impedance

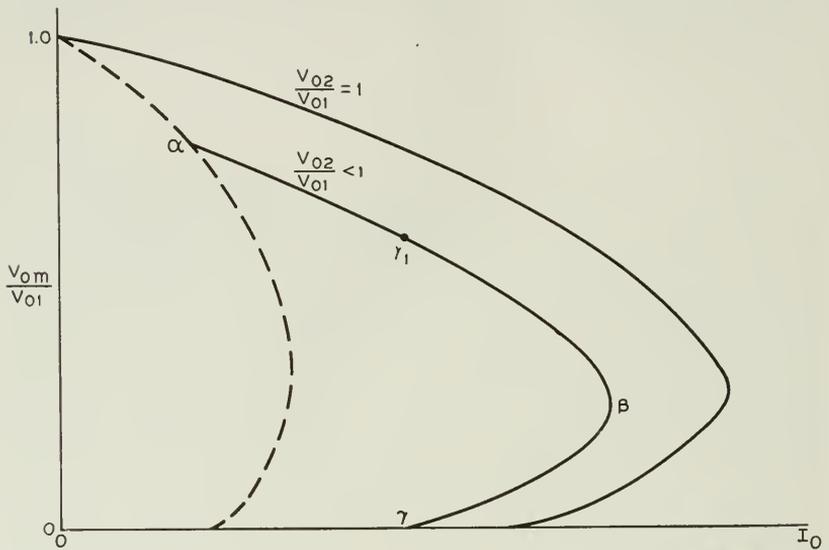


Fig. 2—Variations of the magnitude of the potential minimum as function of injected current.

may then be found by proper application of the general theory developed by Müller and Llewellyn.² The result is that the impedance may be represented by a series combination of a resistance r and a capacitance C having the values

$$r = J_0 \frac{T^4}{\epsilon} \cdot \frac{2 - 2 \cos \theta - \theta \sin \theta}{\theta^4} \quad (1)$$

$$C = C_0 \cdot \frac{1}{1 - J_0 \frac{T^3}{d} \cdot \frac{2 \sin \theta - \theta - \theta \cos \theta}{\theta^3}},$$

² J. Müller, *Hochfrequenztechnik u. Electroakustik*, May 1933; F. B. Llewellyn, *Bell Sys. Tech. Jour.*, October 1935.

where

$$J_0 = \frac{e}{km\epsilon} I_0$$

$$\theta = \omega T$$

$$C_0 = \frac{\epsilon}{d} \text{ is the cold capacitance}$$

T is transit time, seconds

e is electronic charge, coulombs

m is electronic mass, grams

$$\frac{e}{m} = 1.77 \times 10^8 \text{ coulomb/gm.}$$

ϵ is permittivity of vacuum = 8.85×10^{-14} Farads/cm.³

$$k = 10^{-7}.$$

Present interest lies mainly in the range where the transit angle θ is small. Expanding (1) and assuming θ small, we have:

$$\begin{aligned} r &= \frac{J_0 T^4}{12\epsilon} \\ C &= C_0 \frac{1}{1 - J_0 \frac{T^3}{6d}}. \end{aligned} \quad (2)$$

These formulas are also found in Müller's paper.²

With regard to the resistance r it is immediately seen that it is always positive and has the same equation as for complete space charge with current I_0 and transit time T .

From the capacitance equation it is immediately evident that an increase is caused by the presence of electrons. The dielectric constant of space charge under the stipulated conditions is seen to be dependent upon the d-c. conduction current I_0 . As the injected current is increased the capacitance increases initially at a fairly slow rate, but as $J_0 \frac{T^3}{6d}$ becomes comparable with unity the capacitance rises rapidly.

It becomes infinite for

$$J_0 \frac{T^3}{6d} = 1 \quad (3)$$

and if the left member of (3) were to become greater than unity the capacitance would be negative. The possibility of such a condition deserves careful consideration.

² Loc. cit.

From the discussion of the d-c. space charge it follows that the d-c. current has an upper limit; i.e., the Kipp current with the value

$$I_{0K} = \frac{4}{9} \epsilon \sqrt{\frac{2e}{mk}} \frac{(\sqrt{V_{01}} + \sqrt{V_{02}})^3}{d^2}. \quad (4)$$

To determine the relation between the Kipp current (4) and the current (3) required for infinite capacitance, consider the d-c. equations:

$$\begin{aligned} u_b &= u_a + a_a T + \frac{eI_0}{km\epsilon} \frac{T^2}{2} \\ d &= u_a T + a_a \frac{T^2}{2} + \frac{eI_0}{km\epsilon} \frac{T^3}{6}, \end{aligned} \quad (5)$$

where u_b and u_a are electron speeds in cm./sec. at planes b and a respectively and a_a is the acceleration in cm./sec.² at plane a . Eliminating a_a and introducing the values of u_b and u_a in terms of V_{02} and V_{01} we find

$$T^3 - 6\sqrt{\frac{2m}{ek}} \frac{k\epsilon}{I_0} (\sqrt{V_{01}} + \sqrt{V_{02}}) T + \frac{12km\epsilon}{I_0} d = 0. \quad (6)$$

When the transit time T is eliminated between (3) and (6) the result is

$$I_0 = \frac{4}{9} \epsilon \sqrt{\frac{2e}{mk}} \frac{(\sqrt{V_{01}} + \sqrt{V_{02}})^3}{d^2}. \quad (7)$$

But this is precisely the Kipp current as given by (4). Equation (3), therefore, expresses the Kipp relation between current and transit time. It has thus been found that the series capacitance at the Kipp point becomes infinite, whereas the impedance between the two planes is a pure resistance.

Consider next the possibility of $J_0 \frac{T^3}{6d}$ attaining values larger than unity when the d-c. current I_0 is limited to values smaller than the Kipp value (4). The only manner in which this could happen would be for (6) to have one root T_2 such that $T_2 > T_k$ where T_k is the transit time at Kipp. To determine this (6) is transformed by introducing I_{0K} and T_K as parameters. Thus,

$$\left(\frac{T}{T_K}\right)^3 \frac{I_0}{I_{0K}} - 3 \frac{T}{T_K} + 2 = 0. \quad (8)$$

The discriminant D of (8) is

$$D = \left(\frac{I_{K0}}{I_0}\right)^2 \left(1 - \frac{I_{K0}}{I_0}\right) \quad (9)$$

and as $I_0 \leq I_{K0}$ it is clear that

$$D \leq 0. \tag{10}$$

Hence, since the discriminant in general is negative (8) has three real roots. Two of these are positive and one negative. A double root occurs for $I_0 = I_{K0}$, and the value of this root is clearly $T = T_K$. For currents smaller than the Kipp current there are two positive roots T_1 and T_2 such that $T_1 < T_K$ and $T_2 > T_K$.

Thus it becomes evident that space charge conditions corresponding to the roots T_2 result in a negative capacitance. Interpreted in terms

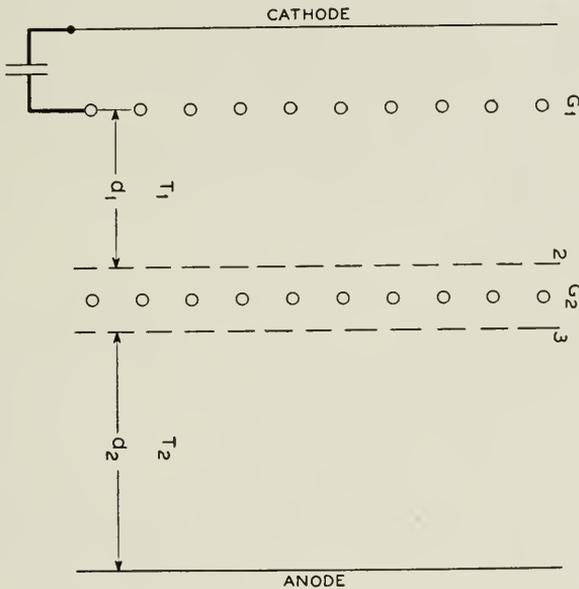


Fig. 3—Schematic of a space charge grid tube.

of Fig. 2, the roots T_1 correspond to the upper branch $\alpha\beta$ and the roots T_2 to the lower branch $\gamma\beta$. Hence, it is evident that space charge corresponding to the lower branch has a negative dielectric constant.

Consider next the system pictured schematically in Fig. 3. It differs basically from that in Fig. 1 in that a negatively polarized grid has been interposed between the planes a and b of Fig. 1. The arrangement shown in Fig. 3 may be considered to be equivalent to a space charge grid tube with the accelerating grid G_1 and control grid G_2 . The grid G_1 is assumed to be by-passed to the cathode for a-c. The planes indicated by 2 and 3 are imaginary planes located

on opposite sides of the control grid wires and are assumed to be sufficiently far away from the grid to insure that the potential distribution at these planes is essentially that of a grid-free space. Moreover, if the grid is fine-meshed these planes may be located quite close together so that the distance between them is negligible in comparison with the distances between G_1 and plane 2, and between plane 3 and the plate. This insures that the potentials of the two planes are practically equal. Since the grid G_2 draws only displacement current, the conduction currents at the planes 2 and 3 are equal. Since the potentials of the two planes are equal, the electron speeds are also equal. The boundary conditions in the plane of G_1 are the same as those used in deriving (1), i.e., constant conduction current and electron-speed. Then, following the method employed by Llewellyn² in his treatment of the negative triode, the input impedance may be found. Since the details are uninteresting the result is merely quoted. On the assumption that the plate is short-circuited to the cathode by a large condenser the input impedance between grid and cathode may be written as:

$$Z = Z_o + \frac{A_1 A_2}{\epsilon(A_1 + A_2 + G_1 B_2 + D_1 C_2)}, \quad (11)$$

where Z_o is the impedance (capacitive) between the planes 2 or 3 and G_2 and where A_1, A_2, G_1, B_2, D_1 and C_2 have the values:

$$\left. \begin{aligned} A_1 &= \frac{1}{(i\omega)^4} \left[(i\omega)^3 d_1 + \frac{eI_0}{km\epsilon} (2 - 2e^{-i\theta_1} - i\theta_1 - i\theta_1 e^{-i\theta_1}) \right] \\ A_2 &= \frac{1}{(i\omega)^4} \left[(i\omega)^3 d_2 + \frac{eI_0}{km\epsilon} (2 - 2e^{-i\theta_2} - i\theta_2 - i\theta_2 e^{-i\theta_2}) \right] \\ G_1 &= \frac{eI_0}{km\epsilon u_{02} (i\omega)^2} (1 - e^{-i\theta_1} - i\theta_1 e^{-i\theta_1}) \\ B_2 &= \frac{1}{(i\omega)^3} [a_{03}(i\theta_2 e^{-i\theta_2} + e^{-i\theta_2} - 1) + u_{02} i\omega (e^{-i\theta_2} - 1)] \\ D_1 &= \frac{1}{(i\omega)^2} \left[1 - e^{-i\theta_1} - \frac{a_{02}}{i\omega u_{02}} (1 - e^{-i\theta_1} - i\theta_1 e^{-i\theta_1}) \right] \\ C_2 &= \frac{eI_0}{km\epsilon (i\omega)^2} (i\theta_2 e^{-i\theta_2} + e^{-i\theta_2} - 1) \end{aligned} \right\} \cdot \quad (12)$$

In (12):

- u_{02} is the d-c. speed at planes 2 or 3.
- a_{02} and a_{03} are the d-c. accelerations at the planes 2 and 3 respectively.

² Loc. cit.

When the transit angles θ_1 and θ_2 are very small and when the effect of the space between control grid and anode may be ignored, evaluation of (11) yields the result:

$$Z = Z_g + r_1 \frac{1 - \frac{4}{3} \frac{J_0 \frac{T_1^2}{2u_{02}} \cdot \frac{1 - J_0 \frac{T_1^3}{6d_1}}{J_0 \frac{T_1^3}{6d_1}}{1 - J_0 \frac{T_1^2}{2u_{02}}}}{1 - J_0 \frac{T_1^2}{2u_{02}}} + \frac{1}{i\omega} \frac{1 - J_0 \frac{T_1^3}{6d_1}}{C_0 \left(1 - J_0 \frac{T_1^2}{2u_{02}} \right)}, \quad (13)$$

where

$$\left. \begin{aligned} r_1 &= \frac{J_0 T_1^4}{\epsilon \cdot 12} \\ C_0 &= \frac{\epsilon}{d_1} \end{aligned} \right\} \quad (14)$$

C_0 is the capacitance in the absence of electrons.

The input impedance is thus seen to be a series circuit composed of a resistance and a capacitance, i.e.

$$Z = \rho + \frac{1}{i\omega C}, \quad (15)$$

where

$$\frac{1}{\rho} = \frac{1}{r_1} \cdot \frac{1 - J_0 \frac{T_1^2}{2u_{02}}}{1 - \frac{4}{3} \frac{J_0 \frac{T_1^2}{2u_{02}} \cdot \frac{1 - J_0 \frac{T_1^3}{6d_1}}{J_0 \frac{T_1^3}{6d_1}}{J_0 \frac{T_1^2}{2u_{02}}}}, \quad (16)$$

$$\frac{1}{C} = \frac{1}{C_g} + \frac{1}{C_0} \frac{1 - J_0 \frac{T_1^3}{6d_1}}{1 - J_0 \frac{T_1^2}{2u_{02}}}. \quad (17)$$

In studying the capacitance and resistance as functions of space charge it is evident from (16) and (17) that space charge enters essentially through the functions $\frac{J_0 T_1^2}{2u_{02}}$ and $\frac{J_0 T_1^3}{6d_1}$. In what follows

the capacitance C_0 is assumed to be so large that the first term in (17) may be ignored. The ratio between "hot" and "cold" capacitance may then be written:

$$\frac{C}{C_0} = \frac{1 - J_0 \frac{T_1^2}{2u_{02}}}{1 - J_0 \frac{T_1^3}{6d_1}} \quad (18)$$

Write the conductance $1/\rho$ as

$$\frac{1}{\rho} = \frac{1}{r_1} \cdot F \quad (19)$$

and note that the first factor of the right member is always positive. F may be termed the relative input conductance.

Before the theoretical curves are discussed a few words about the functions $\frac{J_0 T_1^2}{2u_{02}}$ and $J_0 \frac{T_1^3}{6d_1}$ are in order. They will be expressed in terms of two parameters, i.e., φ and ξ where

$$\varphi = \frac{V_{02}}{V_{01}} \quad (20)$$

and ξ is a constant of integration which assumes different values depending upon the type of space charge present.³ In terms of these parameters one may show that:

$$\left. \begin{aligned} J_0 \frac{T_1^3}{6d_1} &= \frac{[\sqrt{\xi^{1/2} + 1} - \sqrt{(\xi\varphi)^{1/2} + 1}]^3}{(\xi^{1/2} - 2)\sqrt{\xi^{1/2} + 1} - ((\xi\varphi)^{1/2} - 2)\sqrt{(\xi\varphi)^{1/2} + 1}} \\ J_0 \frac{T_1^2}{2u_{02}} &= \frac{[\sqrt{\xi^{1/2} + 1} - \sqrt{(\xi\varphi)^{1/2} + 1}]^2}{(\xi\varphi)^{1/2}} \end{aligned} \right\} \begin{array}{l} \text{Type} \\ D_1 \\ 0 < \xi < \infty, \end{array} \quad (21)$$

$$\left. \begin{aligned} J_0 \frac{T_1^3}{6d_1} &= \frac{[\sqrt{\xi^{1/2} - 1} - \sqrt{(\xi\varphi)^{1/2} - 1}]^3}{(\xi^{1/2} + 2)\sqrt{\xi^{1/2} - 1} - ((\xi\varphi)^{1/2} + 2)\sqrt{(\xi\varphi)^{1/2} - 1}} \\ J_0 \frac{T_1^2}{2u_{02}} &= \frac{[\sqrt{\xi^{1/2} - 1} - \sqrt{(\xi\varphi)^{1/2} - 1}]^2}{(\xi\varphi)^{1/2}} \end{aligned} \right\} \begin{array}{l} \text{Type} \\ D_2 \\ \frac{1}{\varphi} < \xi < \infty, \end{array} \quad (22)$$

³ The parameters $1/\alpha$ and $1/\beta$ used by Fay, Samuel and Schockley are related to ξ as follows:

$$\xi = \frac{1}{\beta} \text{ Type } D_1$$

$$\xi = \frac{1}{\alpha} \text{ Type } D_2.$$

$$\left. \begin{aligned}
 J_0 \frac{T_1^3}{6d_1} &= \frac{[\sqrt{\xi^{1/2} - 1} + \sqrt{(\xi\varphi)^{1/2} - 1}]^3}{(\xi^{1/2} + 2)\sqrt{\xi^{1/2} - 1} + ((\xi\varphi)^{1/2} + 2)\sqrt{(\xi\varphi)^{1/2} - 1}} \\
 J_0 \frac{T_1^2}{2u_{02}} &= \frac{[\sqrt{\xi^{1/2} - 1} + \sqrt{(\xi\varphi)^{1/2} - 1}]^2}{(\xi\varphi)^{1/2}}
 \end{aligned} \right\} \begin{array}{l} \text{Type} \\ C_1 \end{array} \quad (23)$$

$$\frac{1}{\varphi} < \xi < \left(1 + \frac{1}{\sqrt{\varphi}}\right)^2$$

The parameter ξ is without physical significance for the D solutions but serves merely as a convenient constant of integration. For the solutions of C type, however, ξ is of direct physical significance. For here ξ represents the ratio $\frac{V_{01}}{V_{0m}}$. In the language used in the previous qualitative discussion of space charge it is clear that $\xi = \frac{1}{\varphi}$ is the point where a potential minimum sets in and $\xi = \left(1 + \frac{1}{\sqrt{\varphi}}\right)^2$ is the Kipp point. Calculations for C_2 solutions are not included.

The result of the calculations of capacitance and conductance are shown on Figs. 4 to 8. Figure 8 is an enlargement of Fig. 7 for small

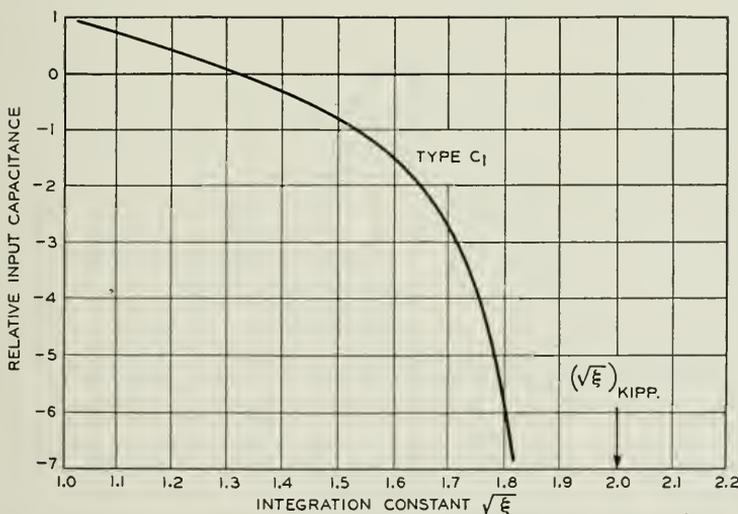


Fig. 4—Relative input capacitance of an idealized space charge grid tube for different space charge conditions (voltage ratio $\varphi = 1.0$).

values of F . Three values of the parameter φ were selected, i.e., 0.04, 0.25 and 1. The curves demonstrate regions of negative capacitance as well as of negative conductance. Note that as the parameter φ is made smaller both the capacitance and conductance pass through

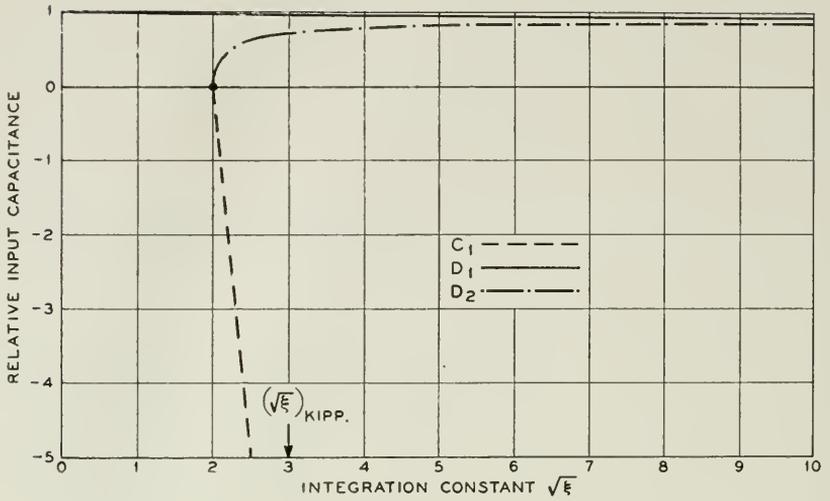


Fig 5—Relative input capacitance of an idealized space charge grid tube for different space charge conditions (voltage ratio $\varphi = 0.25$).

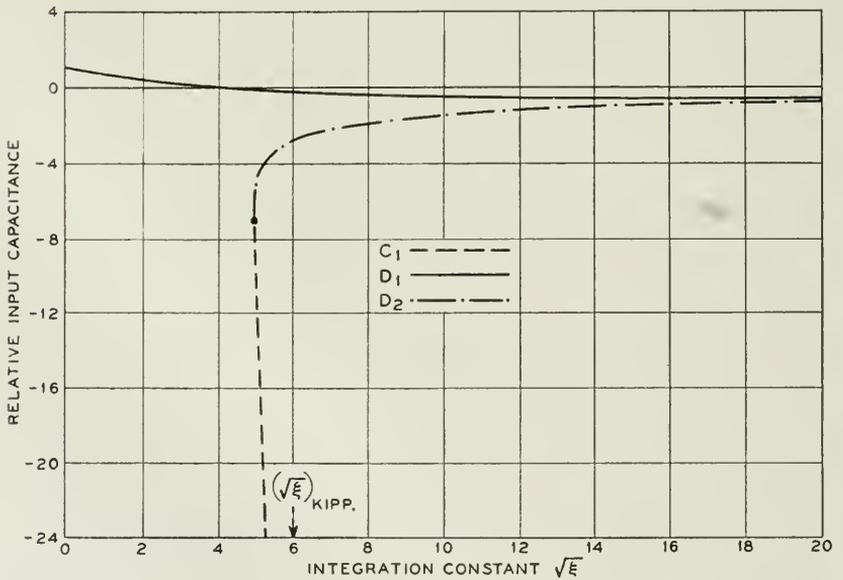


Fig. 6—Relative input capacitance of an idealized space charge grid tube for different space charge conditions (voltage ratio $\varphi = 0.04$).

zero for smaller values of ξ . Consider for instance the case for which $\phi = 0.04$. The capacitance is here negative even in regions of small space charge and it is seen that over a wide range of the parameter ξ the capacitance is nearly constant. Both capacitance and conductance

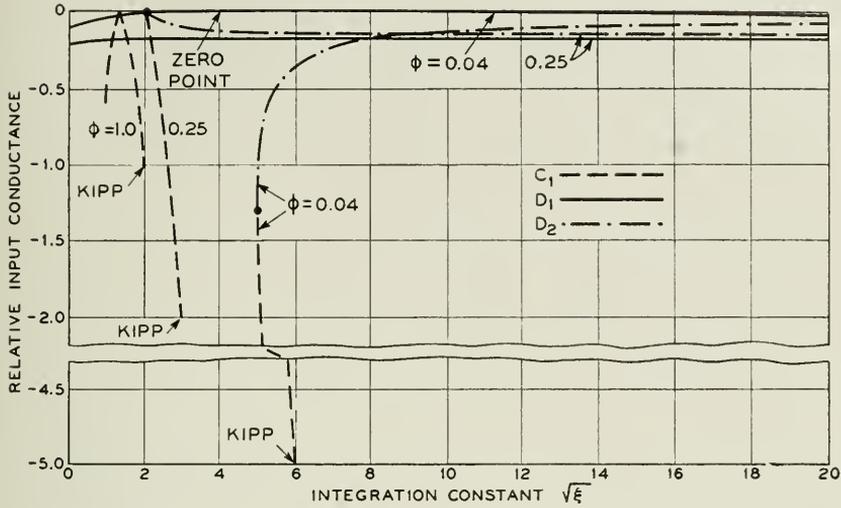


Fig. 7—Relative input conductance of an idealized space charge grid tube for different space charge conditions.

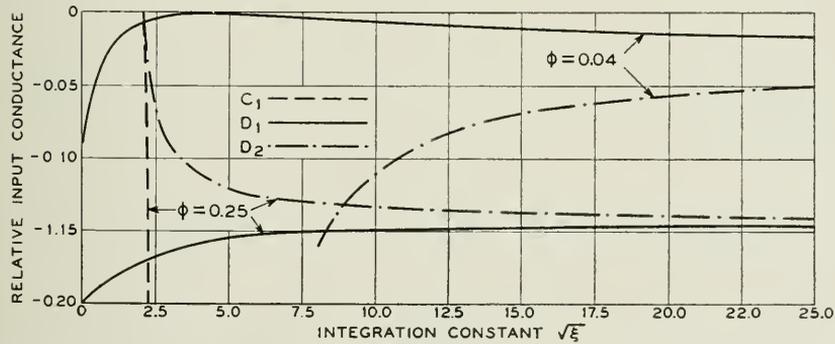


Fig. 8—Relative input conductance of an idealized space charge grid tube for different space charge conditions.

pass through zero for the same value of ξ and one sees that in case of $\phi = 0.04$ no critical adjustment seems necessary. The conclusion is thus arrived at that the idealized space charge grid tube may be made to operate without input capacitance or loading up to moderately

high frequencies. In addition, the possibility of designing an electronic negative capacitance is opened up and this is a highly desirable objective.

However, it must be kept in mind that in the discussion, effects such as velocity distributions, electron deflections and dispersion forces have been neglected. Furthermore, in an actual tube the capacitance C_o must also be considered. At the Kipp the capacitance C is equal to $-\infty$; therefore, as the Kipp point is approached a positive capacitance is expected.

The capacitance and conductance both pass through zero when

$$J_0 \frac{T_1^2}{2u_{02}} = 1, \quad (24)$$

or when

$$u_{02} = J_0 \frac{T_1^2}{2}. \quad (25)$$

But in general

$$u_{02} = u_{01} + a_{01}T_1 + J_0 \frac{T_1^2}{2}, \quad (26)$$

where u_{01} and a_{01} are the d-c. speed and acceleration in the plane of grid G_1 , Fig. 3.

The relation between initial speed and acceleration for zero capacitance and conductance is, therefore:

$$u_{01} + a_{01}T_1 = 0 \quad (27)$$

and since both u_{01} and T_1 are inherently positive, the initial acceleration must be negative. For the capacitance to be negative the requirement is obviously

$$u_{01} + a_{01}T_1 < 0. \quad (28)$$

Necessary requirements for a negative capacitance are thus a finite electron speed and a retarding field at the plane of injection.

PART II

EXPERIMENTAL

In this section some experimental results will be discussed. The measurements all refer to the capacitance between control grid and ground of some experimental tubes. The tubes were cylindrical in structure and contained two positive grids close to the cathode followed by a negative control grid. The first positive grid has the essential function of controlling the magnitude of the current whereas the

second determines the initial speed with which the electrons enter the space adjacent to the control grid.

In Fig. 9 the measured input capacitance and plate current are shown as functions of the voltage $V_{\theta 1}$ of the first grid. It is seen that as the plate current increases the capacitance gradually decreases, passes through zero somewhat before the plate current has reached its maximum and then becomes negative; it passes through a minimum and then gradually assumes a positive value equal to about twice the cold capacitance. This behavior of the capacitance is typical of the formation of a virtual cathode, which in the present instance appears

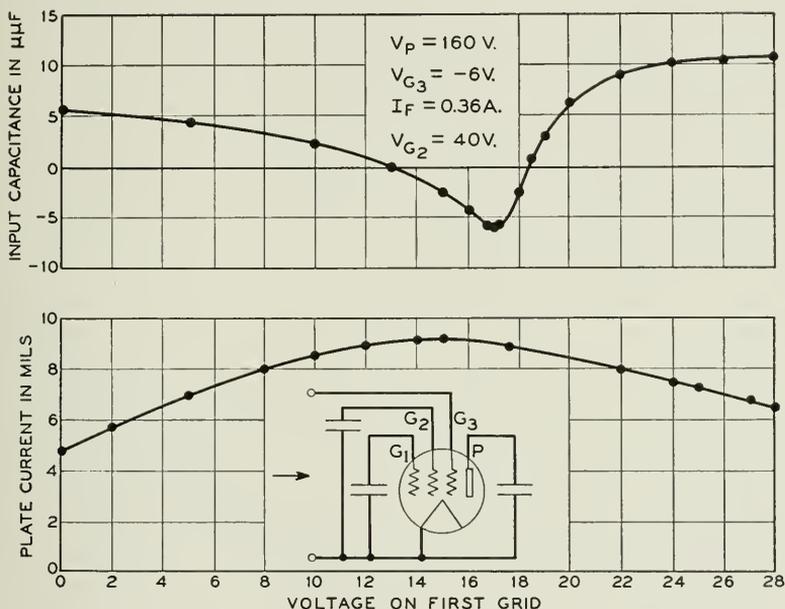


Fig. 9—Measured input capacitance of experimental vacuum tube no. 55 (grid signal = 0.18 volts r.m.s. at 50 kc.).

to be gradual. The negative capacitance is present in a small interval of the voltage $V_{\theta 1}$ immediately before and perhaps also during a part of the virtual cathode formation.

Figure 10 shows the measured input capacitance as a function of the control grid bias $V_{\theta 3}$ for several values of the voltage $V_{\theta 1}$. For comparison purposes the corresponding plate currents are also shown. In the region of low plate current where a virtual cathode is present the capacitance is positive and larger than the cold capacitance. Where the plate current curve starts to bend over, that is, where the

virtual cathode starts to release, the capacitance decreases. It is negative in a small domain and turns ultimately positive. The plate current curve corresponding to $V_{o1} = 30$ volts indicates that the

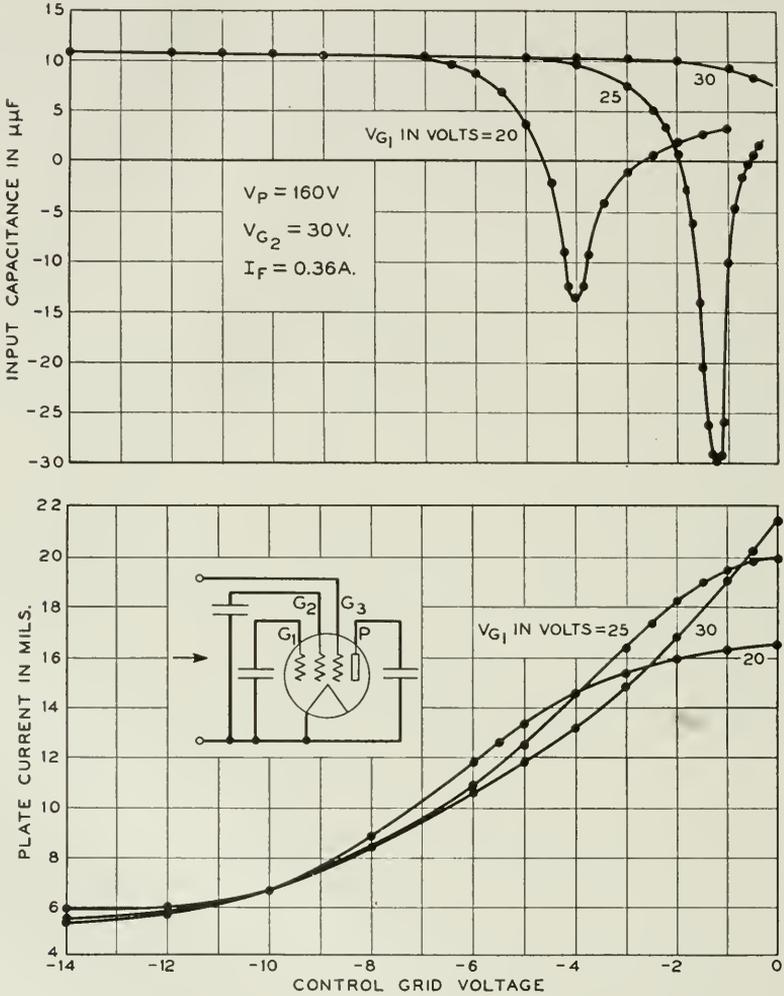


Fig. 10—Measured input capacitance of experimental vacuum tube no. 55 (grid signal = 0.18 volts r.m.s. at 50 kc.).

virtual cathode is present throughout the range of negative bias and the corresponding capacitance curve is positive everywhere.

Figure 11 refers to capacitance measurements on a tube in which a

Kipp occurred. Again the capacitance behaves essentially the same except that it suddenly jumps from a large negative value to the positive value corresponding to virtual cathode operation.

In comparing the theoretical and experimental results, it is seen that the theory gives predictions which are broadly in accord with experiments.

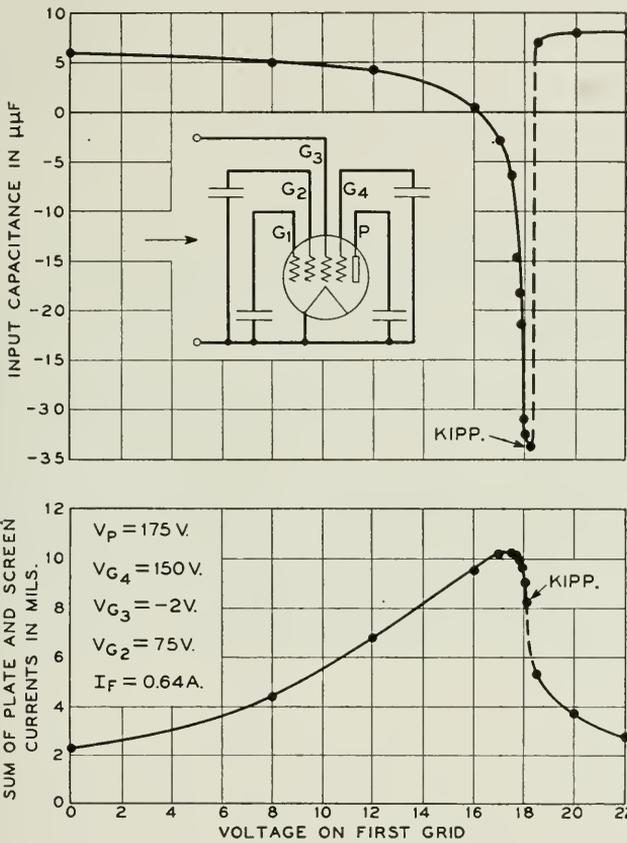


Fig. 11—Measured input capacitance of experimental vacuum tube no. 59 (grid signal = 0.05 volts r.m.s. at 50 kc.).

ACKNOWLEDGMENT

I am indebted to Mr. E. J. Buckley for assistance in the experimental work, to Miss M. Packer for the numerical calculations, to Mr. C. A. Bieling for the mechanical tube design, and last but not least to Dr. F. B. Llewellyn for stimulating discussions.

Plastic Materials in Telephone Use*

By J. R. TOWNSEND and W. J. CLARKE

ORGANIC plastics are used extensively in the manufacture of telephone apparatus and equipment. They belong to the class of materials known as insulators but are very often employed not only for their electrical properties but for their unique manufacturing and structural possibilities. Good insulating materials are very important in the telephone field although the voltage and current used are much smaller than in the power field. Progressive improvement in transmission, especially for long distance telephone service, has required that the telephone industry as a whole provide sensitive instruments and that there be a minimum loss of the electrical impulse due to leakage through insulating materials.

Rubber became at one time the most universally used insulating material in telephone apparatus. Where superior insulating properties are required, rubber has been employed not only in the soft vulcanized form as a covering for wire but as hard rubber. Its use was considerably curtailed as a molding material during the period of the world war due to the high price of rubber. This stimulated the substitution of phenol plastics which were found to produce more permanent parts. Although rubber must be classed as an organic plastic, it will not be dealt with here except in passing since it comprises a large field in its own right and quite distinct from that of the synthetic plastics. In recent years rubber has been greatly improved in life, stability, light sensitivity and resistance to cold flow so that its technical uses in the telephone plant are again increasing.

Shellac and asphalt plastics, both natural materials, were among the early important plastics employed in the telephone art. A shellac compound is still the best material for a panel system commutator where there are many long and delicate contact segments and where the principal problem is to obtain accurate location between the insulation and the brass segments together with uniform wear. The low molding temperatures and pressures for the shellac mica compound contribute to the success of the manufacture of this part. (See Fig. 1.)

Other early plastics that have found some limited uses are cellulose

* Presented before the Organic Plastics Section of the Paint and Varnish Division of the American Chemical Society, Baltimore, Md., April 3-7, 1939.

nitrate and casein. However, the cellulose nitrate plastics were only sparingly employed for telephone construction because of the serious fire hazard. Casein has long been used for key buttons and similar minor applications.

The great expansion of the use of molded plastics in the telephone plant really began with the development of organic materials which had superior manufacturing and structural characteristics over other materials. The newer plastics are of value, therefore, as much from the economies of manufacture as from their superiority over the previously used materials.

Plastics are conventionally divided into two groups: (1) the thermoplastics and (2) the thermosetting plastics. The first, considered as

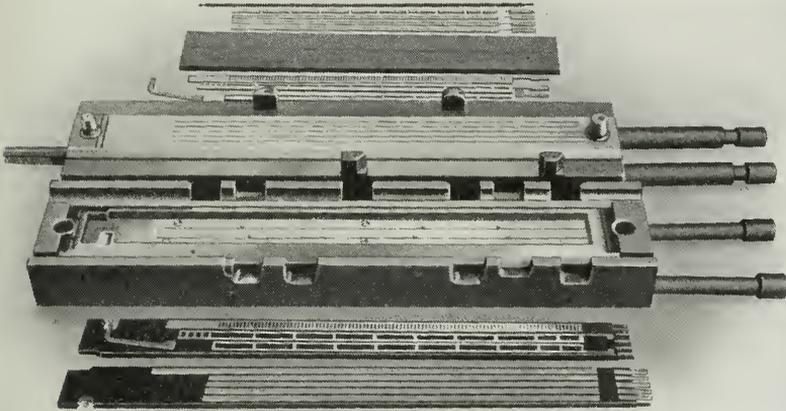


Fig. 1—Panel system commutator (flash type mold).

organic materials, are permanently soluble and fusible as well as fairly rigid at normal or working temperatures and may be deformed under heat and pressure. The second are initially thermoplastic and become insoluble and infusible after a period of time upon application of heat and pressure. These important properties are due to the chemical nature and molecular structure of the materials. All synthetic plastics are polymeric substances, that is, they are the result of a polymerization or condensation from simple organic molecules by linkage of the molecules in fairly definite ways. By a polymerization reaction is meant a reaction in which a more or less considerable number of molecules unite to form larger complexes of the same chemical composition. A condensation reaction on the other hand is one

in which molecules join together to give a larger complex but during the reaction there is a separation of a small amount of water, alcohol or some other substance, so that the final chemical composition is not quite the same as at the start.

Those materials which are thermoplastic and readily soluble owe these characteristics to the fact that the molecules are linked essentially in chain-like fashion. The forces holding the chains together are of a secondary valence character, that is, the chains are free to move apart when heat is applied or a solvent is present. Vinyl, acryl and styryl resins are typical thermoplastics and incidentally each is formed from a monomeric material containing the characteristic ethylene grouping $\text{CH}_2 = \text{C} <$. The properties of synthetic materials of the thermoplastic type vary with the average chain length and distribution, and with the nature of any side groups which may be attached to the main hydrocarbon chains. The materials do not have sharp melting points as do more simple organic substances but soften gradually when they are warmed. Usually on further heating decomposition occurs to the monomeric form before any rapid flow point is reached.

The molecules of thermosetting plastics are initially in chain-like form also, although the chains are generally much shorter in length than those of the thermoplastics. On heating, the material fuses and the chains become cross-linked sufficiently to give a permanent and rigid three-dimensional structure. In the case of a phenol-formaldehyde resin the linkage between chains is directly through CH_2 groups (or according to some investigators through $-\text{O}-\text{CH}_2$ or $-\text{CH}_2-\text{O}-\text{CH}_2$ -groups) and the force necessary to separate the chains is therefore high and of an order equal to that needed to break up any complex organic molecule. Infusibility and inability to go into solution are consequently the prominent characteristics of a phenolic resin in the heat-hardened form. In other materials belonging in this class the cross-linkage may be brought about through oxygen or sulphur atoms though the hardening action sometimes occurs more slowly. Certain oxygen convertible alkyd resins for example are particularly useful as organic finishes because they may be greatly hardened and in this case very much toughened by a baking process, the necessary oxygen for cross-linkage of the polymeric resin being absorbed from the surrounding air.

The properties of the more modern materials including the phenol plastics, the cellulose derivatives and the ethenoid (vinyl, acryl and styryl) type plastics, have yet to be fully evaluated but certain electrical and mechanical characteristics of these materials have already resulted in their adoption to a greater or lesser extent in the telephone

plant. Phenol plastics have been employed in the molding of the regular telephone hand set, recent production being in excess of 1,000,000 units per year. Cellulose acetate is widely used in foil form. Plastics in the form of synthetic organic finishes are used for protection, decoration and insulating purposes on apparatus and equipment. For the purpose of discussion, plastics in the telephone plant may be grouped as follows:

1. Molding plastics.
2. Sheet materials (phenol fiber, acetate foil, etc.).
3. Synthetic organic finishes, adhesives and miscellaneous special items.

OBJECTIVES OF IDEAL TELEPHONE PLASTICS

Telephone apparatus and equipment are not sold as consumption goods but the service rendered by it is sold to the subscriber. Good service means a minimum of breakdown due to replacement of malfunctioning parts, repairs and maintenance. High maintenance costs are inconsistent with the best service at the lowest cost. Uniformly high quality of materials throughout the economic life of the telephone plant is therefore essential.

The molding plastics and sheet materials account for the bulk of the plastics used in the telephone plant and the objectives of these materials are similar enough to permit them to be listed together. There are given below the general and specific properties that must be considered in such materials when they are to be used in the telephone industry. The level of quality demanded in specific properties will obviously depend on the application.

1. General requirements.
 - a.* Strength, hardness, toughness.
 - b.* Low density (to decrease mechanical inertia, aid manual use).
 - c.* Chemical inertness in air, or in contact with other materials.
 - d.* Resistance to humidity (minimum of swelling and shrinkage with variations of moisture content of the air).
 - e.* Ability to withstand temperature, heat and cold without too great impairment of strength and shape.
 - f.* Ability to reproduce die surface accurately and give good appearance to finished part.
 - g.* Light stability.
 - h.* Relative non-inflammability.
 - i.* No odor, no harm to the skin.
 - j.* Resistance to insect attack.

2. Specific mechanical properties.
 - a. Transverse strength.
 - b. Impact strength.
 - c. Cold flow.
 - d. Shrinkage.
 - e. Wear resistance.
 - f. Machinability.
3. Specific electrical properties.
 - a. Insulation resistance
 1. as affected by humidity.
 2. as affected by light.
 - b. Dielectric constant.
 - c. Power factor.
 - d. Dielectric strength.
4. Moldability.
 - a. Free flow at moderate temperature and pressure.
 - b. Favorable setting characteristics.
 - c. Short molding cycle.
 - d. No tendency to stick to die.
 - e. No abrasion of die surface.
 - f. Minimum shrinkage in mold.
 - g. Low bulk factor.
5. Economic considerations.
 - a. Low density materials preferred.
 - b. Cost.
 - c. Die life.
 - d. Utilization of scrap.
 - e. Molding cycle time.
 - f. Trimming and finishing characteristics.
 - g. Refinishing or maintenance.

The objectives of an ideal plastic in the telephone industry depend upon the use to which the material will be put in the telephone plant. The material may be a structural member, an insulator, or both, and may be in the hands of the public or in a telephone exchange. All of the above requirements need not be met but excellence in a majority of these properties is generally desirable.

THE MOLDING OF TELEPHONE PARTS

Molding involves consideration of (1) the molding compound (2) the die (3) the press (4) the heating and cooling system (5) method of ejecting part from mold (6) finning and trimming methods. Regardless of the type of plastic, these operations are necessary.

As to the molding compounds, the Bell System obtains from the suppliers whatever materials are needed and this is generally true for the industry. These range from the plain wood flour-filled phenolics to the various thermoplastics depending on the application. Exact compositions are seldom specified in order that the manufacturer be given all possible opportunity to exercise his ingenuity to produce satisfactory quality material.

The die or mold is such an important item in the molding of a material that several points should be emphasized about it. The dies are always expensive. Every part is an individual design problem, involving flow of material in the cavity, use of inserts, opening and closing, the clamping of die parts under high pressures, and alignment. Everything possible must be done to reduce the complication of the die; eliminate inserts if possible, provide generous fillets, ample taper for removal of parts, and facilitate flow of the compound. The Bell System has found it advantageous to make most of its own dies. The conventional boring, milling and hobbing processes are used. Very little success has been had with other methods, such as casting with hard alloys.

In spite of the expensive and time-consuming effort that must be encountered in designing and building a molding die, the finished die when properly used represents one of the most indestructible tools of modern manufacture. The parts are finished, require little or no surface treatment, the dimensions are accurate and sub-assembly operations may already be completed as the part emerges from the die.

The dies used are of the three general types: the open or flash die, the closed or positive die, and the injection die. The open type consists of two parts which come together at a cutting edge or ridge that surrounds both halves of the die. Since this ridge or cutting edge must withstand the full pressure of the press it is usually about one-eighth of an inch wide. Flash dies are all relatively simple, readily loaded, and the charge need not be measured accurately, any excess being forced out as the two halves close until the cutting edges come into contact. Such molds may be used for any molding compound not requiring high pressure and which does not have high die shrinkage. Shellac-mica commutators (Fig. 1) are manufactured by means of a flash die.

The positive type die consists of a plunger and a cavity shaped to produce the finished part. The plunger may aid in shaping the part. Only enough material is placed in the die to make the part. The material may be weighed in separate charges or preformed by a separate

operation. The cavity is loaded and the plunger forced down, forming the part. Multiple cavity molds cannot be loaded exactly alike and hence some provision must be made for the escape of excess material, forming a flash that must be subsequently removed. Such dies must be carefully designed so that the fin is located for easy trimming and to provide the best appearance. These multicavity molds are frequently called semi-positive molds to differentiate them from a truly positive mold where there is little flash. A typical telephone part made in a semi-positive mold is the handset handle shown in Fig. 2.

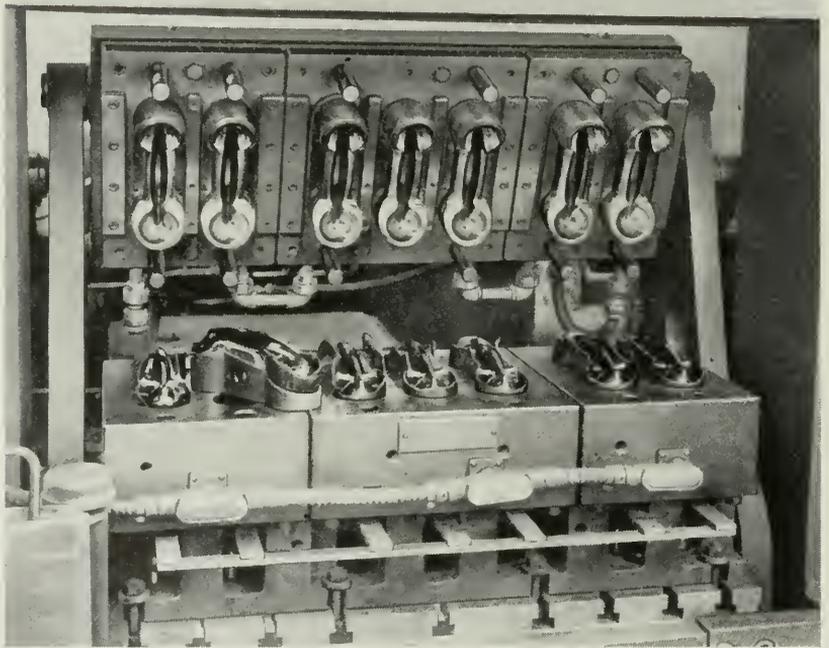


Fig. 2—Mold for handset handle.

The recently developed process of injection molding which consists in forcing plastic material through a nozzle into a completely closed die from an external compression chamber is also being used and promises to alter many of the present operations. Since the opening and the closing of the die are not related to the application of pressure, greater freedom of design is possible. Furthermore, since full pressure is not applied by this method until the die cavity is completely filled with material, the material already in the die tends to support inserts and

hold them in their true position. Hence more delicate inserts are possible by this method of manufacture. Since the material is enclosed in an auxiliary pressure chamber and is not exposed to the atmosphere, greater freedom from room dust is possible, rendering this method ideal for colored plastics. A terminal block used to terminate the subscribers telephone cord is made from thermoplastic molded cellulose acetate compounds as shown in Fig. 3.

Finishing and trimming methods are largely determined by the design and the class of service required of a molded part. In the case of the injected cellulose acetate terminal block mentioned above the gates are trimmed off by a simple trimming punch and the scrap is

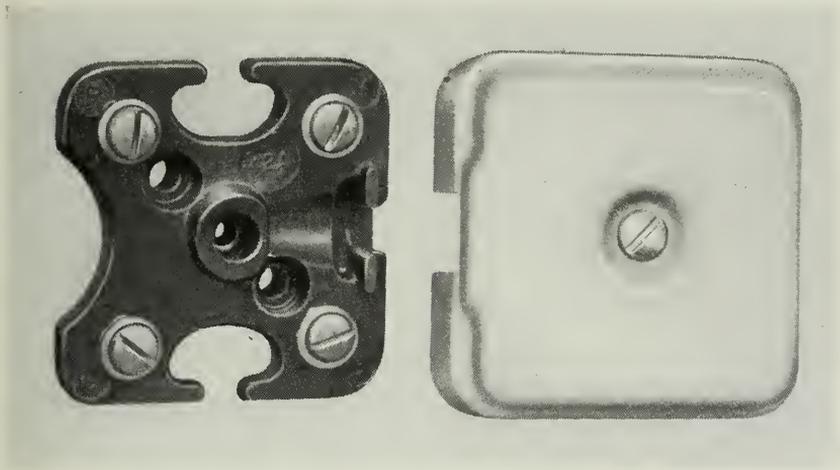


Fig. 3—Terminal block.

reused. This part is covered in service and the appearance of the block is therefore not a major factor.

The telephone handset, since it is in the hands of the public, must be carefully finished for two reasons: (1) to avoid surface roughness and (2) to provide good appearance. The handset handle was originally ground along the fin left by the semi-positive mold and then buffed. The operation was not only expensive but tended to grind off a large portion of the surface of the handle. This removed the resin-rich surface and tended to expose the filler of the phenol plastic molding compound, thus reducing the appearance life of the handle. The more recent product of the Bell System is being grooved along the die parting line. This removes the fin, a minimum of the resin-

rich surface and does not detract from the appearance of the handset. Automatic grooving machines were developed for this purpose. Figure 4 shows a grooved handset handle.

It has been found necessary to pay close attention to the design in order that die parting lines, ejector pin marks, gate marks and the like will appear at points where they may be readily eliminated by simple trimming and grooving operations, or where they may be left without objection to appearance or function of the part.



Fig. 4—Grooved handset handle.

GENERAL TEST METHODS AND REQUIREMENTS

The most satisfactory test is one that can be applied to the finished part to measure the ability of that part to perform its function satisfactorily in service. This ideal is seldom realized, not only because of the difficulty of defining the service requirements but of finding tests that are wholly representative of service conditions. It is customary, therefore, to apply a series of tests whose sum total will approach the ideal as nearly as practicable. Molded organic plastic parts are different from parts made from most other materials in that the molding process may modify them and render them quite different from the raw material. In the case of thermosetting compounds this is particularly true.

Tests are in the main applied, therefore, to a molded part of representative specimen of the fabricated material. In the telephone plant the items that are of most importance are strength, both transverse and impact, permanence of form, appearance, effect of moisture and drying on swelling and shrinkage, insulation resistance, electrical breakdown potential, and reaction on adjacent materials. Methods

of making all of these tests have been worked out and are supplemented by apparatus tests made on the manufactured product.

There is no known test that will measure completely the quality of a phenol plastic part and it is necessary to resort to the expedient of testing standard bars molded under specified conditions. Five bars are molded in a positive type die as shown in Fig. 5. The step arrange-

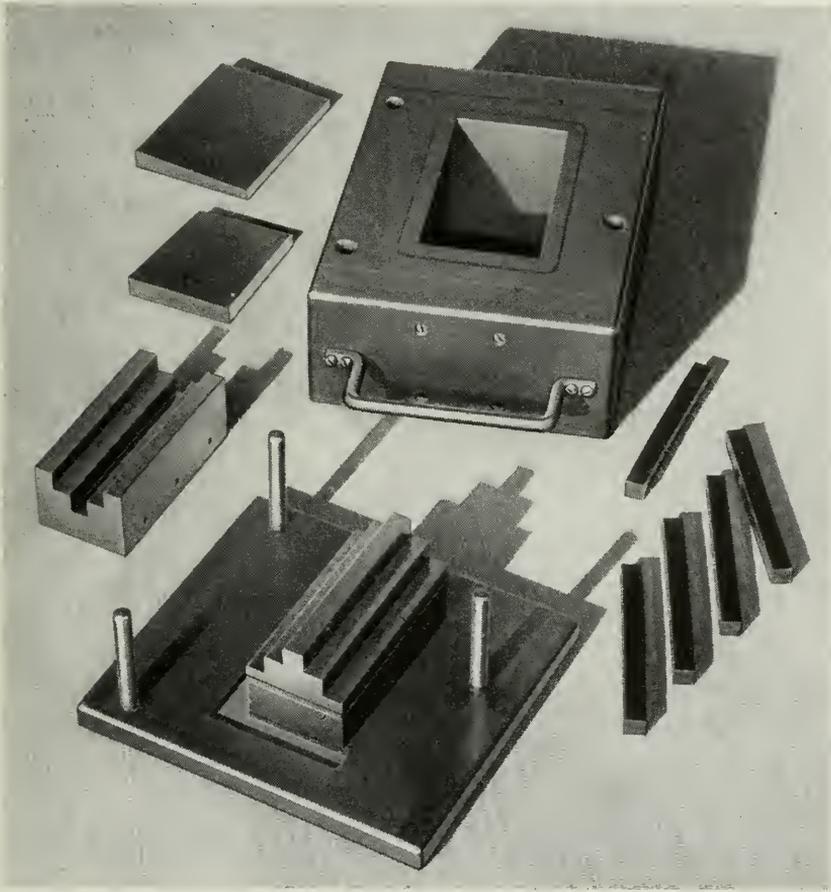


Fig. 5—Specimen mold and bars.

ment provides for flow within the die similar in many respects to the molding of actual parts. The test bars provide transverse or flexural strength, impact insulation and cold flow test specimens. The methods used, whenever possible, are those of the American Society for Testing Materials.

The cold flow test is specially designed to note the distortion of materials which in service are under pressure, such as spring pileups, inserts and apparatus in the form of banks and terminal blocks. The specimen which is $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}$ " in size is first conditioned at 150° F.

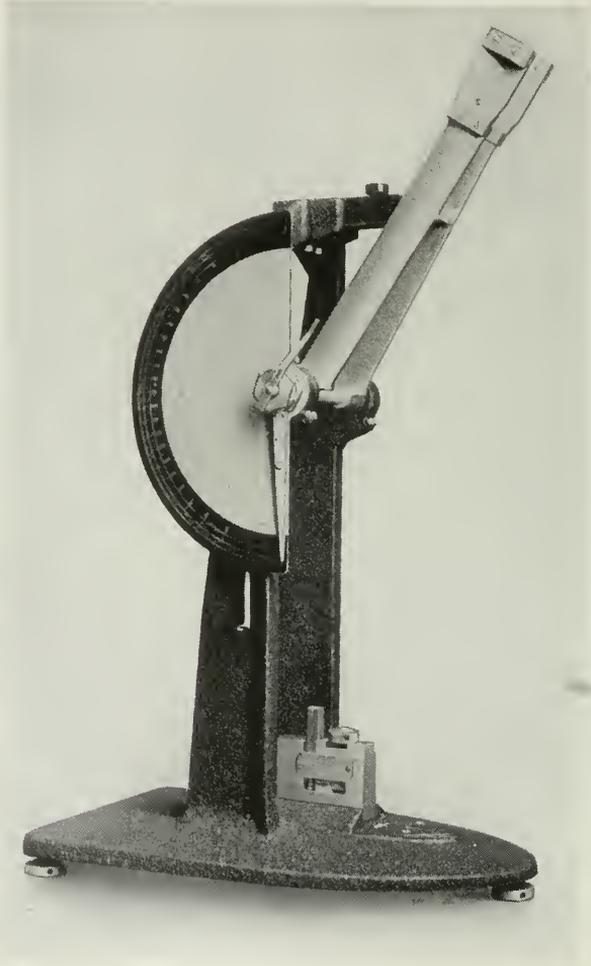


Fig. 6—Impact machine.

for 4 hours and then to 90 per cent R.H. and 85° F. for 68 hours to permit absorption of moisture. It is then held under 4000 lbs. per square inch for 24 hours at 120° F. $\pm 1^\circ$ and the percentage change in thickness determined.

TYPICAL APPLICATIONS OF THERMOSETTING PLASTICS

Phenol Plastics

A typical phenol plastic telephone part is the handle for the handset. This part is molded in a multiple semi-positive die of the kind described above.

In addition to the laboratory tests on the raw material, samples of the molded handles must withstand a dropping test. After being conditioned, handles representative of a given lot are equipped with transmitters and receivers and dropped down a nearly vertical chute to strike on a steel block. The test is made by dropping first at 36 inches and then increasing the drop in increments of 2" until the handle breaks.¹ Normal product handles will withstand a drop of 55 inches on a steel block without failure.

The electrical properties of phenol plastic compounds are adequate for most uses in the telephone plant. Two grades are recognized, however, the mechanical and the electrical. Fully 90 per cent of the uses involve the mechanical grades. For certain high-frequency insulation purposes special mica-filled phenol plastics are used in place of the regular wood and cotton filled varieties.

One of the outstanding disadvantages of a phenol plastic is the ease with which it carbonizes on exposure to electrical arcing. For this reason phenol plastic compounds have only a limited use for commutators and similar applications. However, in addition to handsets they have proved of value for mouthpieces, receiver cases, subset housings, non-magnetic coil forms, coil cases, jack mounting blanks and terminal blocks.

Phenol Fiber

Phenol fiber for telephone apparatus is made of alpha cellulose paper, Kraft paper and rag paper by the usual impregnation with a suitable phenol resin varnish and lamination of a number of sheets under heat and pressure. The most important requirement for the paper is that it shall be pure, clean and free from electrolytes. The paper is carefully tested for chlorides, conductivity of water extract and for alcohol soluble materials. Several grades of phenol fiber are necessary to meet the requirements, some of which are largely mechanical and others electrical.

The principal tests for phenol fiber are cold flow and shrinkage, insulation resistance, corrosion tendency, arc resistance, transverse strength and impact. Arc resistance applies to the case where wiping

¹ "The Impact Testing of Plastics," Robert Burns and Walter W. Werring, *Proc. A.S.J.M.*, 19, Vol. 38, 1938.

contacts cause an arc to flash across the surface of the phenol fiber. This is a condition peculiar to telephone apparatus and a special test has been designed which simulates service conditions. An insulator cam (see Figs. 7 and 8) is prepared as the test specimen. The cams are rotated at a speed of 10.5 to 11.5 revolutions per minute. Two metal cams are attached concentrically with each surface of the phenol fiber cam. The cams are rotated at a speed of 10.5 to 11.5 revolutions per minute. Two metal cams are attached concentrically with each surface of the phenol fiber cam. Attached to a brush that wipes over the cam tensioned to

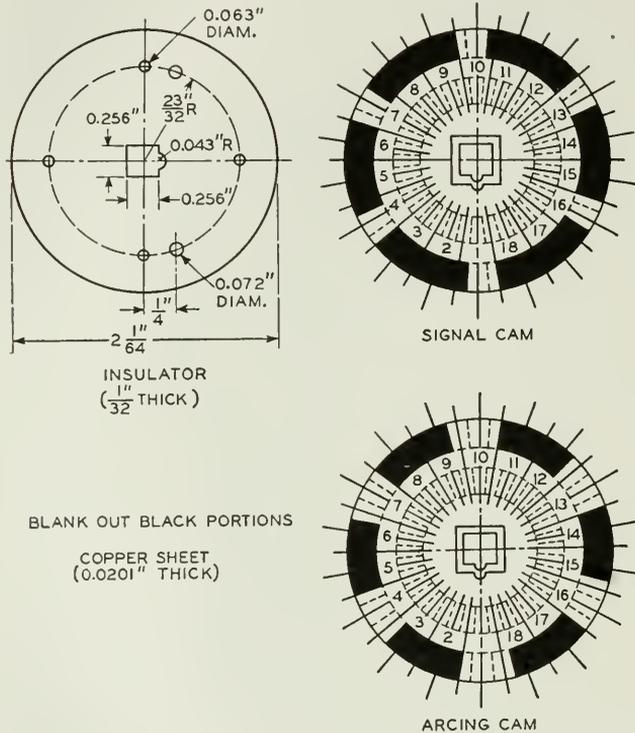


Fig. 7—Arc resistance test cam.

a pressure of 60 grams is a circuit containing $5\frac{1}{2}$ counting relays which supplies a severe inductive load. This is representative of a severe service condition. Failure is indicated when tracking of carbonaceous material shorts a cam segment the distance of 15° or when the $\frac{1}{32}$ " thick material is punctured, and the test is then stopped automatically. A good grade of fiber will resist over 1,800 revolutions whereas a poor grade will fail in 4 to 100 revolutions.

Phenol Fabric

Phenol fabric is similar to phenol fiber except that it is made with fabric instead of paper. It is generally used for its mechanical strength since its electrical properties are inferior to fiber. Phenol fabric is used in tools where high strength and resistance to impact and bending

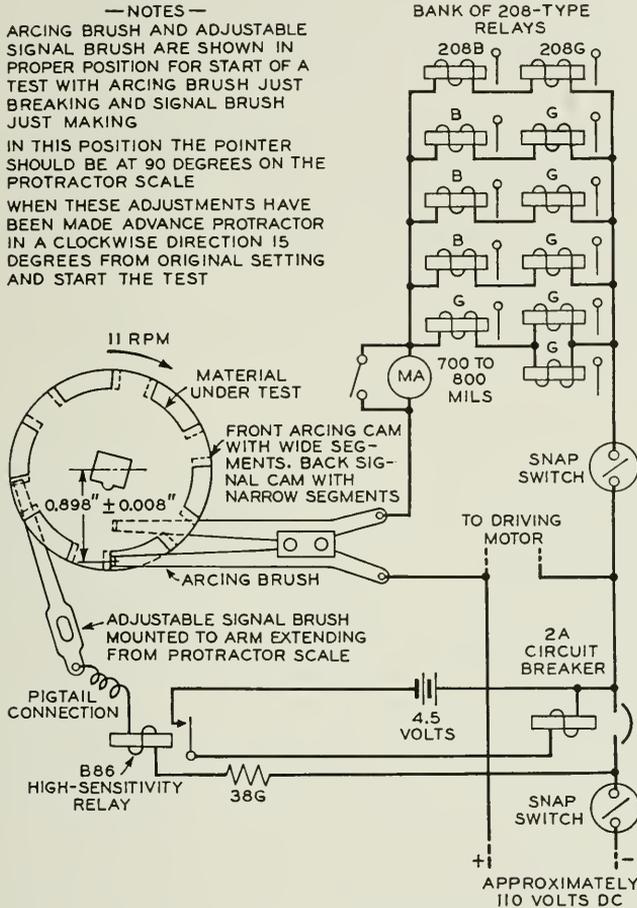


Fig. 8—Circuit for arc resistance test.

are necessary, in terminal plates, cable terminals, gears and in general where phenol fiber does not have suitable structural properties.

Urea-Formaldehyde Plastics

Condensation products made from urea and formaldehyde have attractive possibilities as thermosetting plastics. Relatively light-

fast colored parts with porcelain-like surface luster are possible with these materials. The molding cycle is slightly shorter than for phenol plastics and the material at first is somewhat more fluid. Tight molds are therefore necessary in order to get sufficient pressure.

The principal difficulty with urea-formaldehyde plastics has been that on exposure to heating and cooling or humidification and drying cycles, there is a tendency toward cracking, particularly at changes in section and around inserts. Molded into uniform thin sections without inserts they are reasonably satisfactory plastics.

At present there are practically no urea-formaldehyde plastics employed in telephone apparatus because the wide continental climatic conditions and exacting requirements will not permit their use. Recently there have been improvements made from a stability standpoint and it is believed future application may be found for these plastics, particularly in view of their color permanence.

APPLICATIONS OF THERMOPLASTIC MATERIALS

Cellulose Acetate

The principal use of cellulose acetate is for interleaving in coils. Various relay coils are made of layers of cellulose acetate over which a layer of enamel coated copper wire is wound. Layer upon layer of wire and acetate sheet form a coil. These are then assembled on a core and spoolheads attached. One of the phenol fiber spoolheads has a surface coating of cellulose acetate and the winding is pressed against this spoolhead and dipped in acetone. This dissolves or softens the exposed edges of acetate and the whole coil is firmly secured to the spoolhead.

Cellulose acetate is used for this purpose since it is practically inert as regards corrosion of the fine copper wire in contact with it and in this respect it is superior to any known material. It is permanent, reasonably fireproof and has high insulation resistance. Two grades of cellulose acetate sheet are used in telephone practice. These are the window grade used as a window or covering over designations and an electrical grade for coil use. The principal tests for the electrical grade are insulation resistance, shrinkage, and resistance to burning.

The principal use of molding grade of cellulose acetate is for the terminal block (Fig. 3) mentioned above. Here the application is mainly structural since it has more than adequate electrical insulation. Another application for cellulose acetate is a test strip where a surface layer of acetate over phenol plastic avoids carbonization of the latter.

Acrylate Resins

The clear water-white plastics derived from the polymerization of the esters of acrylic and methacrylic acid are at present being used only for windows, viewing lenses on designation strips and other optical uses in telephone apparatus. This is a new plastic and applications will no doubt develop in time, taking advantage not only of color but of the mechanical and electrical properties and insensitivity to moisture.

Vinyl Resins

Vinyl acetate and vinyl chloride polymers and co-polymer products form interesting thermoplastic resins. Their use in communication work has been limited so far to phonograph records, where the resistance of the co-polymer to warping due to humidity and the superior wear resistance of the plastic have been the important factors. The advantage of non-inflammability imparted by chlorine is more than offset by the acidic nature of the fumes given off from vinyl chloride polymers when heated or burned and this has discouraged use of these materials in the telephone plant.

Polystyrene

There have been no important commercial applications for polystyrene as yet in Bell System telephone communication although much experimentation is being carried on with polystyrene plastic. The low electrical losses of this material make it of special interest in high-frequency work but its mechanical properties have not been satisfactory. In Germany and Italy it is reported that polystyrene has been employed as an insulating plastic in various cable structures for experimental and commercial use. However, in this country the spacing insulators for the coaxial cable from New York to Philadelphia have been made of a special grade of hard rubber which has proved to be a tougher material for the purpose.²

Synthetic Coatings

An important application of synthetic organic materials in the telephone plant is in the finishes that are put on Bell System apparatus. There are three major reasons why such finishes are needed—(1) for the improvement in appearance of certain fabricated parts, especially the exposed portions of subscriber station apparatus, both in private homes and public places, (2) for the mechanical and chemical protec-

² "Systems for Wide-Band Transmission over Coaxial Lines," L. Espenschied and M. E. Strieby, *Bell Sys. Tech. Jour.*, October 1934; and *Elec. Engg.*, Vol. 56, 1937.

tion of the underlying structural material which is usually a metal, and not infrequently, (3) for electrical insulation purposes. Other minor reasons for finishes exist on special apparatus. Many parts are fashioned from such metals as steel, brass, aluminum and zinc alloys. After such practices as punching and die casting the surfaces of these parts are left in an unsightly condition, and furthermore unless protected, they may soon begin to corrode. In certain cases electroplated finishes may be employed to advantage, but organic finishes, because of their low cost and ease of application, find wide use. A good organic finish for telephone apparatus must not only have a lasting decorative value but must also protect the parts against the great variety of conditions to which the apparatus is exposed.

For example, the common black finish which is applied to various parts of subscriber station apparatus, such as the zinc alloy handset mounting, coin collector boxes and metal bell boxes must be sufficiently tough and adherent to withstand perspiration, impact and severe abrasion. Rigorous tests have been applied to find the most durable finishes for such parts. They must maintain their appearance so as to harmonize with the smooth molded black phenol plastic parts. Advantage has been taken of the recent improvements in synthetic resin finishes and a modified alkyd resin vehicle has been employed in the present black enamel. A thorough baking is given to the enamel which results in a more durable finish for telephone apparatus than the former black japan.

There are a number of applications for synthetic finishes where corrosion protection is important from the standpoint of the proper functioning of the apparatus. The aluminum diaphragms in marine and aviation loud speakers and in sound power instruments are protected by a baked finish containing a heat-hardening phenolic resin vehicle into which is incorporated a chemically inhibitive pigment.

The familiar olive-green finish applied to the metal lining of telephone booths is also a synthetic finish. This coating is often subjected to unusual service conditions which only a modern type of finish can withstand. Advantage is also taken of the high initial reflectivity and the retention of light reflection of certain alkyd resins and these are used in the white booth head-lining enamel. Synthetic finishes are generally specified for the finishing of Bell System trucks, etc.

Other applications of finishes include lacquers and wrinkled enamels. A recent interesting development has been the use of ethyl cellulose dipping lacquers to form a continuous, fairly thick, envelope around small telephone parts such as resistances, condensers and the like.

This frequently eliminates a potting operation and provides an excellent mechanical protection for the parts.

The employment of organic coatings for insulation purposes is another important application deserving mention. In general the conditions in the telephone plant are not such as to demand resistance to very high voltages. Millions of feet of copper wire receive a baked clear enamel coating applied in multiple thin coats to assure maximum flexibility and uniformity. Carefully chosen pigmented alkyd baking enamels are used in exchanges as insulating finishes for various hooks, bars and other small parts of the metal framework upon which the exchange wiring is tightly and compactly fastened and from which electrical insulation is needed.

An appreciable amount of clear cellulose acetate lacquer is at present used on switchboard wire. This is applied as a thin coating of a specially plasticized lacquer over a layer of textile insulation, the latter being colored in various ways for ready identification. The requirements of a good lacquer coating material for switchboard wire are chiefly (1) low cost, (2) reasonably good insulation, even under prolonged high humidity conditions such as occur during the summer months in many parts of the country and (3) good transparency so as not to alter the identification colors on the textile serving. Smoothness, flexibility, inflammability and corrosion hazard are other important factors that receive consideration.

A synthetic plastic which has recently found a small but important place in the telephone plant is polybutene. When coated on fabric this plastic has given an excellent membrane material for a new type of handset transmitter. It is dust and moisture-proof, light in weight, flexible and alkali resistant, not impairing in any way the acoustical properties of the instrument.

Synthetic Resins as Adhesives

A growing use for synthetic resins in the telephone plant is in the form of adhesives. The amount of material consumed in this way is not large but the applications are frequently important from the standpoint of the functioning of the apparatus as well as from the economies involved. The older kinds of adhesives such as casein and animal glue are still employed for joining together various large parts (especially wood, as in cabinet work, etc.) but they are brittle and generally unsatisfactory in the assembly of small light parts (metal, phenol fiber, ceramic, etc.) such as go into special communication apparatus.

The trend in the design of most apparatus has been toward smaller lighter parts and at the same time toward more rapid assembly. Synthetic resin adhesives are aiding this trend by avoiding in various places the dependence upon bolts, screws and similar mechanical locking devices. Proprietary resin-cellulosic lacquer adhesives and vinyl and acrylate polymers are proving of value because they give strong tough joints that are affected but little by moisture and are not apt to give trouble from corrosion or growth of mildew. The use of these materials for assembling parts in a thermoplastic manner looks particularly encouraging. When the surfaces which are to be joined are carefully cleaned, then primed with an air-dried coat of a suitable thermoplastic resinous adhesive and finally molded together under heat and pressure, tensile strengths of several tons per square inch are possible between the joined parts. Synthetic resin cements and adhesives are employed in the construction of the handset transmitter, moving coil microphones, loud speakers, switchboard lamps, vacuum tubes and the wood veneer of telephone booths.

CONCLUSION

Many important applications of plastics have been made in the telephone field. These have sprung from the economies of design, methods of fabrication, as well as from the excellent serviceability of the molded plastic products. It might be well to emphasize again the chief limitations of present day plastics which have prevented wider use. When exposed to outdoor conditions which involve the effect of temperature and sunlight, many plastics, particularly the newer thermoplastic materials, are revealed to be insufficiently permanent for telephone use with respect to the physical and chemical characteristics associated with color, distortion at elevated temperatures, surface deterioration due to action of sunlight and brittleness at moderately low temperatures.

Modern trends in stylized designs make it necessary to take advantage of the molding art in order to achieve good ornamentation. This will probably result in the use of plastics on surfaces exposed to light. The effect of light is not confined to direct sunlight for it has been found that daylight filtered by ordinary window glass for long periods will cause reduction in surface electrical resistance of plastics. In fact, the effect of light seems to be over a rather broad range of the spectrum, becoming intensified as the ultra-violet range is approached.

A few of the new thermoplastics are quite inflammable, a characteristic that will be a serious handicap to their extended use in exchanges and other locations where a fire might disrupt the service of a whole community.

There are still certain weaknesses of organic finishes with respect to impact abrasion, perspiration and moisture penetration, although there have been great advances made in this field in recent years.

The Dielectric Properties of Insulating Materials, III Alternating and Direct Current Conductivity

By E. J. MURPHY and S. O. MORGAN

This paper deals with the variation of a-c conductivity with frequency and with that of apparent d-c conductivity with charging time for dielectrics exhibiting anomalous dispersion (i.e., having dielectric constants which decrease with increasing frequency). The a-c conductivity of a dielectric exhibiting simple anomalous dispersion approaches a constant limiting value γ_∞ as the frequency increases. The discussion shows that γ_∞ possesses properties similar to those of the conductivity due to free ions, although in most cases it depends upon the motions of polar molecules or bound ions. It is also shown that the apparent conductivity for constant (d-c) potential approaches an initial value as the charging time is diminished. This initial conductivity γ_0 is demonstrated to be equal to the limiting value of the a-c conductivity attained at high frequencies (γ_∞), a relationship which simplifies the description of the behavior of dielectrics exhibiting simple anomalous dispersion. Dielectrics possessing the property of anomalous dispersion then have *two* conductivities: one is due to local motions of polar molecules or bound ions; the other is due to the migration of free ions to the electrodes.

Both γ_0 and γ_∞ refer to methods of measurement. It is to be noted that in many non-homogeneous dielectrics, especially those in which one part is of much higher resistivity than the remainder, both γ_0 and γ_∞ may be a measure of a free ion conductivity. As the equality of γ_0 and γ_∞ is independent of the nature of the polarization responsible for them, experimental agreement between a-c and d-c measurements cannot be used to distinguish whether the dielectric loss in a material is due to polar molecules, to bound ions, or to free ions present in a non-homogeneous dielectric. However, in homogeneous dielectrics γ_0 (or γ_∞) is a conductivity due to polar molecules or bound ions.

INTRODUCTION

THE preceding paper¹ dealt with the dielectric constant, showing mainly how it varies with the frequency of the applied alternating voltage for those dielectrics which behave in the simplest manner, and indicating the general character of the structural features responsible for this behavior. The discussion is extended here to the conductivity,

¹ Murphy and Morgan, *B. S. T. J.*, 17, 640 (1938).

which is not less important than the dielectric constant as a property of an insulating material. Though general aspects of the conductivity will be described for the sake of completeness, we wish mainly to show that materials which possess the property of anomalous dispersion may be considered to have *two* quite definite conductivities: one of these is the ordinary d-c conductivity due to free ions or electrons; the other is a special value of the a-c conductivity which will be discussed in this paper. We believe that the recognition of the existence in many materials of two conductivities instead of one is of considerable advantage, particularly in interpreting the behavior encountered in direct-current conductivity measurements on insulating materials, a subject upon which there has existed a considerable divergence of opinion.

The measurement of the direct-current conductivity of an insulating material is usually complicated by the fact that the current which flows when a constant potential is applied does not remain constant but decreases with time. The meaning of this variation of the current is open to more than one interpretation. Some investigators consider that its final value, approached asymptotically, and perhaps not closely approximated until a constant potential has been applied for an hour or more, is the proper basis for the calculation of the true conductivity of the material. Other investigators, notably Joffé, consider that the current/time curve should be extrapolated toward the instant of applying the voltage in order to obtain the proper value of the current to use in calculating the true conductivity. On this account the terms *initial conductivity*, *final conductivity* and *true conductivity* frequently appear in papers on the conductivity of insulating materials. While it has been usual to take either the initial or the final conductivity as the true conductivity, rejecting the other, it is shown here that with certain exceptions both conductivities are true conductivities in the sense that they are independent properties of the material having a different, though related, physical significance.

The relationships which will be brought out here depend in an essential way on the nature of the variation of a-c conductivity with frequency for materials which possess the property of anomalous dispersion. The a-c conductivity of a dielectric exhibiting simple anomalous dispersion increases as the frequency increases until the frequency is high as compared with the reciprocal of the relaxation-time; it then approaches asymptotically a constant limiting value. It is shown here that this limiting value of the conductivity, which will be referred to as the *infinite-frequency conductivity*, is a true conductivity of the material, analogous to the ordinary d-c conductivity, and that it is

equal to the initial conductivity obtained by extrapolating the apparent d-c conductivity towards the instant of applying the measuring voltage. We believe that this relationship considerably simplifies the description of the meaning of certain types of measurements upon dielectrics.

In spite of the fact that several terms are already used to distinguish different conductivities, there remains some ambiguity in the meaning of these terms. For example, the physical meaning of the term d-c conductivity when applied to a dielectric is vague. Moreover, it will be evident in the later discussion that the initial conductivity will depend upon free ions for some materials and upon polar molecules for other materials. To avoid this confusion we have found it convenient to use two terms which refer to the nature of the conduction processes rather than to the method of measurement: these are *free ion conductivity* and *polarization conductivity*. The first is the ordinary conductivity due to the drift of free electrons or ions to the electrodes; the second is a conductivity determined by the energy dissipated as heat by the polarization currents in the dielectric. The latter bears the same relation to the neutral polarizable aggregates in the material, which carry the polarization currents, as does the free ion conductivity to the free ions in the dielectric. The terms free ion conductivity and polarization conductivity, or some other terms having approximately the same meaning, are essential to the discussion as they refer unambiguously to two distinct properties of the material, while the terms initial, final, true, infinite-frequency, a-c and d-c conductivity all refer to different methods of measuring these two properties of the material.

The current flowing in a dielectric to which a constant potential is applied often decreases with time for periods of the order of a few minutes or longer measured from the time of applying the potential. This decreasing current is variously referred to as a *residual charging current*, an *absorption current*, an *anomalous conduction current* or an *irreversible absorption current*, depending upon the interpretation given to the phenomenon. We have already indicated that these residual currents complicate the measuring technique in the determination of the d-c conductivity of insulating materials. The most definite kinds of residual currents are those which are simply a manifestation of the structural characteristics which give rise to anomalous dispersion of the dielectric constant. These residual currents and the residual charges associated with them will be referred to here as the *direct-current counterparts of anomalous dispersion* to indicate that they are not independent properties of the material, but necessary requirements of the existence of anomalous dispersion occurring at sufficiently low

frequencies. The information obtainable from the study of such residual currents is the same in kind as that obtainable from the study of dielectric constant and conductivity by means of alternating currents; the residual phenomena, however, provide data regarding polarizations having relaxation-times which are too long for convenient investigation by alternating current methods. Residual currents of this kind have no significance in principle which is different from that of low-frequency a-c measurements.

CONDUCTIVITY AND DIELECTRIC LOSS

The conductivity of a material is usually thought of as a property which depends upon the ease with which electric charge can be transferred through the material by the application of an electric field, though it is recognized that a dissipation of electrical energy as heat occurs in the material through which the current is passing. In these terms we think of the conductivity as a quantity proportional to the current per unit voltage gradient, which in turn is proportional to the number of charge carriers, their mobility, and the magnitude of the charge borne by each carrier. For conductors it does not matter whether we define the conductivity, γ , as the factor by which the voltage gradient, E , must be multiplied to give the current density, I ,

$$I = \gamma E \quad (1)$$

or as the factor by which the square of the voltage gradient must be multiplied to give the heat, W , developed per second in a unit cube of the material,²

$$W = IE = \gamma E^2, \quad (2)$$

for the heat developed by a given voltage is proportional to the current, no matter of what material the conductor is composed. This is due to the fact that the energy obtained by the moving charges from the applied electric field is dissipated continuously to the surrounding molecules or lattice structure as heat, and the electrons or ions then drift with constant average velocity in the direction of the applied field, developing heat at a rate proportional to the current.

However, the proportionality between current and heat developed which is characteristic of conductors does not obtain in dielectrics. When an alternating current flows in a dielectric it dissipates some electrical energy as heat; however, the amount is generally much smaller than would be dissipated by an equal current flowing in a

² Cf. for example, Mason and Weaver, "The Electromagnetic Field," Chicago (1929), p. 233.

conductor and, unlike conductors, the ratio of heat developed to current flowing varies with the material. This is due to the fact that most of the current flowing in a dielectric under ordinary conditions is a polarization current, or rather a sum of several polarization currents of different types, and in general a polarization current dissipates less energy as heat than an equal current flowing in a conductor. In fact a part of the current flowing in a dielectric—the optical polarization current—passes through the dielectric material without developing any heat in it at the ordinary frequencies of electrical transmission. Electrical energy can be transmitted through a good dielectric in a suitable range of frequencies with very little loss; in other words, the dielectric is transparent to currents which have a suitable frequency of alternation. In these circumstances the conductivity of the material as measured on a bridge would be very small though the current density per unit voltage gradient might be quite large. Evidently, then, the view of conductivity as simply a measure of the ease of transfer of electric charge through a material is not in general suitable for application to dielectrics.

The fact is that the *complex* conductivity represents the ease of displacement of electric charge in a dielectric while its real part (i.e., the a-c conductivity as measured on a bridge or equivalent measuring device) is the quantity to which the rate of heat development in the material is proportional. Therefore, in dealing with alternating currents flowing in dielectrics it is usually more convenient to regard the a-c conductivity as the factor which determines the rate of dissipation of electrical energy as heat in the material, rather than as a quantity which is proportional to the current density per unit voltage gradient or to the ease of displacement of electric charge in the material. In a later part of the discussion, however, it will be shown that the limiting high-frequency value of the a-c conductivity may be thought of as representing ease of displacement of electric charge, too, as in a conductor.

The heat developed in a dielectric by polarization currents is called *dielectric loss* and is analogous to the Joule heat developed by free electrons or ions in a conductor; however, it is a property of neutral aggregates of particles, such as polar molecules, rather than of free ions. In the case of a polarization due to polar molecules, for example, the equilibrium distribution of the orientations of the molecules is slightly changed by the application of an electric field. The dielectric constant depends upon the difference between the distribution of orientations with and without the applied field, while the dielectric loss represents the part of the energy of the applied field which is dissipated as heat

because of the "friction" (i.e., the molecular equivalent of macroscopic friction) which the molecules experience as they change from the one equilibrium distribution of orientations to the other. Evidently, the dielectric loss may be quite as characteristic of the structure of the material as is the dielectric constant.

In an ideal insulating material there would be no free ion conduction, but in actual materials there are some free ions or electrons and these produce Joule heat as they drift towards the electrodes in the applied field. The total heat developed is the sum of the dielectric loss and the Joule heat; and, as the latter is proportional to the d-c or free ion conductivity, the dielectric loss is proportional to the total a-c conductivity (as measured on a bridge for example) less the d-c conductivity.

To give the discussion a more concrete basis, let us consider a dielectric which has a dielectric constant ϵ' and a loss-factor ϵ'' (or in other words which has a complex dielectric constant $\epsilon' - i\epsilon''$). Let it be contained in a parallel-plate condenser having a plate separation of d centimeters, and area A cm² for one surface of one of the plates. If a potential difference V is maintained between the plates of this condenser, a charge q per unit area will appear on either plate and a polarization P will be created in the dielectric. The current flowing in the leads to this condenser is $A dq/dt$, if we assume for the present that the conductivity due to free ions may be neglected. The conductivity is then given by

$$\gamma = \frac{1}{E} \frac{dq}{dt}, \quad (3)$$

where $E = V/d$. The charge q can be calculated from the dielectric constant of the material by means of relations which are provided by the general theory of electricity, namely

$$\epsilon E = D, \quad (4)$$

$$D = E + 4\pi P, \quad (5)$$

$$D = 4\pi q \text{ (for a parallel-plate condenser)}. \quad (6)$$

So (3) becomes

$$\gamma E = \frac{dq}{dt} = \frac{1}{4\pi} \frac{dD}{dt} = \frac{\epsilon}{4\pi d} \frac{dV}{dt}, \quad (7)$$

where all of the electrical quantities are expressed in electrostatic units. When the applied potential is alternating, V may be expressed as the real part of $V = V_0 e^{i\omega t}$, where V_0 is the amplitude. The dielectric constant may then be written as the complex quantity $\epsilon' - i\epsilon''$,

as shown in the preceding paper. The current density in the dielectric is then

$$\frac{dq}{dt} = i\omega(\epsilon' - i\epsilon'') \frac{V_0}{4\pi d} e^{i\omega t} \quad (8)$$

$$= \left(\frac{\epsilon''\omega}{4\pi} + i \frac{\epsilon'\omega}{4\pi} \right) E_0 e^{i\omega t} \quad (8a)$$

$$= (\gamma' + i\gamma'') E_0 e^{i\omega t}, \quad (8b)$$

where $\gamma' \equiv \epsilon''\omega/4\pi$ and $\gamma'' \equiv \epsilon'\omega/4\pi$. It is evident that $\gamma' + i\gamma'' (= \gamma)$ is the complex conductivity.

The dielectric constant and conductivity for alternating currents are determined by measurements, made with bridges or by other means, which give the admittance or impedance of the condenser containing the dielectric at the particular frequency at which the measurement is made. This admittance³ may be expressed in terms of the equivalent parallel capacitance (C_p) and conductance (G_p) and an alternative expression for (8) is then

$$\frac{dq}{dt} = \frac{0.9 \times 10^{12}}{A} (G_p + iC_p\omega) V_0 e^{i\omega t}, \quad (9)$$

where G_p is expressed in mhos (or reciprocal ohms) and C_p in farads and 0.9×10^{12} is the ratio of the farad to the electrostatic unit of capacitance and also of the mho to the e.s.u. of conductance.

By comparing (9) with (8), (8a) and (8b) we obtain expressions for γ' , ϵ'' and ϵ' in terms of the quantities C_p and G_p as directly measured on a bridge or similar arrangement. However, the expressions obtained are briefer if we make use of the fact that, when expressed in farads, the capacity C_0 of the empty condenser is

$$C_0 = \frac{A}{4\pi d \times 0.9 \times 10^{12}}. \quad (10)$$

Then it is evident that

$$\epsilon' = C_p/C_0, \quad (11)$$

$$\epsilon'' = G_p/C_0\omega, \quad (12)$$

$$\gamma' = G_p/4\pi C_0 \quad (13)$$

$$= \epsilon''\omega/4\pi = \epsilon''f/2. \quad (13a)$$

³ Measurements on a series bridge give directly the equivalent series resistance R_s and capacitance C_s . These data can be converted into equivalent parallel conductance and capacitance by the general relationships

$$C_p = \frac{C_s}{1 + (\omega R_s C_s)^2}; \quad G_p = \frac{\omega^2 R_s C_s^2}{1 + (\omega R_s C_s)^2}.$$

In equations (11) to (13a) ϵ' , ϵ'' and γ' are expressed in e.s.u., while C_p and C_0 are expressed in farads and G_p in mhos. The substitution of the frequency, f , for ω in (13a) depends upon the fact that $f = 2\pi\omega$.

While it is usual to express ϵ' and ϵ'' in e.s.u., it is more convenient for most purposes to have γ' in the units ordinarily used for specific conductance: thus when expressed in $\text{ohm}^{-1}\cdot\text{cm}^{-1}$

$$\gamma' = \frac{\epsilon''\omega}{4\pi \times 0.9 \times 10^{12}} = \frac{\epsilon''f}{1.8 \times 10^{12}}, \quad (14)$$

$$= \frac{8.85 \times 10^{-2}}{C_0 \text{ mmf}} G_p = \frac{d}{A} G_p, \quad (14a)$$

where $C_0 \text{ mmf}$ is the capacitance in micromicrofarads.

By expressing equation (8) in the equivalent polar form certain quantities appear which are closely related to γ' , ϵ' and ϵ'' and which are commonly used in describing the characteristics of dielectrics. The polar form is

$$\gamma = \gamma_0 e^{i\theta},$$

where $\gamma_0 = (\gamma'^2 + \gamma''^2)^{1/2}$, a quantity which is a measure of the amplitude of the complex current in the dielectric for unit voltage gradient, while $\theta = \tan^{-1} \gamma''/\gamma'$ is its phase angle. It is customary to use the *loss angle* which is defined as $\left(\frac{\pi}{2} - \theta\right) \equiv \delta$, rather than the phase angle in the description of dielectric properties. It is evident that $\delta = \tan^{-1} \gamma'/\gamma'' = \tan^{-1} \epsilon''/\epsilon'$ and that

$$\tan \delta = G_p/C_p\omega. \quad (15)$$

Similarly, the power factor is given by

$$\begin{aligned} \cos \theta &= \gamma'/(\gamma'^2 + \gamma''^2)^{1/2} \\ &= \epsilon''/(\epsilon'^2 + \epsilon''^2)^{1/2} = G_p/(G_p^2 + C_p^2\omega^2)^{1/2}. \end{aligned} \quad (16)$$

When the current given by (8) is multiplied by the voltage, $E_0 \cos \omega t$, we obtain the instantaneous power, and from this the mean power \bar{W} can be obtained by integration over a whole number of half periods. We then obtain

$$\bar{W} \text{ per second} = \gamma' \left(\frac{E_0}{\sqrt{2}}\right)^2 = \frac{\epsilon''\omega}{4\pi} \left(\frac{E_0}{\sqrt{2}}\right)^2 \quad (16a)$$

and

$$\bar{W} \text{ per cycle} = \frac{\epsilon''}{2} \left(\frac{E_0}{\sqrt{2}}\right)^2. \quad (16b)$$

This demonstrates the statements made earlier that γ' is proportional to the heat developed per second and ϵ'' to that developed per cycle in the dielectric. In the above equations \bar{W} is in ergs per second or per cycle when E_0 , γ' , and ϵ'' are in e.s.u.

It can be seen from equation (8) that the total current flowing in the dielectric has a dissipative and a non-dissipative part: ϵ' is proportional to the non-dissipative part, and ϵ'' to the dissipative part. The loss-angle, ϵ''/ϵ' , may be interpreted as the ratio of the dissipative to the non-dissipative current and the power factor as the ratio of the dissipative current to the total current.

THE FREQUENCY-DEPENDENCE OF CONDUCTIVITY

When the dielectric with which we are dealing possesses the property of anomalous dispersion, the expression for the loss factor ϵ'' as a function of frequency is

$$\epsilon'' = \frac{(\epsilon_0 - \epsilon_\infty)\omega\tau}{1 + \omega^2\tau^2}, \quad (17)$$

as was shown in the preceding paper. Substituting this expression for ϵ'' in (14) we obtain:

$$\gamma' = \frac{\epsilon''\omega}{4\pi} = \frac{1}{4\pi} \cdot \frac{(\epsilon_0 - \epsilon_\infty)\omega^2\tau}{1 + \omega^2\tau^2} \quad (18)$$

$$= \frac{1}{4\pi \times 0.9 \times 10^{12}} \cdot \frac{(\epsilon_0 - \epsilon_\infty)\omega^2\tau}{1 + \omega^2\tau^2}, \quad (18a)$$

where γ' is expressed in e.s.u. in (18) and in $\text{ohm}^{-1}\cdot\text{cm}^{-1}$ in (18a), and ϵ_0 is the static dielectric constant, ϵ_∞ the infinite-frequency dielectric constant and τ is the relaxation-time.

Differentiation of (18) with respect to frequency shows that γ' has no maximum when plotted against frequency; *the conductivity of any dielectric to which (18) applies should always increase with frequency, where it changes at all.* On the other hand, differentiation of (17) with respect to ω shows that the dielectric loss-factor has a maximum which occurs when $\omega\tau = 1$. The dielectric constant ϵ' is given by

$$\epsilon' = \epsilon_\infty + \frac{\epsilon_0 - \epsilon_\infty}{1 + \omega^2\tau^2} \quad (19)$$

and it will be seen that it shares with the conductivity the property of having no maximum when plotted against frequency. In Fig. 1 schematic curves are drawn which show the differences in the frequency dependence of γ' , ϵ'' and ϵ' for a material having an absorptive polariza-

tion of relaxation-time τ . The conductivity goes up as the dielectric constant goes down, as if the one were being transformed into the other.

The most interesting feature of (18) is that as the frequency increases γ' approaches a limiting value, and that this limiting value,

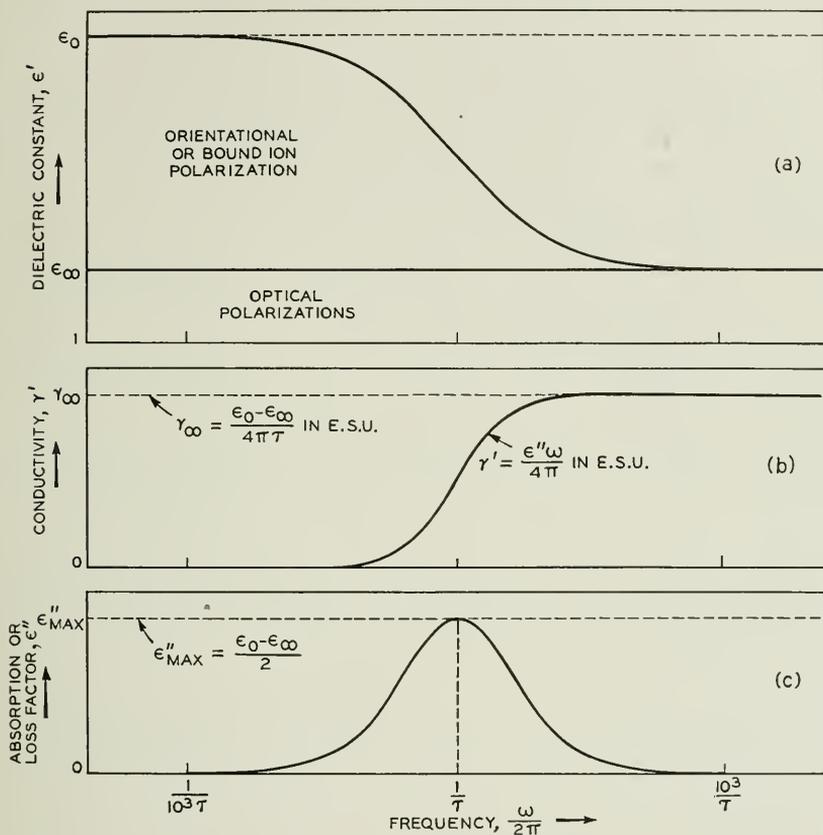


Fig. 1—Schematic diagram comparing the frequency dependence of dielectric constant (ϵ'), loss factor (ϵ'') and conductivity (γ'). This applies to a polarization having a single relaxation-time (τ). The frequency in cycles per second is $\omega/2\pi$.

which will be designated as γ_∞ , has the value

$$\gamma_\infty = \frac{\epsilon_0 - \epsilon_\infty}{4\pi\tau} \quad (20)$$

$$= \frac{\epsilon_0 - \epsilon_\infty}{4\pi \times 0.9 \times 10^{12}\tau}, \quad (20a)$$

where (20) gives γ_∞ in e.s.u. and (20a) in $\text{ohm}^{-1}\cdot\text{cm}^{-1}$. The con-

ductivity of ice provides a good example of this behavior, and Fig. 2 has been plotted to illustrate it, using data obtained by the writers on the conductivity of ice at different temperatures. The leveling-off of the conductivity curve at high frequencies is not observed in all materials. For many materials the conductivity continues to increase as the frequency becomes higher, though the rate of increase is often

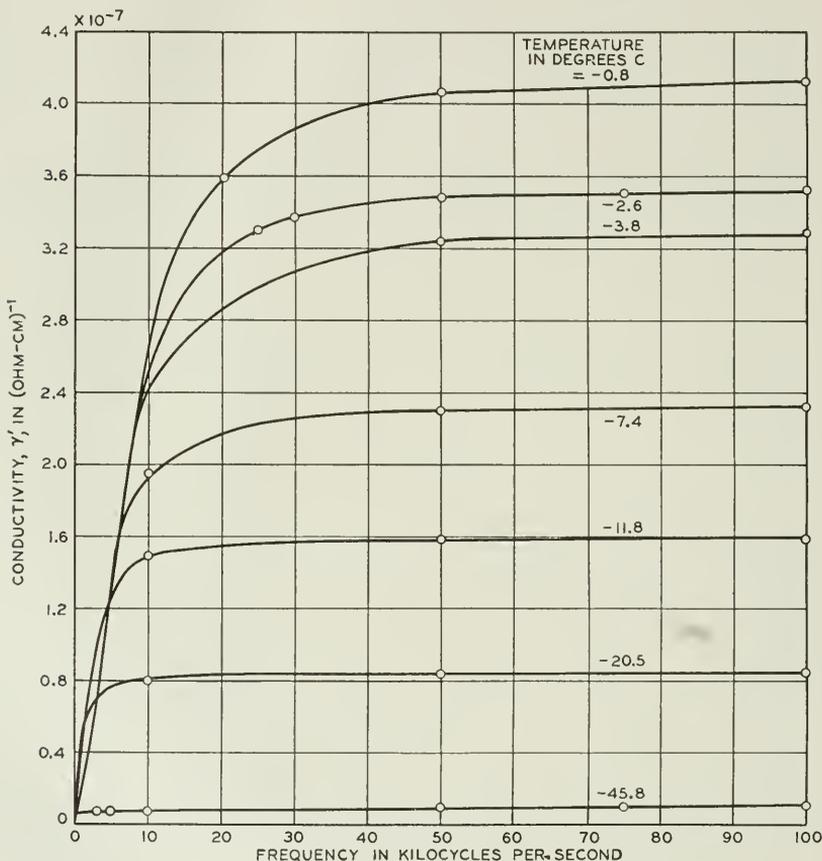


Fig. 2—The dependence of γ' upon frequency for ice at several different temperatures.* The quantity γ' is the a-c conductivity less the d-c conductivity. The curves show that the limiting value γ_{∞} approached by γ' as the frequency increases is lower the lower the temperature.

* The curves of Figs. 2, 4 and 5 are drawn through experimentally determined points. The experimental points are shown only where the curves should theoretically be linear or nearly linear. The fact that the points deviate by only comparatively small amounts from the theoretical curves in their linear sections is a sufficient indication for the present purpose of the agreement between the theoretical curves and the experimental data for ice.

not great. It is to be expected that many complex materials would not have a conductivity which varies with frequency in accurate agreement with (18), because the assumptions from which (18) was derived are perhaps the simplest which could be made regarding the structure of a dielectric.

DISCUSSION OF γ_{∞} IN TERMS OF A MODEL

A convenient way of demonstrating the physical meaning of γ_{∞} and at the same time of showing the general character of the physical mechanism which is responsible for γ' becoming larger as the frequency increases is to consider the operation of the model⁴ used in the preceding paper to develop the formulae for dispersion. This model depends upon the dielectric containing bound ions about which the only things specified were the following:

(1) That the displacement of a bound ion from its equilibrium position in the dielectric by the applied electric field is opposed by a restoring force proportional to the displacement; if the displacement is designated as s the restoring force is fs .

(2) That these ions experience in their motion a frictional force proportional to their velocity in the direction of the applied field; this frictional force is given by $r\dot{s}$, where \dot{s} is the velocity and r is a constant.

(3) That the moving ions have a charge e and a negligible mass. These are the essential features of the model and they can be represented concretely by imagining an ion held in a certain small region, electrically neutral as a whole, in for example a glassy dielectric, by forces characteristic of the structure of the solid. We may suppose that this ion makes small excursions within this region around the point (O in Fig. 3) toward which it is attracted by a force proportional to the displacement, and that its interaction with the molecules which surround it in the dielectric is such that it experiences in effect a frictional force proportional to its velocity in the applied field. The ion of this model is, therefore, subjected to the same sort of frictional resistance as are ions in solution in a liquid.

From the relationship between polarization and displacement which was discussed in the preceding paper, it is evident that the polarization may be considered to be proportional to the displacement of the bound ion of this model. (If P is the polarization per unit volume due to a large number (n) of bound ions, each of charge e , the polarization is evidently proportional to the average displacement \bar{s} per ion, where $\bar{s} = P/ne$, but for the present purpose it will be sufficient to consider the displacement s of a single bound ion.) We may, therefore, discuss

⁴ For further details, see page 652 of the preceding paper, *B. S. T. J.*, 17 (1938).

the dependence of conductivity on frequency in terms of the displacement and velocity of a single bound ion. As we are not concerned here with very high frequencies, we may employ the abbreviated equation of motion given in equation (17) of the preceding paper to

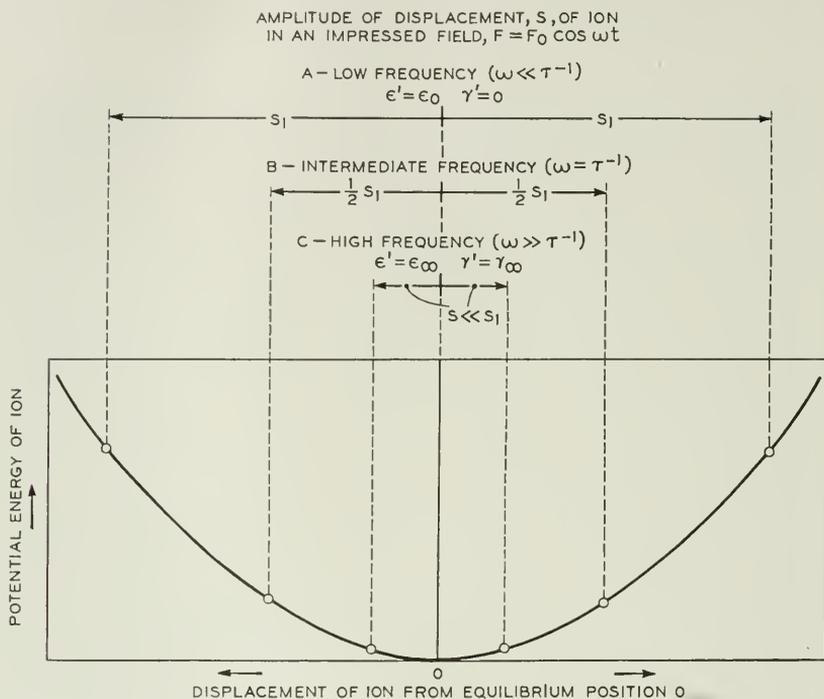


Fig. 3—The mechanism of anomalous dispersion illustrated by a simple model.

The model consists of a single bound ion. The potential energy of this ion increases when it is displaced from its equilibrium position 0. The ion also experiences a frictional force proportional to its velocity, as if it were an ion in solution. The upper part of the diagram shows the way in which the amplitude (s) of displacement of the ion by a given applied field varies with the frequency. It has its maximum amplitude at low frequencies (A in the diagram) and a comparatively negligible amplitude at high frequencies (C in the diagram). In this model the amplitude is a measure of the dielectric constant. The limiting value γ_{∞} of the conductivity prevails under the conditions C where the amplitude is comparatively negligible.

discuss the motion of the bound ion of this model; this equation is

$$rv + fs = eF,$$

where $v = ds/dt$.

When a d-c voltage V_1 is applied to this model, it establishes a field F_1 which displaces the bound ion to a new equilibrium position s_1 if the field is allowed to act upon the ion for a sufficient time. The new

equilibrium position s_1 of the ion corresponds to the static value P_1 of the polarization of the model. When the voltage is varying with the time according to $V = V_0 \cos \omega t$, the greatest amplitude which the displacement can have is s_1 , and in general the amplitude will fall short of this value by an amount which increases with increasing frequency. The value s_1 is then closely approached only when the frequency is low as compared with the reciprocal of the relaxation-time, because at high frequencies the applied field reverses its direction before the ion has had time to reach s_1 . At sufficiently low frequencies, namely where ω is negligible by comparison with $1/\tau$, the frictional dissipation of energy by the moving ion is so small that there is practically no difference between the instantaneous position of the ion when the voltage has any given value and the position it would finally attain upon reaching equilibrium for that voltage. The ion then moves through a succession of near-equilibrium positions, as in a reversible process in thermodynamics. The dielectric constant has its static value and the conductivity is zero unless there is a d-c conduction component in the total conductivity.

At the high-frequency extremity of a dispersion region we see that the situation is simply reversed: the alternations in the direction of the applied field are so rapid that the bound ion does not have time to move an appreciable distance from its equilibrium position before the direction of the applied field is reversed (C, Fig. 3). The amplitude of the displacement of the ion by the applied field is then small as compared with s_1 and the dielectric constant of the material receives practically no contribution from the bound ion of this model in these circumstances. However, though the amplitude of motion of the ion shrinks to a small fraction of s_1 , its velocity is comparatively high and independent of frequency. The conductivity γ_∞ is proportional to the average velocity of the bound ion of the model under these conditions.

As the restoring force is proportional to the displacement, its effect upon the motion of the ion is negligible by comparison with that of the applied force when the displacement s is small as compared with s_1 . On this basis, the fact that the conductivity is an increasing function of frequency may be attributed to the decrease in the influence of the restoring forces as the frequency increases. In fact, when the amplitude of displacement is very small as compared with s_1 , the ion moves as if the only force opposing the applied force were the frictional force; that is, its average velocity is the same as that of a free ion subjected to the same applied field and the same friction.

For many of the purposes of this discussion we could use a model of the dielectric consisting of an air capacity C_s in series with a resistance

R_s both being shunted by a second air condenser C_∞ . In this equivalent circuit, C_s and R_s refer to the polarizations responsible for anomalous dispersion, and C_∞ to the optical polarizations. The frequency-dependence of the equivalent parallel capacitance and conductance of this network is

$$C_p = C_\infty + \frac{C_0 - C_\infty}{1 + \omega^2 T^2} \quad (21)$$

and

$$G_p = \frac{(C_0 - C_\infty)\omega^2 T}{1 + \omega^2 T^2}, \quad (22)$$

where $C_0 \equiv C_s + C_\infty$ and $T \equiv C_s R_s$. In the above expressions $(C_0 - C_\infty)$ and T are analogous respectively to $\epsilon_0 - \epsilon_\infty$ and τ in equations (17) and (19).

The physical basis for the infinite frequency conductivity in this model depends upon the fact that at high frequencies the impedance of C_s is so low that nearly the whole drop in voltage is over the resistance R_s . This simple network is capable of representing the frequency-dependence of materials exhibiting anomalous dispersion due to a polarization having a single relaxation-time. In fact, when the frequency is sufficiently high that it is in the range where the conductivity is independent of frequency, the required network becomes even more simple, for it then reduces to C_∞ shunted by R_s , where the magnitude of C_∞ corresponds to ϵ_∞ and that of R_s to $1/\gamma_\infty$.

POLARIZATION CONDUCTIVITY

The operation of the models which have been discussed above provides a basis for interpreting the physical nature of γ_∞ . The essential characteristics brought out by these models are listed below. They show the justification for considering γ_∞ to be a conductivity in the same sense as the ordinary d-c conductivity.

(1) To obtain γ' in an actual measurement on a dielectric, we subtract the d-c conductivity γ_f from the total a-c conductivity. There is then no contribution from free ion conduction in γ' and consequently none in γ_∞ , its limiting value at high frequencies. Polar molecules or other polarizable aggregates in the dielectric must then be the origin of γ_∞ .

(2) In the second place γ_∞ is independent of frequency, a property which puts it on the same footing as the d-c or free-ion conductivity in at least one respect.

(3) Earlier in this paper it was mentioned that the heat developed in a conductor for a given voltage is proportional to the *total* current,

but that in dielectrics this proportionality does not in general prevail. The current in a dielectric is complex and heat is developed only by its dissipative component. If the expressions for ϵ' and ϵ'' given in (17) and (19) are substituted in (8) we see that when ω becomes large by comparison with $1/\tau$, i.e., when γ' becomes γ_∞ , the imaginary component of the current reduces to $\epsilon_\infty\omega/4\pi$; this is the optical polarization current. If it is subtracted from the total current given by (8), the remaining current contains no imaginary component. This current then develops as much heat in the dielectric as would a current of the same magnitude flowing in a conductor.

(4) In connection with the foregoing we see that unlike lower values of γ' the infinite-frequency conductivity γ_∞ is a measure of the ease with which electrical charge can be displaced in the material by a unit applied field. This characteristic of γ_∞ agrees with our usual conception of the physical basis of the conductivity of an electrolyte or a metal. (We assume in this connection that the optical polarization current $\epsilon_\infty\omega/4\pi$ may be neglected in comparison with the current responsible for γ_∞ . Where this is not the case, appropriate modifications in the above statements are required.)

(5) It is characteristic of a dielectric that when the charged particles which form part of its structure are displaced by a force of external origin, there is a restoring force tending to return them to their initial positions. On the other hand, in an ideal conductor there are, by definition, no restoring forces of this kind. The above discussion of the model shows that in a dielectric possessing the property of anomalous dispersion it is possible to make the influence of the restoring forces on the motion of a bound ion negligible in comparison with that of the applied force by sufficiently increasing the frequency above the value corresponding to the reciprocal of the relaxation-time. This is the condition which prevails when γ' equals γ_∞ . Thus at low frequencies ($\omega \ll \tau^{-1}$) the part of the dielectric structure which is responsible for anomalous dispersion behaves as a dielectric; whereas at high frequencies ($\omega \gg \tau^{-1}$) it behaves as a conductor. A result of this is that a dielectric exhibiting anomalous dispersion of the simple kind conforming to equation (39) of the preceding paper will behave in an electric circuit like a pure capacity shunted by a pure resistance over the whole of that range of frequencies where ϵ' and γ' are both practically independent of frequency and equal respectively to ϵ_∞ and γ_∞ . Pure ice, for example, behaves in this manner over a considerable range of frequencies. (See Figs. 2 and 4.)

(6) The average velocity of the bound ion of our model becomes independent of frequency when ω is large as compared with $1/\tau$. This

constant velocity is equal to that which a free ion would have under the same voltage gradient if it were moving in a medium subjecting it to the same frictional resistance as is experienced by the bound ion of our model.

(7) In ordinary electrolytic conduction the conductivity is usually represented as the product of three factors: the number of ions per unit volume; their valence or charge per ion; and the average mobility of each ion, i.e., the average distance which an ion drifts per second in

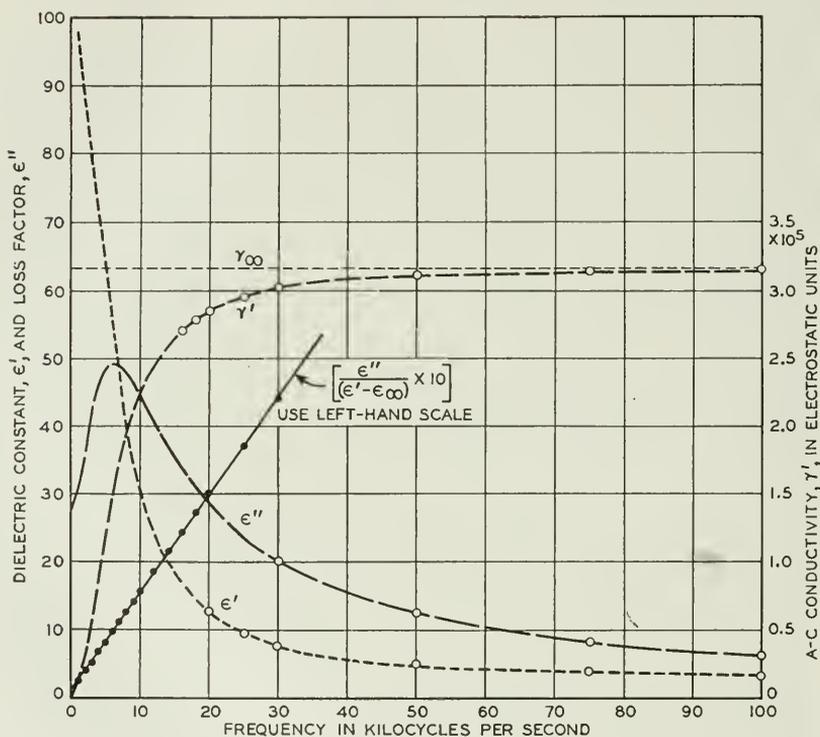


Fig. 4—Dependence of ϵ' , ϵ'' , γ' and $\epsilon''/(\epsilon' - \epsilon_{\infty})$ upon frequency for ice at -2.6°C .

the direction of the applied field. The above statement refers to one species of ion; when two or more species are present in the solution, the total current or conductivity is the sum of the currents or conductivities contributed by each type. The infinite-frequency conductivity γ_{∞} may also be represented as the product of three factors similar in physical meaning to those just mentioned as applying to conduction by ions in solution. Thus reference to Table I, Item 2, will show that

for the model we have employed to illustrate this discussion,

$$\gamma_{\infty} = \frac{1}{4\pi} \cdot \left(\frac{\epsilon_{\infty} + 2}{3} \right)^2 \cdot \frac{ne^2}{r}. \quad (23)$$

The factors in this expression may be seen to have the following general significance: the quantity e is the charge on each bound ion; n is the number of bound ions per unit volume; and $e/4\pi r$ is a measure of the mobility per ion. This mobility does not refer to the motion of an ion which is free to move through the dielectric from one electrode to the other but to the mobility of a bound ion in small, local translational motions or the rotational mobility of a polar molecule. The remaining factor $\left(\frac{\epsilon_{\infty} + 2}{3} \right)^2$ is not of direct significance in the present connection.⁵

We see then that γ_{∞} is also analogous to ordinary electrolytic conduction in that its physical mechanism may be represented as depending upon a mobility, a concentration and a factor such as dipole moment or charge per bound ion. The latter factor has a function in this mechanism which is similar to that of the valence or charge per ion in electrolytic conduction.

These considerations indicate that although γ_{∞} is a property of polarizable units such as polar molecules it has the usual attributes of a conductivity due to free ions or free electrons. A dielectric which exhibits simple anomalous dispersion conforming to equations (17) and (19) then has *two* conductivities. One of these is the conductivity due to free ions; this will be called the *free ion conductivity* and designated throughout this paper by γ_f . The other is a conductivity which is a characteristic of the polarizable complexes responsible for anomalous

⁵ When $\epsilon_{\infty} = 1$, equation (23) reduces to

$$\gamma_{\infty} = \frac{ne^2}{4\pi r},$$

showing thereby that the factor $\left(\frac{\epsilon_{\infty} + 2}{3} \right)^2$ would be absent if the material possessed no optical polarizations. Evidently γ_{∞} depends upon the optical refractive index $\sqrt{\epsilon_{\infty}}$, as well as upon the characteristics of the absorptive polarization. In mixtures it may be possible to vary these two factors independently. Since optical polarization currents make no *direct* contribution to the energy dissipation in the dielectric, even up to the highest radio frequencies, it is interesting to observe that they make an *indirect* contribution according to equation (23). Their indirect action takes place by virtue of their effect on the actual internal field which acts upon each polarizable aggregate in the dielectric. The effect of the interaction of the optical polarization with the absorptive polarizations is to increase the apparent mobility of the polarizable complexes responsible for anomalous dispersion.

lous dispersion; this will be called the *polarization conductivity*⁶ and designated by γ_{pol} in this paper.

The magnitude of the polarization conductivity of a material is proportional to the number of polarizable units of structure such as polar molecules or bound ions per unit volume which contribute to anomalous dispersion. It also depends upon the mobility which these polarizable units have in the local translational or rotational motions in which they engage in consequence of thermal agitation. It finally depends upon the permanent dipole moment of the polar molecules or upon the charges upon the bound ions.

The concentration of ions able to contribute to conduction in dielectrics is generally low because in many cases the free ion conductivity depends mainly upon a small percentage of impurity in the material. On the other hand, the concentration of polarizable units which are able to contribute to the polarization conductivity may be much larger and in fact even equal to the total number of molecules per unit volume. Consequently, the polarization conductivity γ_{pol} may often be more reproducible in measurements upon different specimens of the same dielectric than is the free ion conductivity.⁷ It may well be that in many materials diffusion coefficients, thermal conductivity, mechanical dissipation and other similar properties which might be expected on theoretical grounds to be related to electrical conductivity will bear a simpler or more easily demonstrated relationship to polarization conductivity than to free ion conductivity.

An example of the advantage of using the infinite-frequency conductivity instead of the d-c conductivity appears in measurements of the conductivity of ice. In Fig. 5 the infinite-frequency conductivity (or polarization conductivity) of ice is plotted against the reciprocal of the absolute temperature, using unpublished data of the writers. The data for γ_{∞} are reproducible and the curve shows a relation similar to that usually observed for the d-c conductivity of solids. Direct-current measurements on the same specimens on the other hand yielded very erratic results. It may be seen from Fig. 5 that the polarization conductivity is much higher than the d-c conductivity.

⁶ We suggest for this conductivity the name polarization conductivity because it is a property of polarizable units of structure. In cases where the polarization is due to the change of orientation of polar molecules, we might instead refer to it as an *orientational conductivity* or a *polar molecule conductivity*, contrasting it thereby with the translational aspect of ordinary conduction by free ions. As ions which are loosely bound to some stationary or moving unit of the dielectric structure are often capable of producing anomalous dispersion, at least two types of polarization conductivity are possible; these may be described as the orientational conductivity and the bound ion conductivity.

⁷ Joffé has obtained evidence that the initial conductivity, which we show here to be in some cases a polarization conductivity, is often superior in reproducibility to the final conductivity. (Cf. Reference 15.)

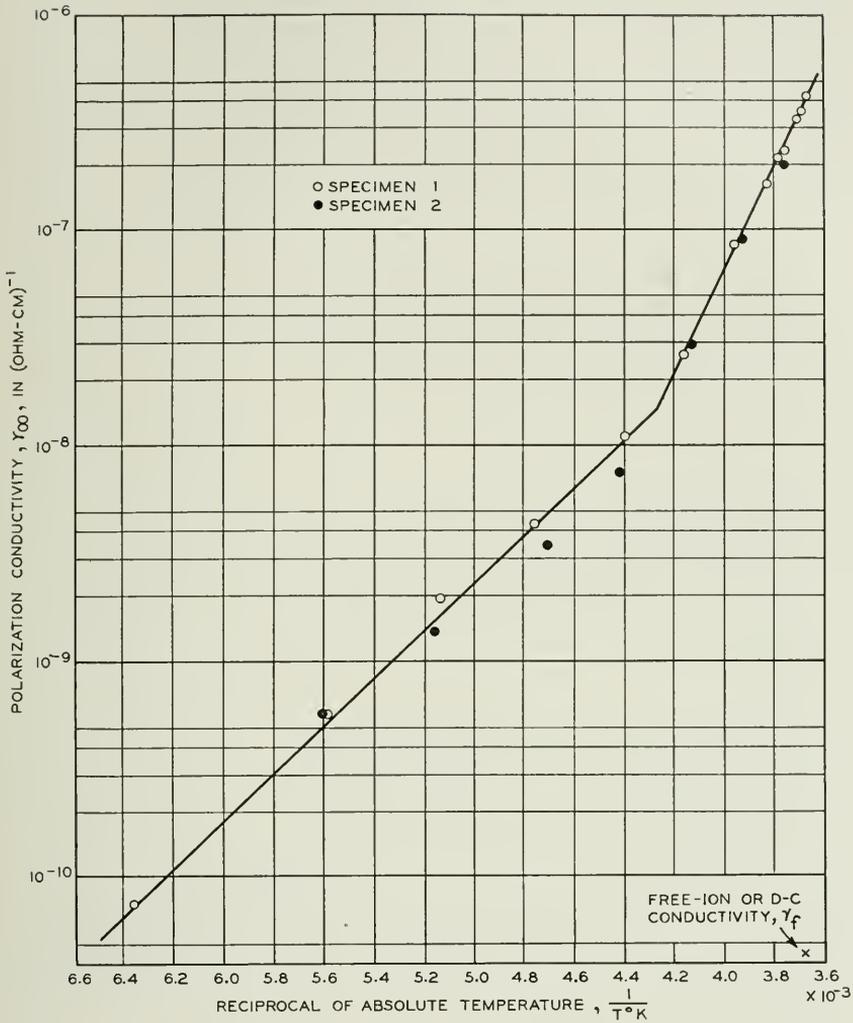


Fig. 5—An illustration of polarization conductivity. The temperature-dependence of the polarization conductivity of pure ice in the range -0.8°C to -190°C . The free-ion (or d-c) conductivity is also shown for a single temperature.

TYPES OF INFINITE-FREQUENCY CONDUCTIVITY

In Table I of the preceding paper there were listed several types of polarization capable of yielding anomalous dispersion curves distinguishable from each other only by the values of the constants. Corresponding to each is a different expression for the polarization conductivity and these expressions are listed in Table I of this paper.

This list shows that there are at least three main types of infinite-frequency conductivity. The first of these is the type which depends upon the change of the orientation of polar molecules according to the Debye theory. This type of polarization conductivity is of more theoretical interest than any of the others and perhaps also of more practical importance; as already mentioned, it may be described as an orientational conductivity to emphasize that no translational mobility is necessary for it to occur.

TABLE I
EXPRESSIONS FOR THE CONSTANT VALUE γ_∞ APPROACHED BY THE A-C
CONDUCTIVITY AS THE FREQUENCY INCREASES

Type of Polarization	
1. The Orientational Polarization due to Polar Molecules.....	$\gamma_\infty = \frac{L\mu^2}{12\pi\eta a^3} \cdot \left(\frac{\epsilon_\infty + 2}{3}\right)^2$
2. A Distortional Polarization having a Relaxation-Time given by $\tau = r/f$	$\gamma_\infty = \frac{ne^2}{4\pi r} \cdot \left(\frac{\epsilon_\infty + 2}{3}\right)^2$
3. Polarizations due to Spatial Variations of Conductivity and Dielectric Constant	
(a) A two-layer dielectric, layers 1 and 2 having respectively static dielectric constants ϵ_1 and ϵ_2 and free-ion conductivities γ_1 and γ_2	$\gamma_\infty = \frac{(\epsilon_1\gamma_2 - \epsilon_2\gamma_1)^2}{(\epsilon_1 + \epsilon_2)^2(\gamma_1 + \gamma_2)}$
(b) Special case of (3a) where γ_1 is much larger than γ_2	$\gamma_\infty = \left(\frac{\epsilon_2}{\epsilon_1 + \epsilon_2}\right)^2 \gamma_1$
(c) Special case of (3b) where $\epsilon_1 \cong \epsilon_2$	$\gamma_\infty = \gamma_1/4$
(d) Special case of (3a) consisting of a high resistance transition layer at the dielectric/electrode interface, where γ_1 is the conductivity of the dielectric.....	$\gamma_\infty = \gamma_1$
(e) Conducting spheres dispersed in an insulating medium of the same dielectric constant.....	$\gamma_\infty = p\gamma_1$

Note: The infinite-frequency conductivity γ_∞ is given here in e.s.u. (See equation 20.) In Table I of the preceding paper (*B. S. T. J.*, 17, 640 (1938)) the values of $(\epsilon_0 - \epsilon_\infty)$ and τ are given for the polarizations listed above; the expressions given there for τ should be divided by 4π in the case of Items 3a, c, d and e, as should also the expressions for $\epsilon_0 - \epsilon_\infty$ in the case of Item 2. The quantities which appear in the above Table are defined in an appendix to the preceding paper. γ_1 and γ_2 are expressed in e.s.u. in this table.

The remaining members of the list of Table I originate in the properties of ions rather than those of molecules. These ions must be more or less bound in order to have an infinite-frequency conductivity differing from the zero-frequency conductivity.⁸ The nature and

⁸ It will be recalled that the terms infinite-frequency and zero-frequency are not used here in their general meaning but merely as a convenient way to indicate opposite directions of extrapolation of dispersion curves. They refer respectively to the high-frequency extremity of a dispersion curve where γ' becomes practically independent of frequency and to the low-frequency extremity where the dielectric constant becomes practically independent of frequency.

strength of the binding forces may vary widely amongst the types of polarization which have an infinite-frequency conductivity. An ion will be regarded as bound if its potential energy increases when it is displaced from an equilibrium position by an applied field.⁹

Included among the infinite-frequency conductivities which depend upon the presence of ions in the dielectric is a type for which macroscopic inhomogeneities in the dielectric are responsible; for example, a two-layer or multiple layer laminated dielectric, or a dielectric in which space-charges¹⁰ form because of spatial variations in its resistivity or because of a transition layer of high resistance at the contact between dielectric and electrode. Examples of the infinite-frequency conductivity due to this type of mechanism are given in Items 3*a*, *b*, *c* and *d* of Table I.

This type of infinite-frequency conductivity is of little interest in principle, but in practice there may be many instances in which the measurement of the infinite-frequency conductivity provides a convenient means of determining the conductivities of the constituents of these non-homogeneous systems. For example, when one layer of a two-layer dielectric has a much higher conductivity than the other, γ_{∞} assumes a value which is related simply to the free ion conductivity of the layer which has the higher conductivity (see Item 3*b*, Table I). If the dielectric constants of the two layers are equal, γ_{∞} is equal to one-quarter of the conductivity of the high-conductivity layer. The conductivities γ_1 and γ_2 of the two layers are considered in the present connection to be free ion conductivities. A more complicated situation is possible where γ_1 and γ_2 are in part polarization conductivities due to polar molecules or bound ions.

The special case of a space-charge caused by a thin layer of high resistance at one or both of the electrodes is of interest in connection with the methods recommended by Joffé for the measurement of the true conductivity of crystals. This will be discussed in more detail later but for the present it may be noted that the infinite-frequency

⁹ It is necessary to confine the application of the last statement to direct voltages or to frequencies lower than those for which γ' is equal to γ_{∞} . When the frequency is high enough for the latter condition to prevail the amplitude of displacement of the ion becomes so small that the applied field produces no appreciable increase in the average potential energy of the ion; this is illustrated in Fig. 3 at *C*.

¹⁰ The external effects of a space-charge occurring in a dielectric because of spatial variations in its resistivity may be reproduced by a uniformly distributed polarization of suitably adjusted magnitude and relaxation-time. Several different polarizations of different magnitudes and relaxation-times would in some instances be required. In referring to such a space-charge as a polarization we may think of the term as applying to the uniform distribution of polarization which could replace the space-charge in its external effects.

conductivity for this system has the simple value

$$\gamma_{\infty} = 1/4\pi C_0 R = \gamma_1$$

as shown in Item 3*d*, Table I. Here γ_1 is the free ion conductivity of the main part of the dielectric, R is its resistance and $4\pi C_0 (= A/d)$ is the ratio of thickness to length of specimen. (In Table I of the preceding paper the symbol C_{∞} is used in place of C_0 .) *The infinite-frequency conductivity in a non-homogeneous system of this type is a free-ion conductivity.*

Bound ions may also be distributed with macroscopic uniformity in a dielectric. An example of this type of bound ion conductivity¹¹ is one due to conducting particles dispersed uniformly in a relatively non-conducting medium. This is the case referred to in Item 3*e* of Table I. Macroscopic uniformity is obtained in this case by the random distribution of a large number of particles. However, there are some general experimental indications that the distribution of bound ions may in some materials depend upon the basic internal structure of the dielectric and involve some regular geometrical configuration repeated throughout the material.

In certain dielectrics which absorb an appreciable amount of water when in a humid atmosphere, conduction takes place in aqueous conduction paths permeating the solid. Examples of these materials are cotton, paper, silk and wool. This property is probably shared by many other polymeric substances. The water in these materials is distributed in minute capillaries, the dimensions and other characteristics of which probably determine the form and distribution of the conduction paths.¹² There exists in these materials a condition capable of producing a bound ion conductivity inasmuch as there are indications that the conducting paths are not of uniform cross-sectional area. Evidence for the existence of a bound ion conductivity in the kind of material to which we have just referred is provided, for example, by conductivity measurements on cotton.¹³ Raw cotton contains salts which can be removed by extraction leaving the material otherwise practically unchanged. These salts are likely to be distributed in the material with macroscopic uniformity as they form part of its natural structure. The fact that the removal of these salts decreases the

¹¹ As already mentioned we shall call any infinite-frequency conductivity which is caused by a macroscopically uniform distribution of bound ions a *bound ion conductivity*. In some places this term will be applied to any conductivity due to bound ions irrespective of whether or not that conductivity is the limiting high-frequency value.

¹² One of the basic structural units of cellulose and other similar materials is the *micelle*. This usually contains a large number of molecules and the capillaries we refer to may correspond to the intermicellar spaces.

¹³ Murphy, *Journal of Physical Chemistry*, 33, 200 (1929).

dielectric loss indicates that the material possessed a bound ion conductivity before the salts were removed.

Some of the materials belonging to the class of dielectrics which we have just discussed are closely related in chemical and in physical structure to compounds which are important biologically and in the study of plant and seed structure. Many are also of commercial importance as insulating materials.

It is not necessary that the conduction paths be composed of aqueous solutions: in some materials plasticizers or products of pyrolysis are sufficiently conducting for this purpose. The dielectric behavior of certain plastics may be interpreted as evidence for the existence of such non-aqueous conduction paths in the material, producing a free ion conductivity, a bound ion conductivity and a contribution to the dielectric constant. Imperfections of structure occurring in crystals are able to produce a bound ion conductivity and there is experimental evidence that these imperfections do occur.¹⁴ The regular lattice ions in an ionic crystal have too high a binding energy, and dissipate too little energy in their motions in a radio frequency electric field to produce a bound ion conductivity.

The polarization which is responsible for the bound ion conductivity is of the interfacial, or Maxwell-Wagner, type. This type of polarization may be of importance in materials with a cellular structure and in materials which may be described as interstitially conducting dielectrics.

In the above discussion we have outlined the character of three widely different types of infinite-frequency conductivity:

(a) An orientational conductivity depending upon the small changes which an applied field produces in the average orientation of polar molecules.

(b) A bound ion conductivity depending upon the displacement of uniformly distributed bound ions.

(c) An infinite-frequency conductivity which is proportional to the free ion conductivity of one of the constituents of a dielectric consisting of two or more layers of widely different conductivities.

THE RELAXATION-TIME

The relaxation-time is closely related to the infinite-frequency conductivity. This may be seen by reference to equation (20), which shows that the relaxation-time is given by

$$\tau = \frac{\epsilon_0 - \epsilon_\infty}{4\pi\gamma_\infty}. \quad (24)$$

¹⁴ See, for example, A. Smekal, *Zeits. f. techn. Physik*, 8, 561 (1927).

This equation shows that specifying the values of τ and $(\epsilon_0 - \epsilon_\infty)$ gives as much information as specifying γ_∞ and $(\epsilon_0 - \epsilon_\infty)$. In some applications there are advantages in using τ but in other applications greater simplicity of description is gained by using γ_∞ .

There are several convenient ways of calculating the relaxation-time. The more familiar ones depend upon the position of maxima which occur in certain dielectric properties when they are plotted against the frequency: there are maxima in the loss factor vs. frequency curve, in the tangent of the loss angle vs. frequency curve and in the power factor vs. frequency curve. As these maxima occur at different frequencies, the corresponding expressions for the relaxation-time are also different. They are listed in Table II. It will be ob-

TABLE II

LIST OF FORMULAE FOR CALCULATING THE RELAXATION-TIME (τ)

1. The frequency at which the maximum in loss factor (ϵ'' or $\epsilon' \tan \delta$) occurs is $\omega_{\max(1)}$ $\tau = 1/\omega_{\max(1)}$
2. The frequency at which the maximum in loss angle (ϵ''/ϵ' or $\tan \delta$) occurs is $\omega_{\max(2)}$ $\tau = \sqrt{\frac{\epsilon_\infty}{\epsilon_0}} \frac{1}{\omega_{\max(2)}}$
3. The frequency at which the maximum in power factor $\epsilon''/(\epsilon'^2 + \epsilon''^2)^{1/2}$ occurs is $\omega_{\max(3)}$ $\tau = \sqrt{2} \sqrt{\frac{\epsilon_\infty}{\epsilon_0}} \frac{1}{\omega_{\max(3)}}$
4. The quantity $\epsilon''/(\epsilon' - \epsilon_\infty)$ is a linear function of ω $\tau = \frac{d}{d\omega} (\epsilon''/(\epsilon' - \epsilon_\infty))$
5. The relaxation-time is proportional to the ratio of the absorptive part $(\epsilon_0 - \epsilon_\infty)$ of the static dielectric constant to the infinite-frequency conductivity (γ_∞)..... $\tau = (\epsilon_0 - \epsilon_\infty)/4\pi\gamma_\infty$

Note: An example of the application of these formulae is provided by the curves of Fig. 4. The value of τ for ice at -2.6°C is 25.8 microseconds as calculated from the position of the maximum in ϵ'' , 24.6 microseconds as calculated from $(\epsilon_0 - \epsilon_\infty)/4\pi\gamma_\infty$, where γ_∞ is in e.s.u., and 23.1 microseconds as calculated from the slope of $\epsilon''/(\epsilon' - \epsilon_\infty)$.

served that the simplest of these formulae for the relaxation-time is the one involving the maximum in the loss factor.

The function $\epsilon''/(\epsilon' - \epsilon_\infty)$ is a linear function of ω with slope equal to τ . This property provides an alternative method of calculating the relaxation-time. An example of its application to an actual material is provided by the data for ice plotted in Fig. 4.

It is interesting that $\epsilon''/(\epsilon' - \epsilon_\infty)$ has no maximum while ϵ''/ϵ' has a maximum. The physical basis of this is that in subtracting ϵ_∞ from ϵ' we remove the contribution to ϵ' made by optical polarizations. What is left represents only the dielectric constant due to the polarization responsible for anomalous dispersion. Consequently the tangent of the loss angle of the polarization current responsible for anomalous

dispersion has no maximum when plotted against the frequency and its behavior is, therefore, in contrast with the tangent of the loss angle of the *total* polarization current, i.e., the sum of the optical polarization current and the absorptive polarization current.

From the above discussion it will also be evident that the physical basis for the maximum in $\tan \delta$ is different from that of the maximum in ϵ'' . As we have just shown, the maximum in ϵ''/ϵ' ($= \tan \delta$) depends upon the inclusion of optical polarizations in ϵ' . On the other hand, there would be a maximum in ϵ'' even if the part of the dielectric constant which is due to optical polarizations (ϵ_∞) were neglected or considered to be zero. This will be evident by differentiation of equation (17).

The maximum in ϵ'' is an intrinsic property of the absorptive polarization. The general nature of the mechanism by which it occurs is as follows: ϵ'' is proportional to γ'/ω ; as the frequency increases γ'/ω at first increases, but when γ' reaches the constant value γ_∞ , further increase in frequency causes γ'/ω to decrease.

The quantity τ which we have discussed here is a property of the dielectric as a whole as we indicated in the preceding paper. This quantity is connected with the relaxation-time τ' of the individual polarizable units by the relation

$$\tau = \frac{\epsilon_0 - 2}{\epsilon_\infty + 2} \tau', \quad (25)$$

when the material is of cubic or isotropic structure. This relationship is a consequence of the fact that the actual force acting upon a particle within a dielectric depends not only upon the applied field of external origin but also upon a force exerted by the polarization induced in the dielectric.

THE RELATIONSHIP BETWEEN DIELECTRIC CONSTANT AND DIELECTRIC LOSS

If ϵ'_{\max} is the value of the dielectric constant at the frequency where the loss factor ϵ'' is at a maximum when plotted against frequency, we have

$$\epsilon'_{\max} = \epsilon_\infty + \frac{\epsilon_0 - \epsilon_\infty}{1 + \omega_{\max}^2 \tau^2} = \frac{\epsilon_0 + \epsilon_\infty}{2}, \quad (26)$$

$$\epsilon''_{\max} = \frac{(\epsilon_0 - \epsilon_\infty) \omega_{\max} \tau}{1 + \omega_{\max}^2 \tau^2} = \frac{\epsilon_0 - \epsilon_\infty}{2}. \quad (27)$$

By addition and subtraction of (26) and (27) we obtain the following

relationships:

$$\epsilon_0 = \epsilon'_{\max} + \epsilon''_{\max}, \quad (28)$$

$$\epsilon_\infty = \epsilon'_{\max} - \epsilon''_{\max}, \quad (29)$$

$$\epsilon_0 - \epsilon_\infty = 2\epsilon''_{\max}. \quad (30)$$

The comparison of the last equation with equation (20) brings out an interesting contrast between the maximum dielectric loss per cycle (ϵ''_{\max}) and the maximum dielectric loss per second (γ_∞). *The maximum dielectric loss per cycle is completely determined by the difference between the static dielectric constant and the optical dielectric constant.* On the other hand, the dielectric loss per second depends as well upon the relaxation-time. The relaxation-time usually varies rapidly with temperature, whereas $(\epsilon_0 - \epsilon_\infty)$ changes comparatively slowly with temperature. The temperature-dependence of the maximum dielectric loss per cycle is related to the polarizability of the material, whereas the temperature variation of the maximum dielectric loss per second is primarily a measure of the change of internal friction with temperature.

In the foregoing we have discussed the conductivity, dielectric loss and relaxation-time of dielectrics which have simple properties with respect to the frequency-dependence of these quantities. However, for many dielectrics, particularly solids, the experimental data are not in agreement with the dispersion formulae for a single relaxation-time which has been discussed here. The explanation usually adopted for this discrepancy is that the polarizations induced in the dielectric possess a distribution of relaxation-times.

THE D-C COUNTERPARTS OF ANOMALOUS DISPERSION

A dielectric so constructed that it exhibits anomalous dispersion under an alternating voltage should show some equally characteristic behavior when a direct voltage is substituted for the alternating one. These characteristics may be described as the d-c counterparts of anomalous dispersion. They include certain definite types of variation of current with time under constant applied potential.

In the appendix the d-c counterparts of anomalous dispersion are derived by employing the model used throughout this paper. This enables us to demonstrate an especially simple relationship between the a-c and d-c conductivity.

Equation (16) of the appendix gives the apparent conductivity as a function of charging time. As it has been assumed that $\epsilon_\infty C_0 R \ll \tau$, the first term of equation (16) will quickly become negligible. Then for charging times, measured from the instant of applying the voltage,

such that $\tau \gg t_c \gg \epsilon_\infty C_0 R$, the apparent conductivity $\gamma_c(t_c)$ has the value

$$\gamma_0 = \frac{\epsilon_0 - \epsilon_\infty}{4\pi\tau} = \gamma_\infty. \quad (31)$$

The special value of $\gamma_c(t_c)$ which we have designated as γ_0 in (31) will be called the *initial conductivity*.

Equation (31) states that the infinite-frequency conductivity (γ_∞), obtained by a-c measurements, is equal to the initial d-c conductivity (γ_0) which would be obtained by extrapolating current-time curves toward the instant of applying the voltage. This relationship, which has not been demonstrated previously to our knowledge, is of interest in connection with the interpretation of conductivity measurements.

The model described in Appendix I and indicated schematically in Fig. 3 illustrates the physical nature of the initial conductivity in a simple manner. When a constant voltage V_1 is applied to a dielectric having the properties of this model an effective impressed field F_1 is established in the dielectric. This displaces the bound ion assumed to be responsible for the polarization in this model. The magnitude of the displacement s of this bound ion depends upon the length of time that F_1 is applied. If it is applied for a time which is much longer than the relaxation-time, s approaches a constant value s_1 corresponding to complete polarization of the dielectric. On the other hand, during that stage of the charging process when the charging time t_c is negligible in comparison with the relaxation time τ and at the same time large as compared with the time constant $\epsilon_\infty C_0 R$, the displacement s is negligible in comparison with s_1 . The resultant force tending to displace the ion is then approximately equal to the impressed force; $eF_1 - fs_1 \cong eF_1$.

Near the beginning of the charging process there is a brief interval of time ($\epsilon_\infty C_0 R \ll t_c \ll \tau$), when the motion of the bound ion of our model in the applied field is essentially the same as that of a free ion. During this interval the prevailing conductivity is the initial conductivity defined by equation (31). Although the bound ion of the model is actually subjected to a force tending to restore it to its initial position, this restoring force has not had time during the initial stage of the charging process to build up to a magnitude appreciable in comparison with the applied force. The initial conductivity corresponds to a condition in the dielectric where bound ions act for a brief time as if they were free as far as conduction processes are concerned. The infinite-frequency conductivity corresponds also to this condition. The variation of apparent conductivity with the time of charging is

indicated schematically in Fig. 6. As is there shown, there must be an initial stage, always too brief to be detected experimentally, during which the inductance of the circuit and the inertia of the charges cannot be neglected.

THE TYPES OF INITIAL CONDUCTIVITY

The polarization conductivity γ_{pol} may then be measured in two ways: either as the infinite-frequency conductivity γ_{∞} obtained by a-c measurements, or as the initial conductivity γ_0 obtained by d-c measurements:

$$\gamma_{\text{pol}} = \gamma_{\infty} = \gamma_0.$$

The quantity γ_{pol} refers to a property of the material, whereas γ_0 and γ_{∞} refer to the methods of measuring this property.

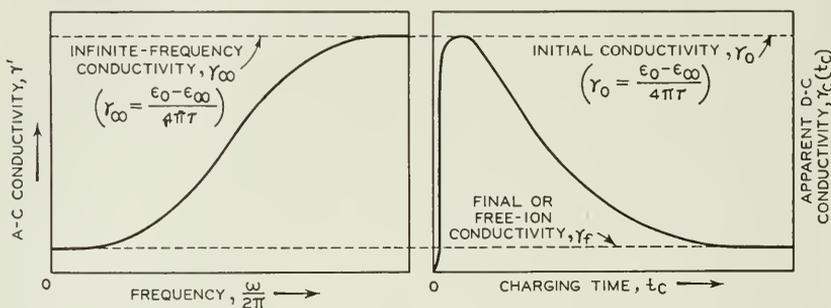


Fig. 6—A-c and d-c methods of measuring conductivity. A schematic diagram comparing the dependence of a-c conductivity (γ') on frequency with the dependence of apparent d-c conductivity $\gamma_c(t_c)$ on charging time (t_c). For homogeneous dielectrics $\gamma_0 = \gamma_{\infty} = \gamma_{\text{pol}}$, where γ_{pol} is a conductivity due to polar molecules or to bound ions. For some non-homogeneous dielectrics $\gamma_0 = \gamma_{\infty} = \gamma_f$, where γ_f is a conductivity due to free ions.

Reference to Table I will show that there are several types of polarization which may be responsible for γ_{∞} . Consequently this will be true also of the initial conductivity. The equality of γ_0 and γ_{∞} applies to all of these types of polarization and, therefore, can not be used to distinguish between them. For this reason experimental agreement between a-c and d-c measurements of conductivity does not enable us to distinguish whether we are dealing with polar molecules, bound ions, or free ions in a macroscopically inhomogeneous dielectric. Thus it is clear that when we are dealing with a polarization due to polar molecules, the *initial conductivity is a true polarization conductivity or a specific dielectric loss*. If the polarization is of the bound ion type discussed earlier in this paper, the initial conductivity

is also a polarization conductivity. However, if we are dealing with a macroscopically non-homogeneous dielectric such as a two-layer dielectric or a material which has a high resistance blocking layer at one of the electrodes γ_0 is a free ion conductivity. This is evident from the previous discussion of the significance of γ_∞ .

These relationships are of interest in connection with the difficulties encountered in the interpretation of d-c conductivity data which were mentioned in the introduction. They apply especially to the methods recommended by Joffé¹⁵ and by Richardson.¹⁶

When a constant potential difference is maintained between the plates of a condenser containing a solid dielectric the current observed does not in general remain constant but usually decreases with time. This decrease may continue for several minutes or hours. The study of these residual currents is of importance in connection with the interpretation of conductivity data on dielectrics. The question arises as to how much of the observed behavior of the residual currents which flow in dielectrics under constant potential may be explained as the d-c counterparts of anomalous dispersion. Many materials of practical importance as insulators exhibit a more complicated type of variation with frequency than is indicated by equations (17) and (19) which refer only to the simplest observed type of dispersion. The more complicated types of behavior observed are usually attributed to the presence of polarizations possessing a wide distribution of relaxation-times. However, other processes may also contribute to the deviation of the experimental curves from the theoretical. One of these is electrolysis which produces changes in the composition, and consequently in the conductivity of the material. Another effect which may contribute is a possible lack of constancy of the relaxation-time. It is evident, therefore, that the d-c counterparts of anomalous dispersion due to a polarization of a single relaxation-time should not be expected to explain all of the observed residual phenomena, particularly in solid dielectrics. However, the quantitative relations derived as the d-c counterparts of anomalous dispersion are applicable to some materials, and for those to which they are not quantitatively applicable, may serve as a useful guide in the interpretation of the residual currents. As another section of this paper is planned in which the influence of residual currents upon conductivity measurements will be discussed further, we have included in the appendix some relationships which are useful in the interpretation of the behavior of these currents.

¹⁵ A. Joffé, *Ann. d. Physik*, 72, 481 (1923); "Physics of Crystals," New York (1928); *Zeits. f. Physik*, 62, 730 (1930). See also Sinjelnikoff and Walther, *Zeits. f. Physik*, 40, 786 (1927).

¹⁶ S. W. Richardson, *Proc. Roy. Soc.*, 107A, 102 (1925).

APPENDIX

THE D-C COUNTERPARTS OF ANOMALOUS DISPERSION

Having developed the a-c characteristics of the model, the properties of which were specified in the preceding paper,¹⁷ we now turn to the direct-current characteristics of this model. This involves investigating the characteristics of the currents produced when a constant or direct voltage V_1 is applied to a condenser containing a dielectric having properties which correspond in all respects essential to the discussion to those of the model just described. Let the resistance of the leads to the condenser be R and let the source of the electromotive force V_1 have a negligible internal resistance. These conditions require the following five equations to be satisfied simultaneously:¹⁸

$$r \frac{dP}{dt} + fP - ne^2F = 0, \quad (1)$$

$$F = E + AP_t, \quad (2)$$

$$P_t = k_i F + P, \quad (3)$$

$$D = E + P_t, \quad (4)$$

$$E = E_1 - CR \frac{dD}{dt}. \quad (5)$$

(Rational e.s.u. are used in these equations for convenience but the final relations are converted into electrostatic units.¹⁹) P_t denotes the total polarization; it is the sum of the optical polarizations, given by $k_i F$, and the polarization P which forms comparatively slowly and causes anomalous dispersion. The total displacement D , is given by (4). C_0 is the air capacitance of the condenser. V is the potential drop over the condenser. The separation of the plates of the condenser is d . When the whole drop in potential is concentrated over the condenser the applied field strength E_1 is given by $E_1 = V_1/d$. Equation (5) is obtained by equating the total drop in potential over the leads and the condenser to the applied potential V_1 .

Equations (1) to (5) may be combined to give the following:

$$(1 + 2p_0)\tau'CR \frac{d^2D}{dt^2} + \{(1 - p_0)\tau' + CR(1 + 2p_0 + 2p_1)\} \frac{dD}{dt} + (1 - p_0 - p_1)D - (1 + 2p_0 + 2p_1)E_1 = 0. \quad (6)$$

¹⁷ These properties are also outlined on page 513 of this paper.

¹⁸ Cf. P. Debye, "Polar Molecules," pp. 86-88, where an analogous case is discussed.

¹⁹ For conversion factors see, for example, Mason and Weaver, Reference 2, page 370. The rational electrostatic unit of conductivity is smaller than the e.s.u.; the ratio is 4π . The dielectric constant is unaffected by changing from rational e.s.u. to e.s.u.

In this equation the following abbreviations are used: $\tau' = r/f$, $Ak_i = p_0$, $Ak = p_1$. The last two abbreviations are introduced to facilitate comparison with an analogous derivation given in "Polar Molecules," p. 86. $D = D_0 e^{\alpha t}$ is a solution of the homogeneous differential equation obtained by letting $E_1 = 0$, provided that the following equation is satisfied:

$$\alpha^2 + \left\{ \left(\frac{1 - p_0}{1 + 2p_0} \right) \frac{1}{CR} + \frac{(1 + 2p_0 + 2p_1)}{1 + 2p_0} \frac{1}{\tau'} \right\} \alpha + \left(\frac{1 - p_0 - p_1}{1 + 2p_1} \right) \frac{1}{\tau' CR} = 0. \quad (7)$$

As we are interested here only in the special case where $CR \ll \tau'$, the $1/\tau'$ term in the coefficient of α can be neglected in comparison with the $1/CR$ term, and the roots are

$$\alpha = - \left(\frac{1 - p_0 - p_1}{1 - p_0} \right) \frac{1}{\tau'} \text{ or } \alpha = - \left(\frac{1 - p_0}{1 + 2p_0} \right) \frac{1}{CR}$$

to a degree of approximation which improves the larger the ratio τ'/CR . The general solution of the non-homogeneous equation (6) is

$$D = D_1 e^{-\left(\frac{1-p_0}{1+2p_0}\right)\frac{t}{CR}} + D_2 e^{-\left(\frac{1-p_0-p_1}{1-p_0}\right)\frac{t}{\tau'}} + \left(\frac{1 + 2p_0 + 2p_1}{1 - p_0 - p_1} \right) E_1 = 0. \quad (8)$$

For the special case of a random or cubic distribution of molecules (in which case $A = 1/3$), Eq. (8) can be simplified by means of the following relationships:

$$(1 + 2p_0)/(1 - p_0) = \epsilon_\infty, \quad (9)$$

$$(1 + 2p_0 + 2p_1)/(1 - p_0 - p_1) = \epsilon_0, \quad (9')$$

$$(\epsilon_\infty + 2)/(\epsilon_0 + 2) = (1 - p_0 - p_1)/(1 - p_0), \quad (10)$$

$$\left(\frac{1 - p_0}{1 - p_0 - p_1} \right) \tau' = \left(\frac{\epsilon_0 + 2}{\epsilon_\infty + 2} \right) \tau' = \tau. \quad (10')$$

Equation (8) then becomes

$$D = D_1 e^{-t/\epsilon_\infty CR} + D_2 e^{-t/\tau} + \epsilon_0 E_1 \quad (11)$$

and introducing the initial conditions, $t = 0$, $D = 0$; $t = 0$, $\frac{dD}{dt} = E_1/CR$, we obtain,

$$D = \epsilon_0 E_1 - \epsilon_\infty E_1 e^{-t/\epsilon_\infty CR} - (\epsilon_0 - \epsilon_\infty) E_1 e^{-t/\tau}. \quad (12)$$

Let q_c be the charge per unit area on either of the condenser plates at any stage of the charging process, that is, at any time t_c after the instant at which the voltage was applied. On converting (12) into e.s.u., we have for the charge at any time during the charging process:

$$q_c = \frac{D_c}{4\pi} = \frac{\epsilon_0 E_1}{4\pi} - \frac{\epsilon_\infty}{4\pi} E_1 e^{-t_c/\epsilon_\infty C_0 R} - \frac{(\epsilon_0 - \epsilon_\infty)}{4\pi} E_1 e^{-t_c/\tau}. \quad (13)$$

Therefore the ballistic or d-c dielectric constant at any time of charging t_c , is

$$\epsilon(t_c) = \frac{D_c}{E_1} = \epsilon_0 - \epsilon_\infty e^{-t_c/\epsilon_\infty C_0 R} - (\epsilon_0 - \epsilon_\infty) e^{-t_c/\tau}. \quad (14)$$

The d-c dielectric "constant" appears in this equation as a function of the charging time. Its dependence on charging time is analogous to the dependence of the a-c dielectric "constant" on frequency. The static dielectric constant ϵ_0 is obtained when charging has been continued until $t_c \gg \tau$, that is, when t_c is infinite the dielectric constant has its static value ($t_c = \infty$, $\epsilon(t_c) = \epsilon_0$) and when t_c is zero the dielectric constant is zero ($t_c = 0$, $\epsilon(t_c) = 0$).

The charging *current* is obtained by differentiation of (13) and is

$$\dot{q}_c = \frac{D_c}{4\pi} = \frac{E_1}{4\pi C_0 R} e^{-t_c/\epsilon_\infty C_0 R} + \left(\frac{\epsilon_0 - \epsilon_\infty}{4\pi\tau} \right) E_1 e^{-t_c/\tau}. \quad (15)$$

The charging current per unit voltage gradient, or the apparent conductivity $\gamma_c(t_c)$, is

$$\gamma_c(t_c) = \frac{\dot{q}_c}{E_1} = \gamma_R e^{-t_c/\epsilon_\infty C_0 R} + \gamma_\infty e^{-t_c/\tau}, \quad (16)$$

where $\gamma_R \equiv (4\pi C_0 R)^{-1}$ and $\gamma_\infty = (\epsilon_0 - \epsilon_\infty)/4\pi\tau$. (It will be noticed that if we define a quantity $G_R \equiv 1/R$, it follows that $(4\pi C_0 R)^{-1} = G_R \cdot d/A = \gamma_R$ since $4\pi C_0 = A/d$. The quantity γ_R is the specific conductance which a fictitious material must possess if it were put in a condenser of geometric capacitance C_0 and required to conduct the same current as C_0 when in series with an external resistance R .)

At any stage of the charging process such that t_c is large as compared with $\epsilon_\infty C_0 R$, but at the same time small as compared with τ the apparent conductivity given by (16) reduces to

$$\gamma_c(t_c) = \gamma_\infty.$$

This value of $\gamma_c(t_c)$, where $\epsilon_\infty C_0 R \ll t_c \ll \tau$, will be designated by γ_0 and called the *initial conductivity*. Using this terminology we see that

the initial conductivity γ_0 as determined by *d-c* measurements equals the infinite-frequency conductivity γ_∞ as determined by *a-c* measurements.

If t_d represents the time measured from the instant that the voltage is abruptly reduced to zero, and q_d , D_d , \dot{q}_d represent respectively the charge, displacement, and discharge current at time t_d , we have for discharging after *complete* charging

$$q_c(t_d) = \frac{D_d}{4\pi} = \frac{\epsilon_\infty E_1}{4\pi} e^{-t_d/\epsilon_\infty C_0 R} + \frac{(\epsilon_0 - \epsilon_\infty)}{4\pi} E_1 e^{-t_d/\tau} \quad (17)$$

or

$$\epsilon(t_d) = \epsilon_\infty e^{-t_d/\epsilon_\infty C_0 R} + (\epsilon_0 - \epsilon_\infty) e^{-t_d/\tau}. \quad (17a)$$

At the end of the charging process (cf. (13)) or the beginning of the discharge process (cf. (17)) the charge per unit area per unit applied field strength is $\epsilon_0/4\pi$.

The discharge current for *complete* charge is obtained by differentiating (17) with respect to the time:

$$\dot{q}_d = \frac{\dot{D}_d}{4\pi} = -\frac{E_1}{4\pi C_0 R} e^{-t_d/\epsilon_\infty C_0 R} - \frac{(\epsilon_0 - \epsilon_\infty) E_1}{4\pi\tau} e^{-t_d/\tau} \quad (18)$$

or

$$\gamma_d(t_d) = \frac{\dot{D}_d}{4\pi E_1} = \gamma_R e^{-t_d/\epsilon_\infty C_0 R} + \gamma_\infty e^{-t_d/\tau}. \quad (18a)$$

Comparison of (18) and (15), or (18a) and (16), shows that for complete charging (that is, t_c effectively infinite or $t_c \gg \tau \gg \epsilon_0 C_0 R$) the charging current vs. time curve is identical, except for direction, with the discharge current vs. time curve. *This is true only of the curves for complete charging and is not true if the polarized condition in the dielectric is not fully formed and the polarization currents are not zero.*

The discharge-current time curve for *incomplete* polarization of the dielectric is not as simple as for complete polarization. When the charging process is broken off before completion, the initial conditions for the discharge are not the same as when charging is complete. The charge at time t_d during a discharge following a charging process which is broken off at t_c is

$$q_d = \frac{D_d}{4\pi} = \frac{\epsilon_\infty E_1}{4\pi} e^{-t_d/\epsilon_\infty C_0 R} - \frac{\epsilon_\infty E_1}{4\pi} e^{-(t_c+t_d)/\epsilon_\infty C_0 R} + \left(\frac{(\epsilon_0 - \epsilon_\infty)}{4\pi} E_1 e^{-t_d/\tau} - \frac{(\epsilon_0 - \epsilon_\infty)}{4\pi} E_1 e^{-(t_c+t_d)/\tau} \right). \quad (19)$$

This is an example of the *superposition principle for residual charges*.

The discharge current for incomplete charge is given by

$$- \dot{q}_d = \frac{-\dot{D}_d}{4\pi} = \frac{E_1}{4\pi C_0 R} e^{-t_d/\epsilon_\infty C_0 R} - \frac{E_1}{4\pi C_0 R} e^{-(t_c+t_d)/\epsilon_\infty C_0 R} + \left(\frac{\epsilon_0 - \epsilon_\infty}{4\pi\tau} \right) E_1 e^{-t_d/\tau} - \left(\frac{\epsilon_0 - \epsilon_\infty}{4\pi\tau} \right) E_1 e^{-(t_c+t_d)/\tau}. \quad (20)$$

Or if

$$\frac{-\dot{D}_d}{4\pi} = -\dot{q}_d \equiv I_d$$

$$I_d = \gamma_R E_1 e^{-t_d/\epsilon_\infty C_0 R} - \gamma_R E_1 e^{-(t_c+t_d)/\epsilon_\infty C_0 R} + \gamma_\infty E_1 e^{-t_d/\tau} - \gamma_\infty E_1 e^{-(t_c+t_d)/\tau}. \quad (20a)$$

Equation (20) or (20a) is an example of the superposition principle for the residual currents in a dielectric having a single absorptive polarization with a relaxation-time τ . It is evident that in the early stages of the charging process the electronic or instantaneous polarizations responsible for ϵ_∞ have the same external effect as an absorptive polarization because of the fact that the current by which they are formed must flow through the lead resistance R . Thus the condenser acts as though it contained a polarization yielding a dielectric constant with a relaxation-time $\epsilon_\infty C_0 R$.

As the time-constant $\epsilon_\infty C_0 R$ is generally small, the first two terms on the right of (20a) may usually be neglected and we have

$$I_d = \gamma_\infty E_1 e^{-t_d/\tau} - \gamma_\infty E_1 e^{-(t_c+t_d)/\tau}. \quad (21)$$

The first term on the right of (21) is the discharge current corresponding to the discharge of the condenser after the residual polarizations have been *fully* formed. The second term gives the value which the charging current would have if the charging process were continued for the interval of time $t_c + t_d$ instead of being discontinued after t_c seconds. When the charging time is large as compared with τ the second term may be neglected and the magnitude of the discharge current is the same function of the discharge time as is the charging current of the charging time; the charging curve and discharge curve can then be superimposed on one another if we disregard the direction of the current. There is only one complete charging current curve and *only one discharging current curve corresponding to complete polarization of the dielectric at any given applied potential. There is, however, an infinite number of discharge curves corresponding to incomplete polarization of the dielectric, that is, to any time of charging which is shorter than that necessary for complete polarization.*

The superposition principle states that any of these discharge curves may be derived from the discharge curve for complete polarization by subtracting from its ordinates the values which the charging current would have if it had continued during the discharge. From the method of deriving equation (21) it is clear that the superposition principle is a necessary consequence of an assumed exponential growth and relaxation of the residual polarizations, as required by the theory of simple anomalous dispersion. If these in fact do not vary exponentially with the time, whatever function they do follow appears in general to obey an empirical superposition rule. Reference to (20a) will indicate that if there are m polarizations of different relaxation-times which are quite far apart, each polarization will simply contribute two terms to the expression for I_d ; that is,

$$I_d = \sum_{j=1}^m (\gamma_{\infty j} E_1 e^{-t_d/\tau_j} - \gamma_{\infty j} e^{-(t_c+t_d)/\tau_j}) \\ + \gamma_R E_1 e^{-t_d/\epsilon_{\infty} C_0 R} - \gamma_R e^{-(t_c+t_d)/\epsilon_{\infty} C_0 R}. \quad (22)$$

Thus, the existence of the superposition principle for residual currents as an empirical law suggests that the individual polarizations actually vary exponentially with the time, though direct measurement of the total discharge current seldom gives a single exponential curve. This is, however, not the only possible interpretation.

Abstracts of Technical Articles from Bell System Sources

*Radio Telephone System for Harbor and Coastal Services.*¹ C. N. ANDERSON and H. M. PRUDEN. Radio telephone service with harbor and coastal vessels is now being given through coastal stations in the vicinities of seven large harbors on the Atlantic and Pacific coasts with additional stations planned. The system is designed to be as simple as possible from both the technical and operating standpoints on both ship and shore.

Recent developments in the shore-station design eliminates all manipulations of the controls by the technical operator. This is made possible principally because of crystal-controlled frequencies on shore and ship, a "vogad" which keeps the transmitting volume of the shore subscriber constant, and a "codan" incorporated in the shore radio receiver which will operate on signal carrier but is highly discriminatory against noise. A signaling system permits the traffic operator to call in an individual boat by dialing the assigned code which rings a bell on the particular boat called. The ship calls the shore station by turning on the transmitter. The radio signal operates the codan in the shore receiver which in turn lights a signal lamp in the traffic switchboard.

Gradually the system has been taking on more and more the aspects of the wire telephone system.

*Ship Equipment for Harbor and Coastal Radio Telephone Service.*² R. S. BAIR. The ultimate objective in the design of radio telephone apparatus for use on ships is to provide equipment which is as convenient and simple to operate as the telephone at home. To a considerable degree this has been accomplished in the new 15- and 50-watt ship sets that have recently been designed for use on harbor craft and coastwise vessels.

The requirements for sets of this type are discussed and the new equipment is described in this paper.

*Protective Coatings for Metals.*³ R. M. BURNS and A. E. SCHUH. This book is one of the American Chemical Society Series of Scientific and Technologic Monographs. The chapter headings are: *Protective*

¹ *Proc. I. R. E.*, April 1939.

² *Proc. I. R. E.*, April 1939.

³ Published by Reinhold Publishing Corporation, New York, N. Y., 1939.

Coatings and the Mechanism of Corrosion—Surface Preparation for the Application of Coatings—Types of Metallic Coatings and Methods of Application—Zinc Coating by Hot-Dipping Process—Zinc Coating by Electroplating and Cementation—Protective Value of Zinc Coatings—Cadmium Coatings and their Protective Value—Tin Coatings—Nickel and Chromium Coatings—Coatings of Copper, Lead, Aluminum and Miscellaneous Metals—Coatings of Noble and Rare Metals—Methods of Testing Metallic Coatings—Composition of Paints and Mechanism of Film Formation—The Durability and Evaluation of Paints—Paint Practices—Miscellaneous Coatings.

“The active interest manifested during the past few years in investigations on the general subject of the corrosion of metals has led to the carrying-out of long-time exposure tests which yielded much new basic information having a direct bearing on our knowledge of the useful life of coatings and coated metals. The authors have wisely incorporated a great deal of this information in the discussion of the different types of coatings. Likewise, it has been deemed desirable to devote considerable space to the preliminary preparation of metal surfaces before the application of the coating since the quality of any coating is so dependent upon this factor.

“The new monograph, therefore, covers a much broader field than did the previous one which was really a pioneer in the field of metallic coatings. The investigator of the abstruse problems of corrosion as well as the materials engineer seeking practical help in combating this problem by preventing corrosion by protecting the surface will find this volume a veritable mine of information on all phases of the subject.”

*A Synthetic Speaker.*⁴ HOMER DUDLEY, R. R. RIESZ, and S. S. A. WATKINS. This synthetic speaker is an electrical device manipulated by keys and levers for the production of synthetic vocal sounds and their combination into speech. The device was developed as an interesting educational exhibit by the Bell System at the San Francisco Exposition and the New York World's Fair.

From a buzzer-like tone and a hissing noise as raw material, the operator skillfully shapes speech by manipulating the controls to give inflection and the sound spectrum that differentiates one speech sound from another.

This paper covers the development of the device and the training of the operators to demonstrate it.

⁴ *Jour. Franklin Institute*, June 1939.

*Remotely Controlled Receiver for Radio Telephone Systems.*⁵ H. B. FISCHER. New radio receiving equipment for shore station used in ship-to-shore telephone circuits has been developed. This equipment is designed to operate on a remotely attended basis and may be located a considerable distance from the telephone terminal equipment. The radio receiver forming a part of the equipment has a codan circuit which operates reliably under high noise conditions and does not require adjustments to compensate for variations in the noise level. An emergency battery power-supply system is provided which is automatically connected to the receiver when the primary alternating-current power supply fails. Power failures are indicated at the telephone central office. A test oscillator which is controlled from the telephone central office is provided which may be used to check the operation of the receiver or to measure the frequency deviations of the incoming signals. The various apparatus units are mounted in two weather-proof cabinets which may be fastened to the same telephone pole which supports the receiving antenna.

*Analysis and Measurement of Distortion in Variable-Density Recording.*⁶ J. G. FRAYNE and R. R. SCOVILLE. Several types of non-linear distortion in variable-density recording are discussed and methods of measurement outlined. The two-frequency inter-modulation method is described. Mathematical and experimental relationships between per cent inter-modulation and per cent harmonic distortion are established. The inter-modulation method is applied to film processing for the determination of optimal negative and positive densities and overall gamma. Variance of these parameters from those indicated by classical sensitometry are traced to halation in the emulsion and to processing irregularities. The use of special anti-halation emulsions appear to reduce residual distortion effects and tend to bridge the gap between inter-modulation and sensitometric control values.

*Rubbed Films of Barium Stearate and Stearic Acid.*⁷ L. H. GERMER and K. H. STORKS. Films of barium stearate and of stearic acid have been prepared on polished chromium and on smooth natural faces of silicon carbide crystals. After these films have been rubbed with clean lens paper, electron diffraction patterns are obtained from them by the reflection method. *Well rubbed films* give patterns characteristic of a single layer of molecules standing with their axes approximately normal

⁵ *Proc. I. R. E.*, April 1939.

⁶ *Jour. S. M. P. E.*, June 1939.

⁷ *Phys. Rev.*, April 1, 1939.

to the surface; the hydrocarbon chains of barium stearate are found to be more precisely oriented than those of stearic acid; exactly the same difference exists between unrubbed single layers of molecules of barium stearate and of stearic acid deposited by the Langmuir-Blodgett method. Thickness of rubbed films on chromium has been found, by the Blodgett optical method, to be the same as that of unrubbed single layers of molecules. *Lightly rubbed films* may be thicker than a single layer of molecules. The arrangement of barium stearate in such thicker films has been found to have been somewhat altered by the rubbing. The axes of the hydrocarbon chains still stand normal to the surface, but lateral arrangement is less regular than it is in unrubbed films of equal thickness. In the case of stearic acid, molecules left on top of the first layer after light rubbing in one direction are found to lie inclined by about 8° to the surface and to point outward against the rubbing direction (Fig. 7); they are arranged in crystals having a structure different from that of the film before rubbing. Such "up-set" films of stearic acid are completely removed by very light rubbing in the direction opposite to that of the original rubbing, but they are rather resistant to light rubbing in the same direction.

*Diffraction and Refraction of a Horizontally Polarized Electromagnetic Wave over a Spherical Earth.*⁸ MARION C. GRAY. Formulas are derived for the electromagnetic field at a point on or above the surface of a spherical earth due to the presence of a vertical magnetic dipole. It is shown that the resultant field resembles that due to a vertical electric dipole above a spherical earth of low conductivity, and that in the magnetic case the values of the earth constants are of much less importance than in the electric. Curves are included showing the variation of the field with distance and with height.

*Inductive Coordination with Series Sodium Highway Lighting Circuits.*⁹ H. E. KENT and P. W. BLYE. This paper describes the wave-shape characteristics of the sodium-vapor lamp and discusses the relative inductive influence of various series circuit arrangements in which such lamps are employed. A method is outlined by means of which the noise to be expected in an exposed telephone line may be estimated. Measures are described which may be applied in the telephone plant or in the lighting circuit to assist in the inductive coordination of the two systems. These measures need be considered only when a considerable number of lamps is involved, since noise induction is negli-

⁸ *Phil. Mag.*, April 1939.

⁹ *Electrical Engineering*, Transactions Section, July 1939.

gible when there are only a few lamps as, for instance, at highway intersections.

*Sound Picture Recording and Reproducing Characteristics.*¹⁰ D. P. LOYE and K. F. MORGAN. In the improvement of sound motion pictures, the trend has been to make the response of all parts of the recording and reproducing circuits as nearly "flat" as possible. In some cases, however, this has resulted in unnatural sound, and therefore certain empirical practices have been adopted in the studios and theaters to make pictures sound best.

This paper describes the results of a study the purpose of which has been to evaluate the factors which affect the quality of speech as recorded and reproduced, from the vocal cords of the actor on the sound-stage to the brain of the listener in the theater. The characteristics of the various factors have been determined and combined with dialog, voice effort, and other equalizers designed to produce an overall characteristic "subjectively flat" at the brain of the theater patron. These factors, as well as others which are now in the process of being studied, are presented in this paper.

One of the most important characteristics studied is that of the change in voice quality with a change in the effort on the part of the speaker. This is described in detail in this paper. The stage and set acoustic characteristics, microphone characteristic, and dialog equalization to compensate principally for the hearing characteristic of the average theater listener, are among the factors described herein.

*A Dynamic Measurement of the Elastic, Electric and Piezoelectric Constants of Rochelle Salt.*¹¹ W. P. MASON. The elastic, electric and piezoelectric constants of Rochelle salt have been measured at low field strengths by measuring the resonant frequencies and impedance of vibrating crystals. It is shown experimentally that the resonant and anti-resonant frequencies of the crystal are both considerably below the natural mechanical resonant frequency of the crystal in disagreement with the usual derivation of the frequencies of a piezoelectric crystal. By assuming that the piezoelectric stress is proportional to the charge density on the electrodes rather than the potential gradient as usually assumed, theoretical frequencies are obtained which agree with those found experimentally. This theoretical derivation together with the measured frequencies supply values for the piezoelectric constants. The elastic constants measured dynamically show some differences from those measured statically. A large difference is

¹⁰ *Jour. S. M. P. E.*, June 1939.

¹¹ *Phys. Rev.*, April 15, 1939.

found for the dynamically measured piezoelectric constants from those statically measured, which may be attributed to the finite relaxation time for the piezoelectric elements.

*A Vogad for Radio Telephone Circuits.*¹² S. B. WRIGHT, S. DOBA, and A. C. DICKIESON. Commercial radio telephone connections must generally be accessible to any telephone in an extensive wire system. Speech signals delivered to the radio terminals for transmission to distant points vary widely in amplitude due to the characteristics of the wire circuits and individual voices. To provide the best margin against atmospheric noise, it is usually the practice to equalize this wide range of speech amplitudes and thus drive the radio telephone transmitter at its full capacity.

Many devices have been proposed to adjust automatically the gain in a circuit to equalize speech volumes. The difficulties of providing a device which will respond properly over a wide range to the complex qualities of a speech signal have only recently been overcome to a satisfactory degree.

The voice-operated gain-adjusting device, or "vogad," described in this paper is a practical design based upon more than a year's experience with one of the most promising devices made available by earlier development effort. A trial installation of this latest vogad is now under way at Norfolk in connection with a new radio telephone system for harbor and coastal service.

¹² *Proc. I. R. E.*, April 1939.

Contributors to this Issue

JOHN R. CARSON, B.S., Princeton, 1907; E.E., 1909; M.S., 1912, D.Sc. (Honorary), 1936. American Telephone and Telegraph Company, 1914-34; Bell Telephone Laboratories, 1934-. As Transmission Theory Engineer for the American Telephone and Telegraph Company and later for the Laboratories, Dr. Carson has made substantial contributions to electric circuit and transmission theory and has published extensively on these subjects. The Franklin Institute of Philadelphia recently awarded him the Elliott Cresson Medal. He is now a research mathematician.

J. G. CHAFFEE, S.B., Massachusetts Institute of Technology, 1923. Western Electric Company, Engineering Department, 1923-25; Bell Telephone Laboratories, 1925-. Mr. Chaffee has been engaged in the study of radio problems at ultra-high frequencies.

W. J. CLARKE, B.Chem., Cornell University, 1924; M.A., Columbia University, 1932. Research Laboratory, Devoe and Reynolds Company, 1924-30. Bell Telephone Laboratories, 1930-. Mr. Clarke was at first engaged in studies of organic finishes for telephone equipment, particularly on the compounding of improved finishing materials. More recently his work has been concerned with investigations of molding plastic materials.

VICTOR E. LEGG, B.A., 1920, M.S., 1922, University of Michigan. Research Department, Detroit Edison Company, 1920-21; Bell Telephone Laboratories, 1922-. Mr. Legg has been engaged in the development of magnetic materials and in their applications, particularly for the continuous loading of cables, and for compressed dust cores.

S. O. MORGAN, B.S. in Chemistry, Union College, 1922; M.A., Princeton University, 1925; Ph.D., 1928. Western Electric Company, Engineering Department, 1922-24; Bell Telephone Laboratories, 1927-. As Dielectric Research Chemist, Dr. Morgan is concerned with the relation between dielectric properties and chemical composition.

E. J. MURPHY, B.S., University of Saskatchewan, Canada, 1918; McGill University, Montreal, 1919-20; Harvard University, 1922-23. Western Electric Company, Engineering Department, 1923-25; Bell

Telephone Laboratories, 1925-. Mr. Murphy's work is largely confined to the study of the electrical properties of dielectrics.

LISS C. PETERSON. Chalmers Technical Institute, Gothenburg, 1920; Technical Universities of Charlottenburg and Dresden, 1920-23. American Telephone and Telegraph Company, 1925-30; Bell Telephone Laboratories, 1930-. Mr. Peterson is engaged in amplifier research.

JOHN R. TOWNSEND, Brooklyn Polytechnic Institute. Bethlehem Steel Company, 1915-17. Mathematics and Dynamics Branch, U. S. Ordnance Department, 1917-19. Member of Technical Staff, Western Electric Company, 1919-25; Bell Telephone Laboratories, 1925-. He is now Materials Standards Engineer. He is the author of "Fatigue Studies of Telephone Cable Sheath Alloys," "Physical Properties and Methods of Test for Some Sheet Non-ferrous Metals," and also of many other articles published in technical magazines and discussed before engineering societies. Awarded Dudley Medal, A.S.T.M., 1930.

THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS
OF ELECTRICAL COMMUNICATION

Experience in Applying Carrier Telephone Systems to Toll
Cables—*W. B. Bedell, G. B. Ransom and W. A. Stevens* 547

The Toronto-Barrie Toll Cable
—*M. J. Aykroyd and D. G. Geiger* 588

The Computation of the Composite Noise Resulting from Ran-
dom Variable Sources—*E. Dietze and W. D. Goodale, Jr.* 605

Load Rating Theory for Multi-Channel Amplifiers
—*B. D. Holbrook and J. T. Dixon* 624

The Quantum Physics of Solids, I. The Energies of Electrons
in Crystals—*W. Shockley* 645

Dial Clutch of the Spring Type—*C. F. Wiebusch* 724

Abstracts of Technical Papers 742

Contributors to this Issue 747

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

THE BELL SYSTEM TECHNICAL JOURNAL

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

EDITORIAL BOARD

F. B. Jewett

H. P. Charlesworth

W. H. Harrison

A. F. Dixon

O. E. Buckley

O. B. Blackwell

S. Bracken

M. J. Kelly

G. Ireland

W. Wilson

R. W. King, *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, 1939
American Telephone and Telegraph Company

The Bell System Technical Journal

Vol. XVIII

October, 1939

No. 4

Experience in Applying Carrier Telephone Systems to Toll Cables

By W. B. BEDELL, G. B. RANSOM and W. A. STEVENS

THE application of carrier telephone systems to toll cable conductors, particularly those conductors in existing cables, is expected to become an important means of providing additional long distance telephone circuits. Eight hundred and thirty-seven route miles have been equipped in the United States to date, and 17 twelve-channel systems have been placed in service, providing a total of about 58,000 circuit miles. Late in 1939, 200 additional route miles are expected to be completed which, together with additional systems on existing routes, will add nine systems and about 48,000 circuit miles to the above figures.

The type of carrier system which has been installed is that described by Messrs. C. W. Green and E. I. Green before the American Institute of Electrical Engineers in 1938, and which is now designated as type K.¹ The problems incident to the application of this system to toll cable conductors may be of general interest and it is the purpose of this paper to describe some of these. This description will start at the point where traffic needs have indicated that additional circuits should be provided along a given route and economic and other considerations have shown that they should be provided by means of type K cable carrier telephone systems. For specific examples, reference will be made to the New York-Charlotte and Detroit-South Bend projects, in which sections the application of the initial type K carrier systems has been completed. Figure 1 shows the geographical location of these installations, as well as some of those sections where type K carrier systems will probably be installed in the future.

¹ For references see end of paper.

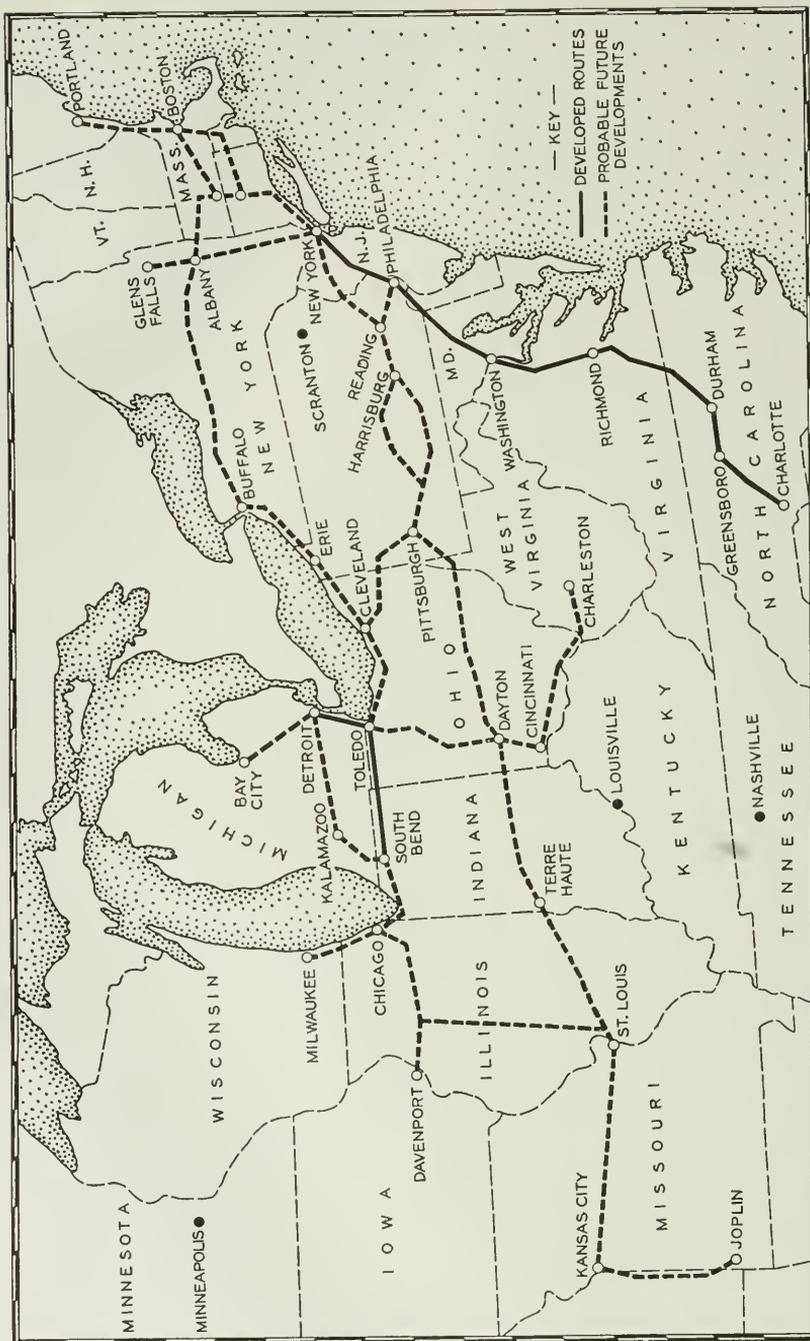


Fig. 1—Routes of existing and probable future type K cable carrier developments.

SELECTION OF CABLES TO BE EQUIPPED

Where more than two cables existed on a route selected for carrier operation, a number of factors influenced the selection of cables to be equipped initially. Among these were the ages of the cables, their makeup, specific route, number of branch cables and open wire junctions, and lengths of underground cable involved. Between Detroit and South Bend there were but two cables on the route selected and hence no selection was necessary. Between New York and Washington on the New York-Charlotte route all cables are underground, and since there were from three to six cables in each repeater section, the two cables which it was decided to employ were selected because they were relatively new and had the smallest number of branches. From Washington to Petersburg, Va., there were two cables, while between Petersburg and Charlotte there was but one cable and it was necessary to install a small second cable chiefly for carrier operation.

The Petersburg-Greensboro section of this second cable was installed one year ahead of the carrier application in order to make use of part of its conductors which were loaded for voice frequency operation. This cable is made up in most sections of 32 quads of 19-gauge conductors, of which 20 are loaded with H-88-50 loading units, leaving 10 quads non-loaded for carrier use and two for maintenance purposes.

The second cable in the Greensboro-Charlotte section was installed coincidentally with the installation of the initial carrier systems. This cable contains 61 non-quadded pairs throughout, except in certain sections where it also contains some loaded conductors for short voice frequency circuits. Paired construction was used because it was expected to be slightly more economical and temporary voice usage of the conductors was not planned.

One additional factor which, in special cases, influences the selection of cables is that of carrier repeater spacing. This is brought about by the fact that on multi-cable routes all of the cables may not follow exactly the same route. For example, one cable may be aerial and the other underground, and the two may be separated in some sections; or underground cables, for conduit reasons, may follow different routes. It is desirable that the two cables used be near each other at repeater points.²

One interesting feature in the construction of the Greensboro-Charlotte Cable is the method by which the cable was attached to the messenger. The cable was lashed to the messenger by means of a galvanized steel wire continuous between poles, as shown in Fig. 2. This method of installation is expected to reduce buckling, ring cuts,

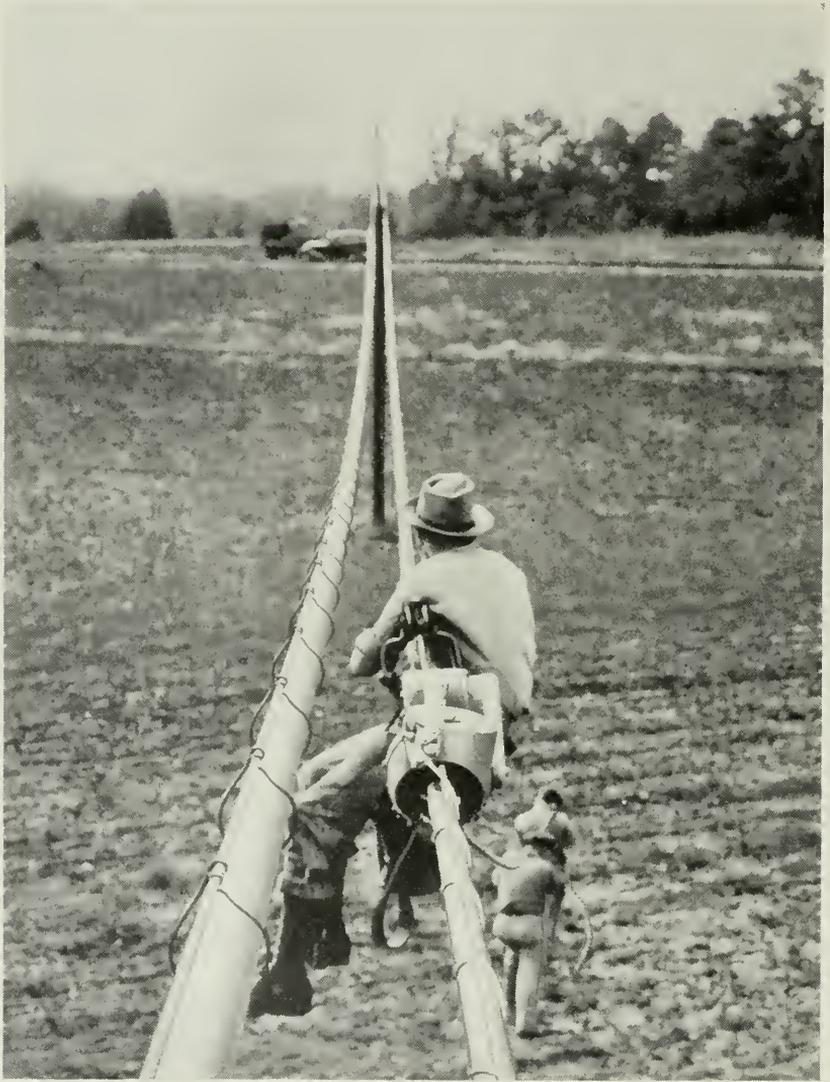


Fig. 2—Method of lashing cable to messenger with galvanized steel wire shown in progress.

and jumping, and avoids the necessity of splicing the cable under tension.

SELECTION OF CONDUCTORS FOR CARRIER USE

In general, non-loaded cable pairs are not available in existing cables and it is necessary to remove voice frequency loading from pairs

intended for carrier operation. As has been described previously in a paper in this *Journal*² the crosstalk mitigation plans in connection with type K carrier are designed on the assumption that cable pairs will be developed in units of 20 pairs for each direction of transmission. Further, the design of this carrier system contemplates the use of 19-gauge pairs.

Ten quads (20 pairs) were, therefore, selected in each cable in which carrier operation was planned. These quads were selected, for crosstalk reasons, from a large voice complement. Two-wire facilities may be used for carrier where a sufficiently large complement exists. This results, however, in the loss of twice as many voice circuits as compared to unloading four-wire quads. In the sections unloaded to date it has been impracticable to unload two-wire facilities.

Where four-wire facilities were used, five quads from the groups designed for each direction of voice frequency transmission in each cable were selected. Over 20,000 circuit miles of four-wire facilities have been unloaded for carrier use. Of this total, H-174-63 loading units were removed from 2,280 circuit miles, and H-44-25 units were removed from the remainder. The H-44-25 loading units removed from 2,475 miles of four-wire circuits were transferred to two-wire 16-gauge quads loaded with H-174-63 units in the same cable, and these latter units released, thus providing at small cost transmission improvement on a total of 4,950 circuit miles.

PREPARATION OF CABLE CONDUCTORS

Coincidentally with the removal of the loading from the quads selected for carrier operation, special splicing work was performed for crosstalk and transmission reasons. The exact method of making these splices depended upon the layup of the cable involved. For example, if the cable involved concentric segregation, the five former east bound quads were spliced at random to the five former west bound quads and vice versa at each loading point; in cables involving split layer segregation, the ten quads were spliced at each loading point in a planned random manner.

The removal of the loading at the point nearest the center of each carrier repeater section was left until last, so that a special splice, called a poling splice, based on measurements of within-quad admittance unbalances, might be made at each such point.⁵ These measurements could not be made until all loading coils on the carrier pairs in the repeater section had been removed. Using these measurements as a guide the quads in one half-section were connected to quads in

the other half-section so that the unbalances in the two sections tended to compensate. Table 1 shows for a typical type K repeater section

TABLE 1
MEASUREMENTS OF MUTUAL INDUCTANCE UNBALANCE (G) AND CAPACITANCE UNBALANCE (C) BEFORE AND AFTER POLING ON QUADS IN THE PHILADELPHIA-PHILADELPHIA KN SECTION OF THE NEW YORK-PHILADELPHIA E CABLE

Pairs	Before Poling		After Poling	
	G Micromho	C Mmf.	G Micromho	C Mmf.
1-2	.16	110	.12	50
3-4	.12	40	.01	30
5-6	.12	80	.01	45
7-8	.01	40	.02	0
9-10	.07	85	.05	5
11-12	.06	0	0	0
13-14	.10	10	.04	25
15-16	.13	65	0	45
17-18	.11	30	0	20
19-20	.08	25	0	10

the unbalance measurements before poling and the final results after the poling splice was made. These measurements were made at voice frequencies, since as discussed in a previous paper² satisfactory results were obtained at these frequencies. It will be noted that poling reduced markedly the unbalances in the quads in the section shown. This is particularly true of the mutual inductance unbalances, indicated in the table by G, the reduction of which is important in reducing crosstalk at carrier frequencies.

After carrying out this and the balancing operations described later for the reduction of crosstalk, it was still necessary to take other steps to reduce the within-quad crosstalk. The recurrence of within-quad coupling between carrier systems which may be assigned to two of the pairs in a 20-pair carrier complement, has been reduced by means of a splicing plan so worked out, that two given carrier systems will operate on the same quad as infrequently as possible. This was accomplished by splitting the quads at the ends of each carrier repeater section on a planned basis. The plan is shown in Table 2. Nineteen types of splices are shown. This table is used as a guide in performing the splice between the balancing bays and the input sealed test terminal. For example, in performing splice type 8, sealed test terminal jacks K-1 are connected to the pair designated 8 at the balancing bay cable terminal, jacks K-2 to the pair designated 16, jacks K-3 to

TABLE 2

DETAILED PLAN FOR SPLITTING QUADS IN NINETEEN CONSECUTIVE CARRIER REPEATER SECTIONS FOR A TEN-QUAD GROUP ARRANGED FOR TYPE K CABLE CARRIER OPERATION

Pair Designation at Sealed Test Terminal Jacks	Planned Splice Type Number																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
K-1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
K-2	2	3	5	7	9	12	14	16	18	20	1	2	4	6	8	10	13	15	17
K-3	3	4	7	10	14	17	20	3	6	9	19	13	16	19	2	5	8	12	15
K-4	4	5	9	14	18	2	6	10	15	19	3	3	7	12	16	20	4	8	13
K-5	5	6	12	17	2	7	13	18	3	8	17	14	19	4	9	15	20	5	10
K-6	6	7	14	20	6	13	19	5	12	18	5	4	10	17	3	9	16	2	8
K-7	7	8	16	3	10	18	5	13	20	7	15	15	2	9	17	4	12	19	6
K-8	8	9	18	6	15	3	12	20	8	17	7	5	14	2	10	19	7	16	4
K-9	9	10	20	9	19	8	18	7	17	6	13	16	5	15	4	14	3	13	2
K-10	10	11	11	11	11	11	11	11	11	11	9	11	11	11	11	11	11	11	11
K-11	11	12	2	13	3	14	4	15	5	16	14	6	17	7	18	8	19	9	20
K-12	12	13	4	16	7	19	10	2	14	5	8	17	8	20	12	3	15	6	18
K-13	13	14	6	19	12	4	17	9	2	15	16	7	20	13	5	18	10	3	16
K-14	14	15	8	2	16	9	3	17	10	4	6	18	12	5	19	13	6	20	14
K-15	15	16	10	5	20	15	9	4	19	14	18	8	3	18	13	7	2	17	12
K-16	16	17	13	8	4	20	16	12	7	3	4	19	15	10	6	2	18	14	9
K-17	17	18	15	12	8	5	2	19	16	13	20	9	6	3	20	17	14	10	7
K-18	18	19	17	15	13	10	8	6	4	2	2	20	18	16	14	12	9	7	5
K-19	19	20	19	18	17	16	15	14	13	12	12	10	9	8	7	6	5	4	3
K-20	20	1	1	1	1	1	1	1	1	1	10	1	1	1	1	1	1	1	1

Note: Above numbers are pair number designations at balancing bay cable terminals of pairs which are connected to Sealed Test terminal jacks as indicated.

pair 3, etc. Using this plan two systems are exposed to each other on the same quad only once in carrier repeater sections. Beginning with the 20th section the plan is repeated so that between New York and Charlotte two given systems operate together on the same quad in only two repeater sections. Planned splices were made at each end of each carrier repeater section so that after the quads had been split in a definite way at one end of a section, a complementary splice was made at the other end, to rearrange the pairs into a given order as they go through each repeater office. Each carrier pair has been made to appear in the same position at each carrier testboard throughout the 19 types of planned splice sections. This is for convenience in maintaining and identifying them, because carrier systems must be assigned to the same carrier pair in a series throughout the 19 types of planned splice sections if the quad splitting plan is to be completely effective.

The poling splice and the planned splices were, of course, not re-

quired in the paired cable placed for carrier operation between Greensboro and Charlotte.

Far-end crosstalk was still further reduced by means of balancing coils installed at the end of each repeater section connected to the repeater inputs.^{2, 8} After the splicing operations just discussed were completed and the balancing coils were installed and connected to the carrier pairs, the coupling between each pair and each other pair of the carrier complement was reduced to the lowest value practicable by adjustment of these coils. This was accomplished by sending a disturbing testing tone on one pair, receiving on each other pair in turn, and adjusting the coil which couples each combination of two pairs until a minimum amount of the testing tone was measured on the disturbed pair. Figure 3 shows these adjustments in progress while Table 3 shows a summary of the final crosstalk measurements made after the adjustments. About 19,000 balancing coils have been installed on the carrier routes equipped to date.

At two points on the New York-Charlotte route, retardation coils which are described later were used to increase the attenuation in potential crosstalk paths. At Petersburg, Va., and Burlington, N. C., 60 and 4 voice quads, respectively, connected direct from one carrier cable to the other. These quads, through secondary induction, were likely to serve as crosstalk paths at carrier frequencies between the two cables. Retardation coils were installed in each quad to limit crosstalk currents. Two other situations, where quads connected between the carrier cables, were eliminated by cable rearrangements.

LATERAL CABLES

At each carrier repeater station four lateral cables were installed to bring the carrier pairs into the repeater building. One of these was required for each direction of transmission for each cable; that is, two input cables and two output cables were required. As a result of the transposition of directions of transmission at each repeater point, the two input cables connect to one toll cable and the two output cables to the other. Figure 4 shows these cables installed at an aerial cable repeater point. All lateral cables were installed for the probable ultimate capacity for carrier systems of the cables being developed; that is, 100 systems on the Detroit-South Bend route, 100 systems north and 60 systems south of Richmond, Va., on the New York-Charlotte route.

THE LOCATION OF CARRIER REPEATER STATION SITES

The next important step in the development of a route for type K carrier operation is the selection of points at which intermediate



Fig. 3—Adjustment of coils in crosstalk balancing panel.

TABLE 3
FINAL FAR-END CARRIER CROSTALK MEASUREMENTS AFTER BALANCING
COIL ADJUSTMENTS

Section	New York Toward Charlotte				Charlotte Toward New York					
	Ca.	Crosstalk Units 39.85 kc		Crosstalk Units 28.15 kc		Ca.	Crosstalk Units 39.85 kc		Crosstalk Units 28.15 kc	
		R.M.S.	Max.	R.M.S.	Max.		R.M.S.	Max.	R.M.S.	Max.
New York-N. Y. KS.....	E	42	283	36	122	F	33	184	31	100
N. Y. KS-Prin. KN.....	F	37	148	36	122	E	29	95	27	92
Prin. KN-Prin.....	E	38	130	36	114	F	40	212	37	155
Prin.-Prin. KS.....	G	37	212	33	130	E	35	132	36	130
Prin. KS-Phila. KN.....	E	35	148	32	164	G	30	100	28	112
Phila. KN-Phila.....	G	35	250	31	100	E	29	114	29	127
Phila.-Phila. KS.....	D	45	243	44	226	F	37	127	33	122
Phila. KS-Elk. KN.....	F	29	116	28	112	D	45	138	45	122
Elk. KN-Elk.....	D	45	161	45	145	F	36	145	36	130
Elk.-Elk. KS.....	F	36	204	36	164	D	42	217	42	176
Elk. KS-Balt. KN.....	D	38	126	39	129	F	35	207	33	145
Balt. KN-Balt.....	F	35	107	34	114	D	45	224	45	170
Balt.-Wash. KN.....	D	44	241	42	179	E	43	219	38	148
Wash. KN-Wash.....	E	39	179	38	195	D	34	100	36	126
Wash.-Wash. KS.....	A	43	184	43	148	B	44	182	48	158
Wash. KS-Fred. KN.....	B	43	224	40	167	A	47	141	46	155
Fred. KN-Fred.....	A	54	179	50	163	B	44	170	46	208
Fred.-Fred. KS.....	B	45	173	43	152	A	49	219	48	219
Fred. KS-Rich. KN.....	A	47	167	46	164	B	39	167	39	127
Rich. KN-Rich.....	B	43	167	46	176	A	39	200	38	125
Rich.-Rich. KS.....	A	39	179	40	127	B	30	190	39	158
Rich. KS-McK. KN.....	B	48	138	53	155	A	37	265	35	200
McK. KN-McK.....	A	36	130	34	100	B	56	148	64	167
McK.-McK. KS.....	B	57	190	62	174	A	38	152	38	155
McK. KS-Norl. KN.....	A	36	145	35	145	B	56	161	62	187
Norl. KN-Norl.....	B	53	173	56	170	A	41	198	39	190
Norl.-Norl. KS.....	A	34	110	38	134	B	59	167	59	170
Norl. KS-Dur. KN.....	B	56	179	56	195	A	39	142	40	142
Dur. KN-Dur.....	A	42	190	38	190	B	65	200	62	187
Dur.-Dur. KS.....	B	61	155	58	145	A	38	114	43	134
Dur. KS-Gbo. KN.....	A	42	195	42	118	B	59	205	59	187
Gbo. KN-Gbo.....	B	57	219	58	195	A	37	118	36	107
Gbo.-Gbo. KS.....	A	39	145	37	141	B	55	173	50	155
Gbo. KS-Sal. KN.....	B	57	338	54	346	A	40	224	38	161
Sal. KN-Sal.....	A	34	161	35	155	B	43	245	50	300
Sal.-Sal. KS.....	B	54	265	52	245	A	35	126	32	107
Sal. KS-Chlot. KN.....	A	34	141	32	155	B	53	245	53	286
Chlot. KN-Chlot.....	B	43	200	42	155	A	33	167	31	129

Note: These measurements were made in connection with balancing work where relative values are important and do not necessarily represent the absolute magnitude of the crosstalk.

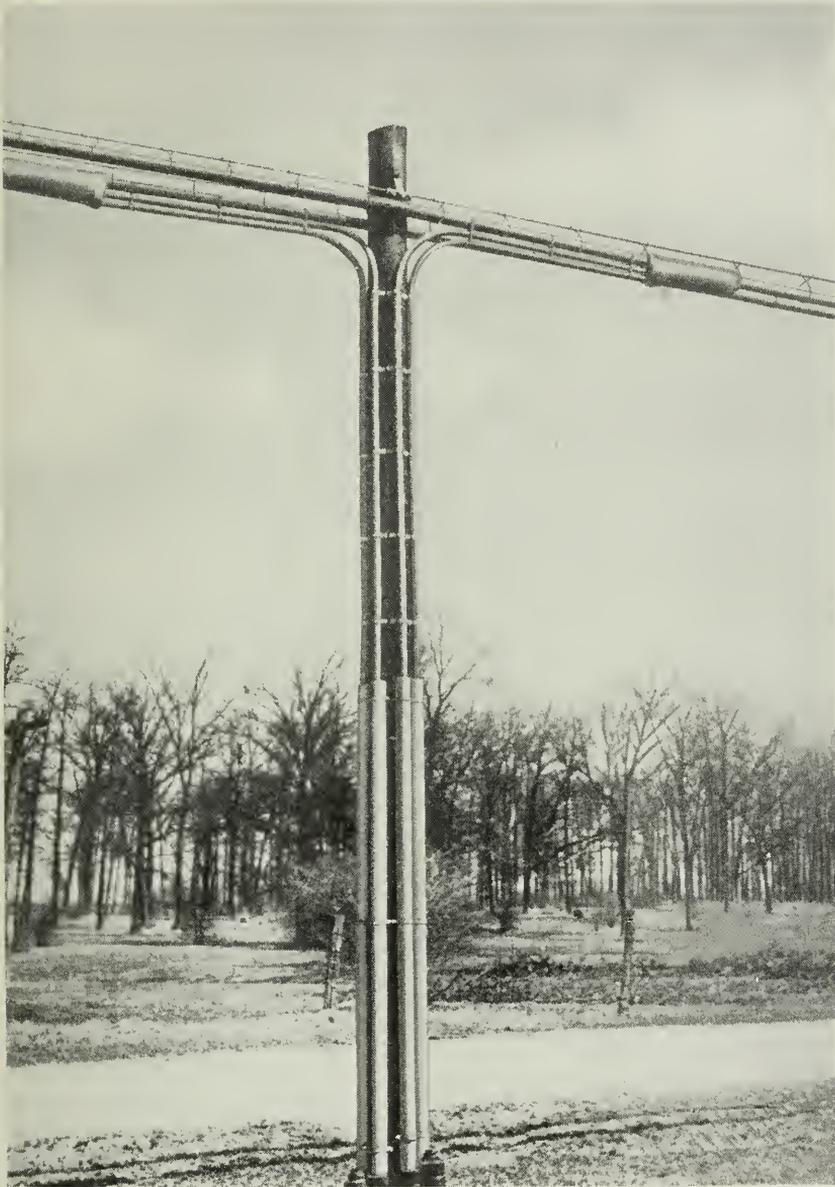


Fig. 4—The four lateral cables containing carrier pairs shown as installed at an auxiliary repeater point on aerial cables.

repeater stations, known as auxiliary repeater stations, will be constructed. The selection of these locations is necessary before the transmission design of a carrier route can be completed. Repeaters for voice frequency circuits are located on existing cable routes at an average spacing of about 45 miles. These same offices are used as cable carrier repeater points, but, because of the high losses at carrier frequencies, two additional carrier repeater stations, on the average, have been provided between each two voice frequency repeater offices.

The cables and routes having been tentatively decided upon, the route records were studied and locations which were the most practicable from a transmission standpoint were selected. These selections were influenced by the expected maximum section losses, taking into account aerial and underground construction. In general, the distances between the existing voice repeater points were divided into the smallest number of equal parts, the lengths of which did not exceed the maximum permissible carrier repeater spacing, and repeater station sites were tentatively located at the junctions of these parts.

A physical inspection of these tentative sites was then made and where necessary an alternate site selected. Such factors as accessibility of site, suitability for building, availability of primary power, cost of real estate, and willingness of owners to sell determined whether or not the tentative site could be used and, if not, what alternate location might be used. Where a suitable existing telephone building happened to be located near a proposed repeater station location, the possibility of using such building was studied. It has been practicable, however, in only one instance to use an existing building in installing the 34 auxiliary stations provided to date. The sites which were considered satisfactory after the physical inspection were then examined to check their suitability from a carrier transmission standpoint. In cases where transmission limits had been exceeded, the sites were reinspected and compromise locations finally agreed upon. In most cases it has not been difficult to find sites which are suitable both from a transmission and a construction standpoint. Of the 34 type K carrier repeater stations which have been built, 20 were constructed at the sites originally selected from a transmission standpoint. In most of the other cases the final sites were within a short distance of the originally selected locations. In a few cases, however, where ideal sites fell in populous centers or comparatively inaccessible wilds, it was necessary to take unusual steps.

In one case the site which had been selected from a transmission standpoint fell at a location where the two cables which had been selected followed different conduit runs and were separated by more

than $2\frac{1}{2}$ miles. Two lateral cables, each of that length, would have been required in order to make use of this site. Moving the location back to the point where the cables came together would have resulted in an excessively long carrier repeater section and would have located the station within the business section of Wilmington, Del. The problem was to find a compromise site between these two points where the lengths of lateral cables would not be excessive and the repeater section could be kept within desirable limits. A tide water stream between these two points complicated the problem.

Nine sites were inspected and four of them studied in detail. Flood and fire hazards, as well as high prices of real estate, were added to the other factors governing the choice. None of the locations was entirely desirable, but a compromise choice was finally made of a location which resulted in the longest repeater spacing in the New York-Charlotte project, but the lengths of the lateral cables were reduced to between eleven and twelve hundred feet.

Table 4 shows the theoretical spacings which, considering the fixed

TABLE 4

COMPARISON OF THE ORIGINALLY SPECIFIED CARRIER REPEATER SPACINGS AND ACTUAL SPACINGS ON THE NEW YORK-CHARLOTTE AND DETROIT-SOUTH BEND CABLE ROUTES

Project	No. of Repeater Sections	Theoretically Best Spacings			Actual Spacings Used		
		Min.	Ave.	Max.	Min.	Ave.	Max.
New York-Charlotte...	38	13.5	16.5	18.8	13.2	16.5	20.2
Detroit-South Bend...	13	14.9	16.1	17.6	13.6	16.1	18.4

location of existing repeater points, were selected as best from a transmission standpoint, and the actual spacings which it was found necessary to use. The theoretically best spacings for the two routes differed because of the difference in spacings of the existing voice frequency repeater points of which use has been made as carrier repeater points. Figure 5 shows the repeater office locations as they are distributed on the New York-Charlotte project. The various types of repeater offices shown are discussed later in this paper.

DIRECTION OF TRANSMISSION

The circuits used for the two directions of transmission of type K carrier systems operate in separate cables on the projects so far completed and, for crosstalk reasons, these two directions of transmission have been transposed between the two cables at each carrier repeater point.² The location of branch cables and taps to open wire have an

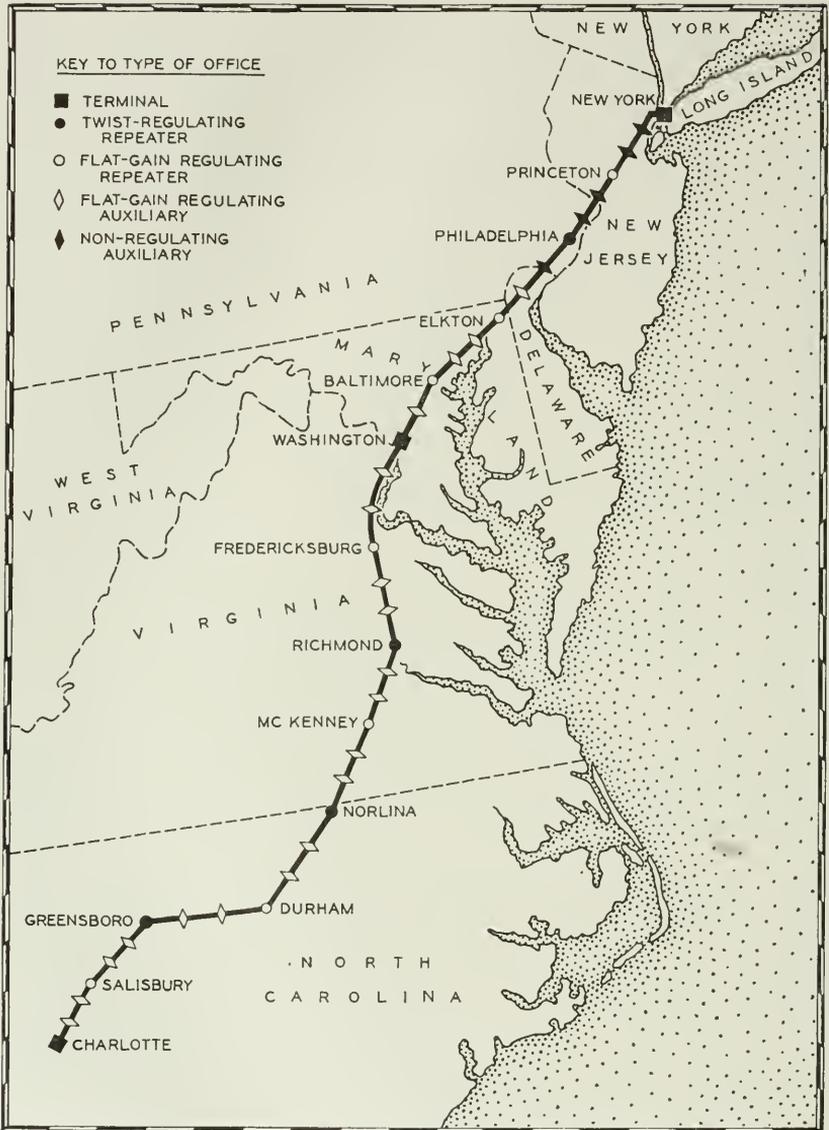


Fig. 5—Route of New York-Charlotte cable carrier development.

important bearing on the selection of the direction of transmission, since it is desirable to assign directions of transmission which will result in a minimum number of taps to open wire and branch cables occurring near carrier repeater inputs. Where the lengths of under-

ground construction adjacent to a repeater station differ on the two cables, it is preferable to have pairs in the cable having the longer section of underground connected to repeater inputs. With these and other factors in mind, tentative directions of transmission were assigned to the cable conductors and computations of expected noise currents made and checked with computations which assumed the directions reversed. The total overall noise currents computed to be 1.25 db better when assuming the directions of transmission finally selected for the New York-Charlotte project than when assuming these directions reversed. This was largely due to the fact that south of Petersburg, where but one cable had existed, a small cable was added to permit carrier operation. Pairs in this cable, because of its small size, are more susceptible to static induction directly into the cable than pairs in larger cables. The effect, therefore, of the greater contribution to overall noise currents, which this small cable tends to cause, was reduced by selecting the directions of transmission so as to take advantage of the increased shielding resulting from underground construction adjacent to repeaters. Tables 5 and 6 show the final noise level computations for the 1000-cycle point of channel 12 (57 kc on the line) of a New York-Charlotte system. It will be noted that the longer repeater sections contribute a great deal of noise as compared to average or shorter sections. Noise measurements which have been made indicate that noise conditions compare favorably with those which it was calculated might be expected.

NON-REGULATING REPEATER POINTS

Examination of the carrier repeater sections on the New York-Charlotte route showed seven to be unusually short and involving all underground cable construction. The usual plan would have been to provide flat gain regulation at each carrier repeater point, but since these seven sections averaged but 14.8 miles in length and the theoretical transmission variation might be but ± 1.42 db, it was obvious that the regulators having a normal range of ± 7.15 db would be required to operate only over a small part of their range. However, the real limitation is not the regulating mechanism but the lower levels to which the line currents without regulated gain would drop during periods of high cable temperature with corresponding impairments in noise levels. In this layout, omission of regulation at one station increases the noise level about the same as lengthening the following repeater section about $\frac{3}{4}$ of a mile. Omission of two successive regulators is approximately equivalent to increasing the second repeater section about $\frac{3}{4}$ of a mile and the third about $1\frac{1}{2}$ miles. Repeater

TABLE 5
NEW YORK-CHARLOTTE TYPE K CARRIER NOISE COMPUTATIONS

Section	Cable	Miles			57 kc Loss Max. Temp. ¹	Estimated 57 kc Noise Level ² at Rept. Output
		U.G.	Aerial	Total		
New York-New York KS.....	E	15.45	—	15.45	58.09	— 8.91
New York KS-Princeton KN.....	F	14.45	—	14.45	54.33	— 9.71
Princeton KN-Princeton.....	E	16.59	—	16.59	62.38	+ 1.10
Princeton-Princeton KS.....	G	13.22	—	13.22	49.71	—13.29
Princeton KS-Philadelphia KN.....	E	13.92	—	13.92	52.34	—10.75
Philadelphia KN-Philadelphia.....	G	14.79	—	14.79	55.61	— 2.19
Philadelphia-Philadelphia KS.....	D	15.14	—	15.14	56.93	—10.07
Philadelphia KS-Elkton KN.....	F	13.49	—	13.49	50.72	— 9.38
Elkton KN-Elkton.....	D	19.70	—	19.70	74.07	+ 8.57
Elkton-Elkton KS.....	F	17.28	—	17.28	64.97	— 2.03
Elkton KS-Baltimore KN.....	D	16.89	—	16.89	63.51	— 3.49
Baltimore KN-Baltimore.....	F	16.91	—	16.91	63.58	— 3.42
Baltimore-Washington KN.....	D	18.61	—	18.61	69.97	+ 2.97
Washington KN-Washington.....	E	18.95	—	18.95	71.25	+ 4.25
Washington-Washington KS.....	A	7.31	11.28	18.59	71.93	+ 5.93
Washington KS-Fredericksburg KN.....	B	—	18.45	18.45	72.69	+ 6.69
Fredericksburg KN-Fredericksburg.....	A	.10	16.54	16.64	65.55	+ 2.95
Fredericksburg-Fredericksburg KS.....	B	.11	18.23	18.34	72.24	+ 6.24
Fredericksburg KS-Richmond KN.....	A	—	18.26	18.26	71.94	+ 5.94
Richmond KN-Richmond.....	B	7.64	10.99	18.63	72.03	+ 5.03
Richmond-Richmond KS.....	A	9.82	7.09	16.91	64.85	— 1.15
Richmond KS-McKenney KN.....	B	10.13	5.85	15.98	61.14	+ 8.54
McKenney KN-McKenney.....	A	.08	15.43	15.51	61.09	— 1.51
McKenney-McKenney KS.....	B	.11	16.87	16.98	66.88	+14.08
McKenney KS-Norlina KN.....	A	—	15.23	15.23	60.01	— 2.59
Norlina KN-Norlina.....	B	.06	14.28	14.34	56.49	+ 3.89
Norlina-Norlina KS.....	A	—	16.36	16.36	64.46	+ 3.66
Norlina KS-Durham KN.....	B	—	16.68	16.68	65.72	+13.12
Durham KN-Durham.....	A	.12	17.62	17.74	69.87	+ 3.87
Durham-Durham KS.....	B	.05	18.05	18.10	71.31	+18.51
Durham KS-Greensboro KN.....	A	—	17.79	17.79	70.09	+ 4.09
Greensboro KN-Greensboro.....	B	2.38	16.25	18.63	72.98	+12.18
Greensboro-Greensboro KS.....	A	5.64	12.18	17.82	69.20	+ 3.20
Greensboro KS-Salisbury KN.....	B	—	16.41	16.41	64.66	+12.06
Salisbury KN-Salisbury.....	A	1.85	14.84	16.69	65.43	— 2.53
Salisbury-Salisbury KS.....	B	1.82	11.92	13.74	53.80	+ 1.2
Salisbury KS-Charlotte KN.....	A	—	14.47	14.47	57.01	— 4.09
Charlotte KN-Charlotte.....	B	3.78	9.96	13.74	53.45	— 8.35
New York-Charlotte Totals.....				627.42		+23.27 ³

¹ Attenuation figures used: 3.76 for U.G. at 73° and 3.94 for Aerial at 110°.

² For top channel, referred to — 9 db switchboard level.

³ Computed on a root-sum-square basis.

sections adjacent to New York and Philadelphia are short and it was decided, therefore, to omit regulation from the two auxiliary stations between New York and Princeton, N. J., two between Princeton and Philadelphia, and one between Philadelphia and Wilmington, Del.

TABLE 6
CHARLOTTE-NEW YORK TYPE K CARRIER NOISE COMPUTATIONS

Section	Cable	Miles			57 kc Loss Max. Temp. ¹	Estimated 57 kc Noise Level ² at Rept. Output
		U.G.	Aerial	Total		
Charlotte-Charlotte KN	A	3.78	9.96	13.74	53.45	- 9.15
Charlotte KN-Salisbury KS	B	—	14.47	14.47	57.01	+ 5.91
Salisbury KS-Salisbury	A	1.82	11.92	13.74	53.80	- 7.60
Salisbury-Salisbury KN	B	1.85	14.84	16.69	65.43	+14.13
Salisbury KN-Greensboro KS	A	—	16.41	16.41	64.66	+ 2.06
Greensboro KS-Greensboro	B	5.64	12.18	17.82	69.20	+ 4.70
Greensboro-Greensboro KN	A	1.66	16.97	18.63	73.10	+ 7.10
Greensboro KN-Durham KS	B	—	17.79	17.79	70.09	+17.29
Durham KS-Durham	A	.05	18.05	18.10	71.31	+ 5.31
Durham-Durham KN	B	.12	17.62	17.74	69.87	+17.07
Durham KN-Norlina KS	A	—	16.68	16.68	65.72	+ 3.12
Norlina KS-Norlina	B	—	16.36	16.36	64.46	+11.86
Norlina-Norlina KN	A	.06	14.28	14.34	56.49	- 6.11
Norlina KN-McKenney KS	B	—	15.23	15.23	60.01	+ 8.91
McKenney KS-McKenney	A	.11	16.87	16.98	66.88	+ .78
McKenney-McKenney KN	B	.08	15.43	15.51	61.09	+ 8.49
McKenney KN-Richmond KS	A	6.18	9.81	15.99	61.89	- .71
Richmond KS-Richmond	B	16.84	.07	16.91	63.60	+ .60
Richmond-Richmond KN	A	7.52	11.08	18.60	71.94	+ 5.94
Richmond KN-Fredericksburg KS	B	—	18.26	18.26	71.94	+ 5.94
Fredericksburg KS-Fredericksburg	A	.10	18.23	18.33	72.21	+ 6.11
Fredericksburg-Fredericksburg KN	B	.10	16.54	16.64	65.55	+ 2.95
Fredericksburg KN-Washington KS	A	—	18.45	18.45	72.69	+ 6.69
Washington KS-Washington	B	8.69	9.91	18.60	71.72	+ 4.72
Washington-Washington KN	D	18.90	—	18.90	71.06	+ 4.06
Washington KN-Baltimore	E	18.64	—	18.64	70.09	+ 3.09
Baltimore-Baltimore KN	D	16.89	—	16.89	63.51	- 3.49
Baltimore KN-Elkton KS	F	16.86	—	16.86	63.39	- 3.61
Elkton KS-Elkton	D	17.33	—	17.33	65.16	- 1.84
Elkton-Elkton KN	F	20.21	—	20.21	75.99	+ 8.99
Elkton KN-Philadelphia KS	D	13.51	—	13.51	50.80	-12.20
Philadelphia KS-Philadelphia	F	16.00	—	16.00	60.16	- 1.34
Philadelphia-Philadelphia KN	E	13.36	—	13.36	50.23	-12.77
Philadelphia KN-Princeton KS	G	13.93	—	13.93	52.38	- 9.34
Princeton KS-Princeton	E	13.49	—	13.49	50.72	- 9.67
Princeton-Princeton KN	F	16.69	—	16.69	62.75	- 2.75
Princeton KN-New York KS	E	14.50	—	14.50	54.52	- 5.38
New York KS-New York	F	15.38	—	15.38	57.88	- .69
Charlotte-New York Totals				627.70		+23.46 ³

¹ Attenuation figures used: 3.76 for U.G. at 73° and 3.94 for Aerial at 110°.

² For top channel, referred to - 9 db switchboard level.

³ Computed on a root-sum-square basis.

Figure 6 shows a comparison of the computed levels for maximum cable temperature conditions in the New York-Philadelphia Cable under conditions of both regulation and non-regulation. Omitting the regulation from a carrier repeater office where noise conditions and

gain requirements are favorable has the advantage of economy in saving the cost of the regulating apparatus and in an expected saving in maintenance.

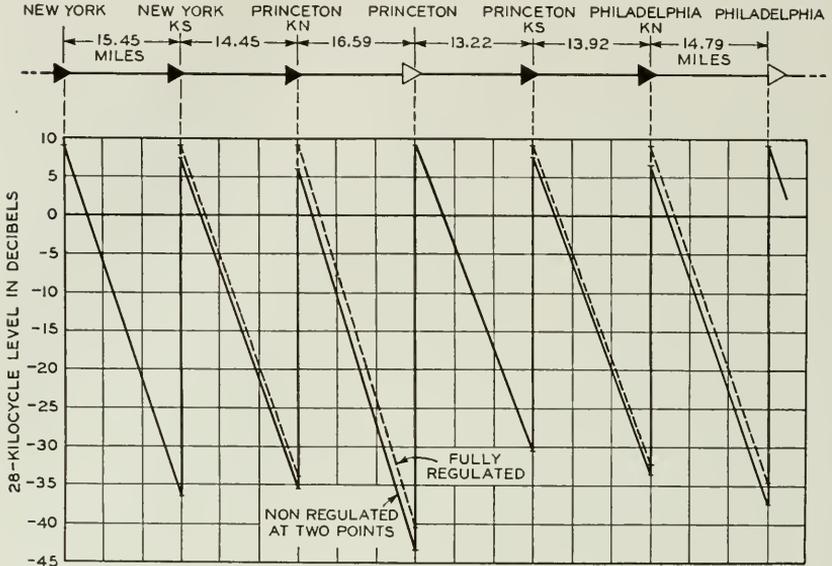


Fig. 6—Level diagram of New York-Philadelphia section, showing theoretical maximum effect on 28 kc levels of omitting regulation at the auxiliary repeater stations.

NOISE SUPPRESSION DEVICES

Two types of noise suppression devices were used on voice frequency circuits to limit noise currents which might enter the cables used for carrier.² The first of these is a retardation coil designed to suppress longitudinal currents but to have a negligible effect on the metallic currents on the side and phantom circuits. These were installed at voice frequency repeater points, in all voice quads in cables which contained carrier pairs connected to carrier repeater inputs. These coils attenuate longitudinal noise currents at carrier frequencies generated in the voice repeater office which might enter the cable over the pairs used for voice circuits and be induced into the carrier pairs. A total of 3,000 retardation coils was installed along the New York-Charlotte route for this purpose. The coils are connected into the cable conductors on the office side of the point at which the carrier lateral cable is connected to the main cable. These coils were installed in the office cable vault or manhole. Figure 7-A shows a number of

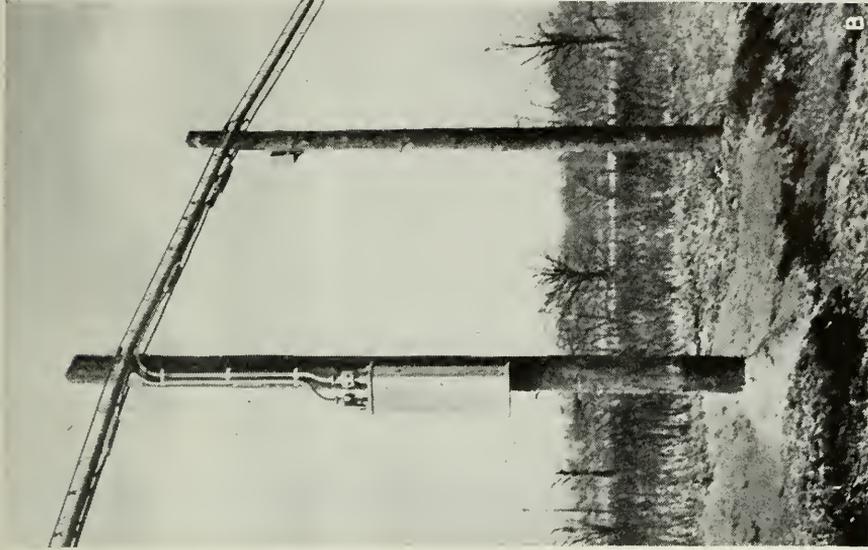


Fig. 7-B—Installation of filters on aerial cable.

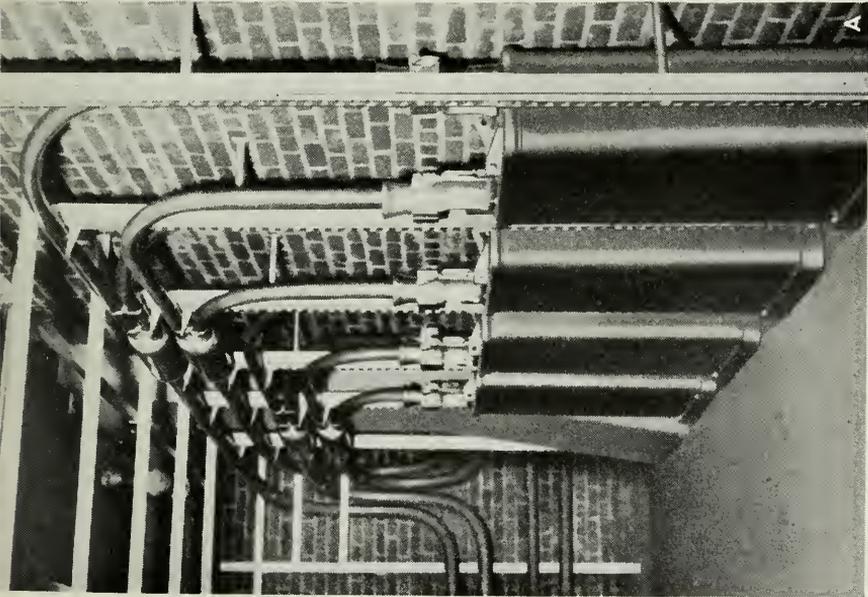


Fig. 7-A—Retardation coil cases installed in cable vault at West Unity, Ohio.

apparatus cases containing these retardation coils installed in the cable vault at West Unity, Ohio.

In addition to this usage, retardation coils were installed in certain instances in the conductors of branch cables and open wire taps. In other branch cables and open wire taps where a higher degree of noise current suppression was required the second noise suppression device was used.⁶ This second device is a filter which provides a considerably greater degree of suppression than the retardation coil. The purpose of these is to attenuate longitudinal noise currents which might enter the main cable over the conductors in the branch cable or open wire tap. A typical installation of filters on aerial cable is shown by Fig. 7-B. The question as to whether a retardation coil or a filter was required in each particular case was determined by computations of expected noise which might be contributed by the conductors entering the main cable. This was done by considering the makeup and length of the branch or tap, use to which it was put, and its location with respect to the nearest carrier repeater input in the cable to which it was connected. These computations, however, were made coincidentally with those described earlier in determining the most desirable directions of transmission.

Five hundred fifty-one retardation coils and 132 filters were installed in the 26 branch cables and open wire taps along the New York-Charlotte route, 11 of which connect directly to open wire. On the Detroit-South Bend project, 12 branch cables and open wire taps were equipped with 44 retardation coils and 124 filters. Nine of the 12 are taps connected directly to open wire.

As a further step toward prevention of noise currents in the carrier pairs, the shielding furnished by the lead sheaths of the cables has been kept effective by maintaining continuous the electrical path through these sheaths by means of shunts consisting of large condensers placed across each insulating joint.⁴

AUXILIARY REPEATER STATION BUILDINGS

Small buildings to house the auxiliary repeaters have been erected at the sites determined to be acceptable from transmission and construction standpoints. These structures are of fire resistive construction with concrete foundations, brick walls, and slate roofs. Since these buildings house equipment which is expected to operate for long periods of time without attention, no openings have been provided in the walls except for an entrance door and ventilating units. Thermal insulation has been provided over the ceiling. Two sizes of buildings have been used. The larger one which is 24 ft. × 24 ft., inside dimen-

sions, is used on routes where an ultimate of 100 systems is expected, while the smaller one, 21 ft. \times 24 ft., is used on a route to be developed for a maximum of 60 systems. The ceiling height in these buildings is sufficient to care for 11'-6" relay racks. Eleven of the small buildings and 22 of the larger ones have been built on the carrier projects so far completed.

The architectural treatment of the exterior of these buildings varies somewhat, depending upon the location of the site selected and the character of the buildings in the immediate neighborhood. The present designs may be classified as three types; i.e., plain brick with no trim, plain brick with limestone trim, and plain brick with limestone trim and artificial windows. In the latter type the window arrangements are obtained by the use of a wooden frame and sash with rough wire glass, backed by the interior brick wall. The brick portion behind the window is painted buff on the upper half facing the window, and black on the lower half, to simulate a true window with the shade half drawn. Typical examples of these types may be seen in Fig. 8. The type of building selected for each station depended upon the locality.

Arrangements have been provided in these buildings for automatically controlling the heating and ventilation by the use of thermostatically controlled electric heater and fan units. Although experience was generally lacking on the heating and ventilating problem for these stations, tentative requirements were set up. A minimum temperature of 40° F. has been considered satisfactory for the operation of the equipment in these stations and the thermostat has been set to turn on the heater unit if the inside temperature drops below that point.

Ventilating equipment consisting of intake and exhaust ventilators, exhaust fan and control equipment has been provided so that advantage may be taken of the effect of cooler outside air when the temperature inside the buildings rises to about 90° F. Consideration was given to the direction of the prevailing winds in locating these units in the building walls. The room side of the intake ventilator unit is equipped with a spun glass filter. These ventilators are equipped with rigid and movable louvers. The movable louvers are actuated by solenoids which are connected to the exhaust fan control which functions by means of a thermostat and a differential temperature control. The latter includes outside and inside temperature compensating elements. The thermostat is set at 90° which, with the differential feature of the control, will cause the louvers to open and the exhaust fan to start only when the inside temperature is more than 10° above that prevailing outside the building. When the inside temperature has been reduced to within 10° of that outside, the control circuit is opened to shut off the fan and close the louvers.



Fig. 8—Auxiliary repeater station buildings: A, without trim; B, with limestone trim; C, with limestone trim and artificial windows.

TERMINAL OFFICE EQUIPMENT

The various major items of equipment which are provided at a terminal office are shown schematically by Fig. 9. Two bays of sealed test terminals—one input and one output—are associated with the pair of cables which bring the carrier pairs into the office and are equipped initially to terminate 40 pairs. The input high-frequency jacks are mounted in high-frequency patching bays adjacent to the input sealed test terminal bays and the output jacks in bays adjacent to the output sealed test terminal bays. At points where more than 50 terminals are expected to be required at some future date plans have been made for one input and two output high-frequency patching bays. The input and output high-frequency patching and sealed test terminal bays for two cable routes, together with a high-frequency transmission measuring bay, are grouped so as to form a desirable arrangement for testing and maintenance purposes. This group of bays serves somewhat the same purpose as a primary testboard on voice frequency facilities. The arrangement of these bays as installed at New York is shown in Figs. 10-A and 10-B.

A portable transmission measuring set has been provided and may be placed on a writing shelf mounted in the high-frequency transmission measuring bay. This set may be connected to the various circuits by means of patching cords.

Line and twist amplifiers with associated flat and twist gain master controller equipment and crosstalk balancing bays also are installed at each terminal office. The arrangement of the amplifier equipment and controllers as installed at New York is shown in Fig. 11.

The terminal equipment for one system consists of six channel modem (modulator plus demodulator) panels, each of which mounts the sending and receiving apparatus for two channels.³ Channel modem equipment for three systems is mounted in two adjacent bays with the first two systems occupying separate bays. One bay of carrier supply equipment for each ten systems provides both regular and emergency units for generating carrier frequencies for the operation of the channel and group modem units and pilot channel equipment. The group modem units, one of which is required for each system, are mounted nine in one bay. Figures 12 and 13, respectively, show the method in which these bays are installed at New York.

The d-c power supply for the carrier equipment at terminal and main repeater offices is obtained from the existing 24-volt and 130-volt office power plants. Two sets of main distributing leads have been provided for each filament and plate power supply. Odd numbered circuits are

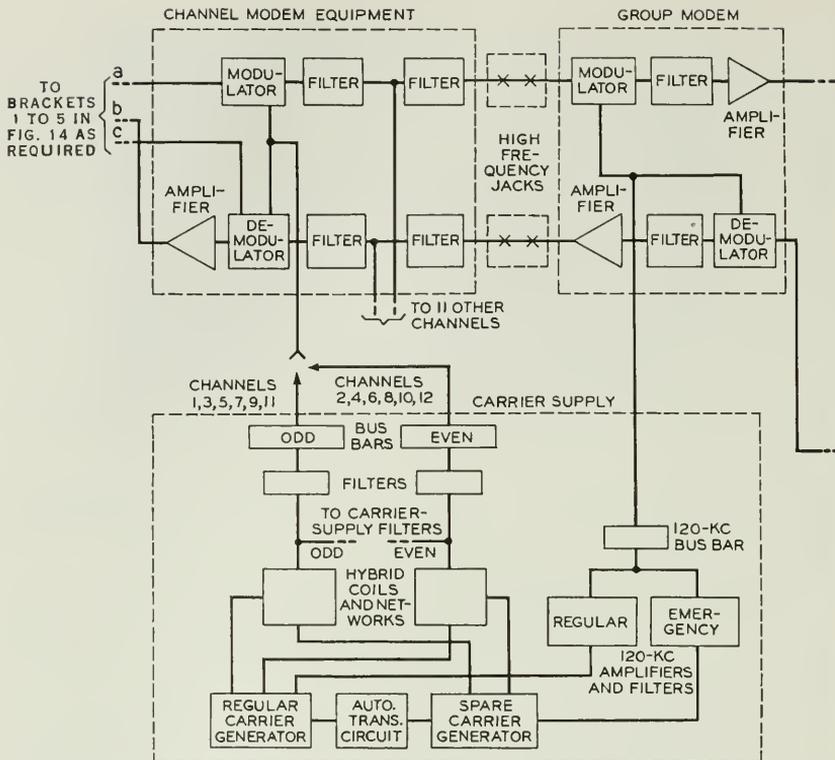


Fig. 9—Schematic showing the order of equipment at a terminal office.

connected to one set of leads for each type of power, and even circuits to another set.

The voice frequency sides of the channel modem units have been terminated in jacks which are located in a four-wire jack field mounted in a voice frequency patching bay. From this point the four-wire jack circuits are connected to the distributing frame for interconnection on a four-wire basis with other channel equipment, voice frequency repeater equipment, or terminating apparatus, as shown schematically in Fig. 14. Voice frequency patching bays shown in Fig. 10-B may be considered the equivalent of a secondary testboard.

Voice frequency transmission measuring apparatus has been mounted with the voice frequency patching bays.

MAIN REPEATER OFFICE EQUIPMENT

There are two types of installations at main repeater offices: (a) flat gain regulation only, and (b) both flat and twist gain regulation

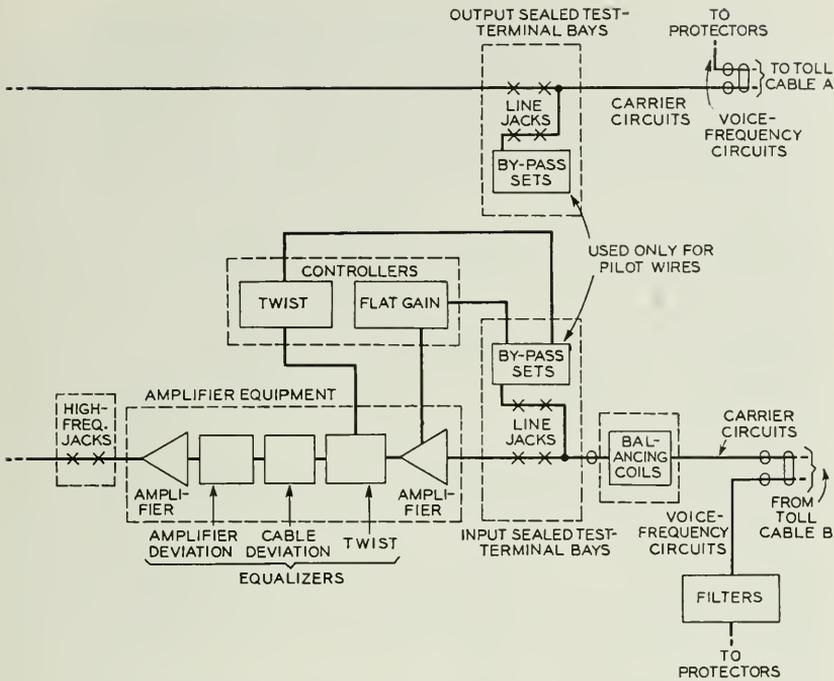


Fig. 9—Continued from page 570.

In general, these offices are attended regularly by maintenance forces. They serve in some cases as control or supervisory points for the auxiliary repeater stations. Since the installations at main stations have been made in existing voice frequency repeater offices, the available power plant is used to furnish filament and plate current for the amplifier equipment.

The input and output sealed test terminal bays with a high-frequency transmission measuring bay have been installed adjacent to each other to form a five-bay unit for testing and patching purposes.

Line amplifiers for each direction of transmission have been grouped together in adjacent relay rack bays. Each bay has a capacity of 20 amplifiers, except the first, in which is mounted the test amplifier associated with the high-frequency measuring system and 19 line amplifiers. Associated with the line amplifiers are the flat gain master controllers and power supply and cable balancing equipment. A schematic arrangement of circuits in a flat gain repeater office is shown in Fig. 15.

Repeater offices giving both flat and twist gain regulation have been

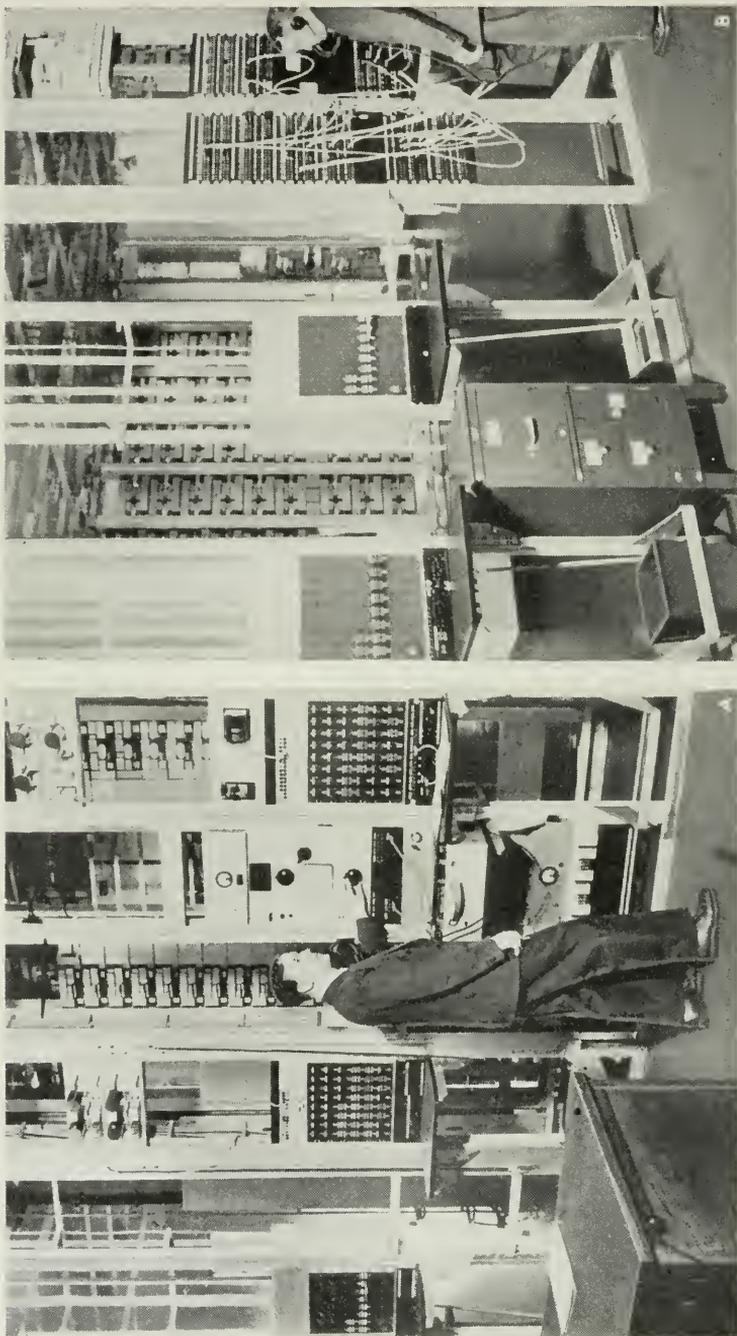


Fig. 10A—High frequency test bays at New York, N. Y.

Fig. 10B—High frequency and voice frequency patching bays at New York, N. Y.

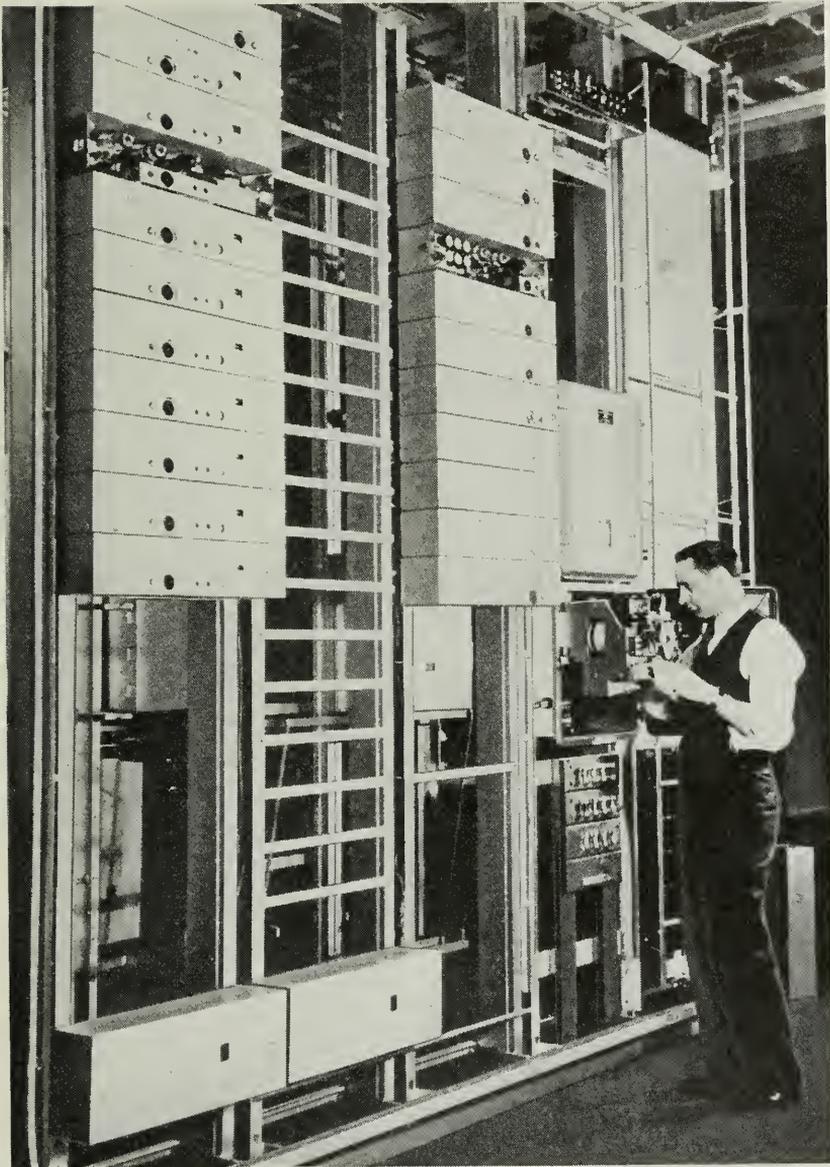


Fig. 11—Line and twist amplifier and controller equipment bays as installed at New York.

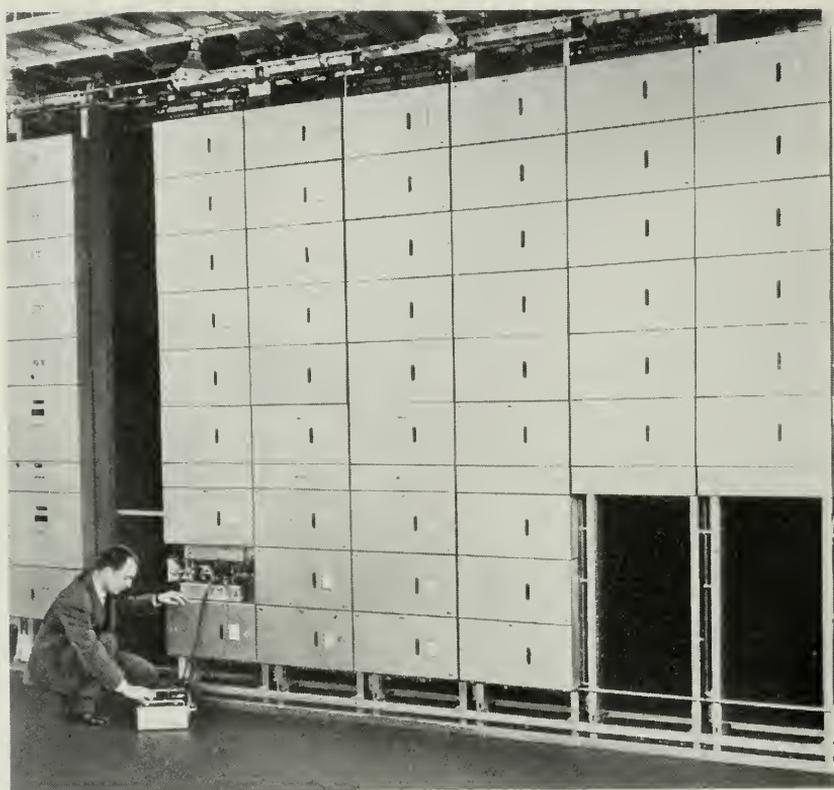


Fig. 12—Arrangement of carrier supply and channel modem equipment at New York, N. Y.

provided at intervals of about 100 miles. They differ from the flat gain regulating repeater office chiefly in that they include twist correction regulation¹ and its associated amplifiers. Provision was made at each twist gain regulating office for the installation of amplifier and cable deviation equalizers. The equalizers were actually connected to the circuits, however, only at such points as were indicated by lineup tests. To permit lineup without delay, spare equalizers were available at points where computations had indicated they might be needed. Amplifier deviation equalizers were required at South Bend, Toledo, Philadelphia, Richmond, and Greensboro on completed projects. It was not found necessary to install cable deviation equalizers at any point. Equalizer, flat gain, and twist gain amplifiers were installed in three-bay groups with the equalizer equipment occupying the center bay, and controller equipment in the fourth bay, similar to the ar-

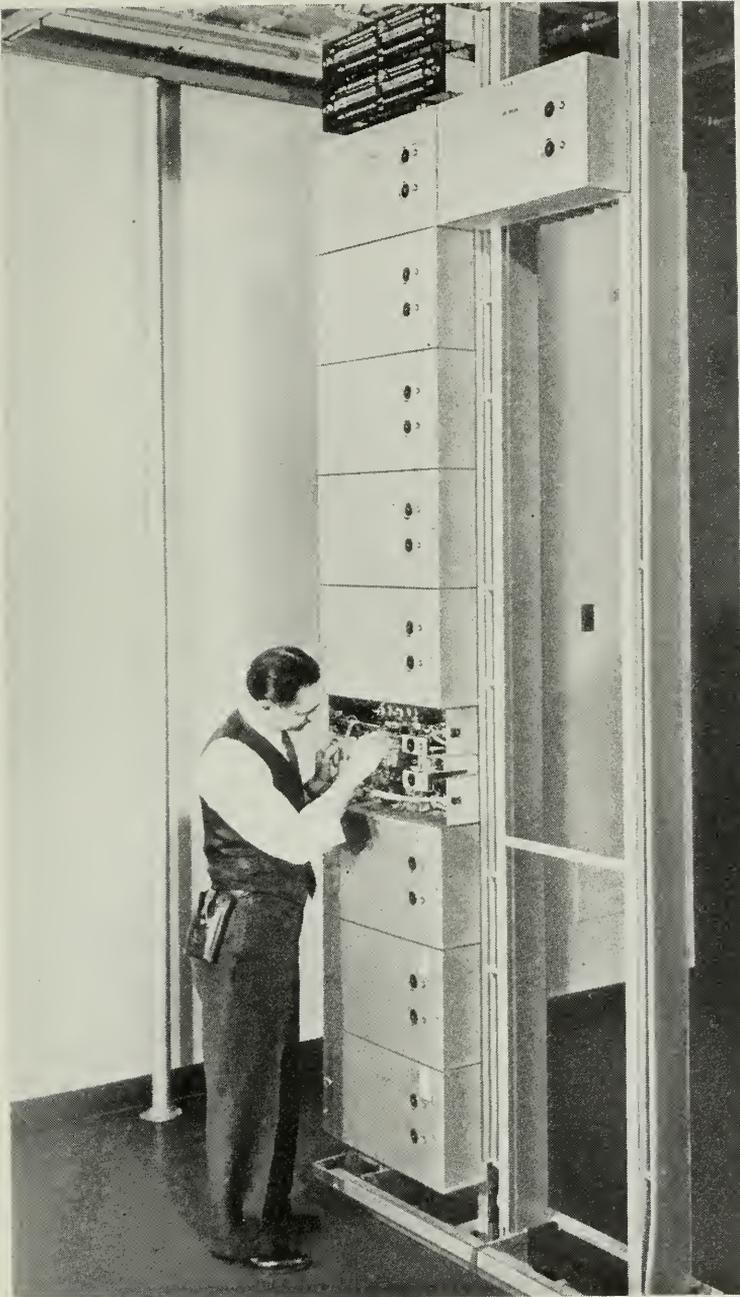


Fig. 13—Bay arrangement for group modem equipment as installed at New York.

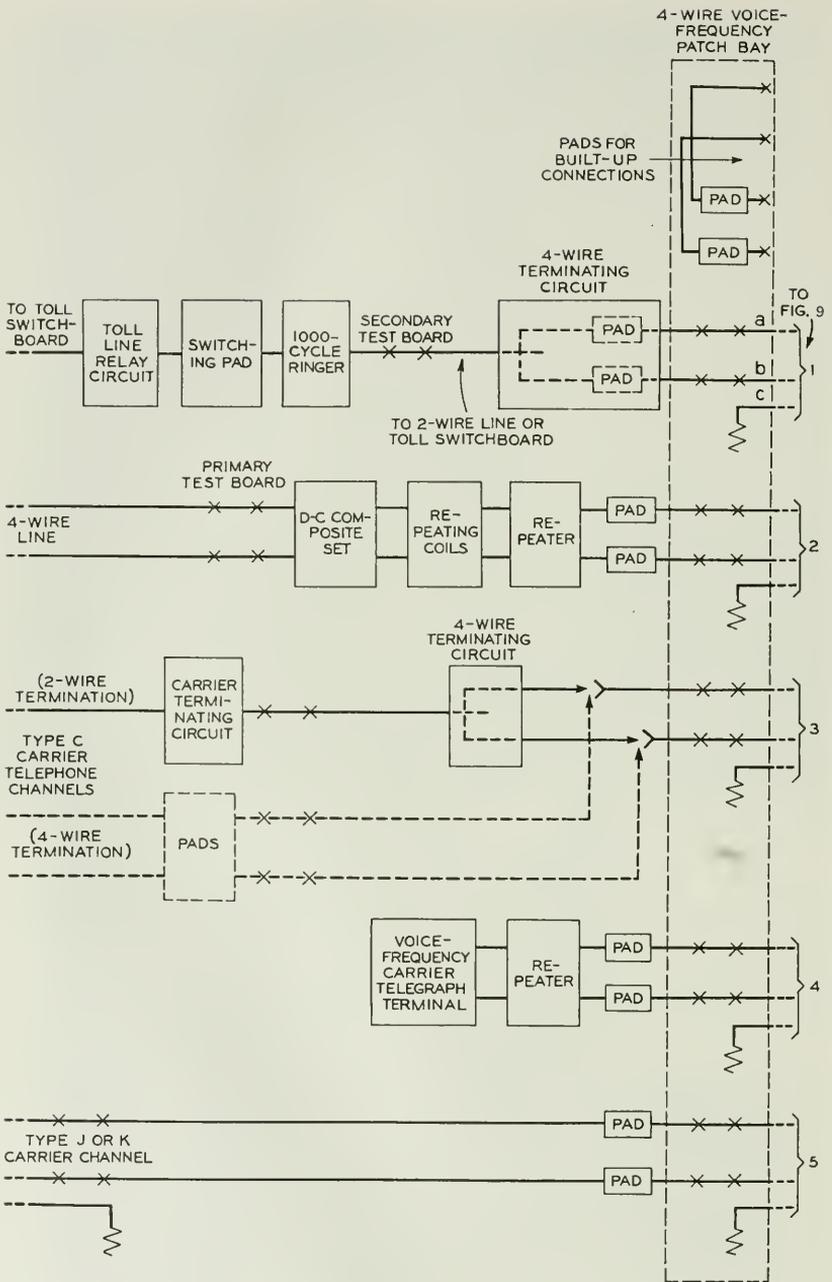


Fig. 14—Schematic showing the various circuit arrangements on the office side voice frequency patching bay.

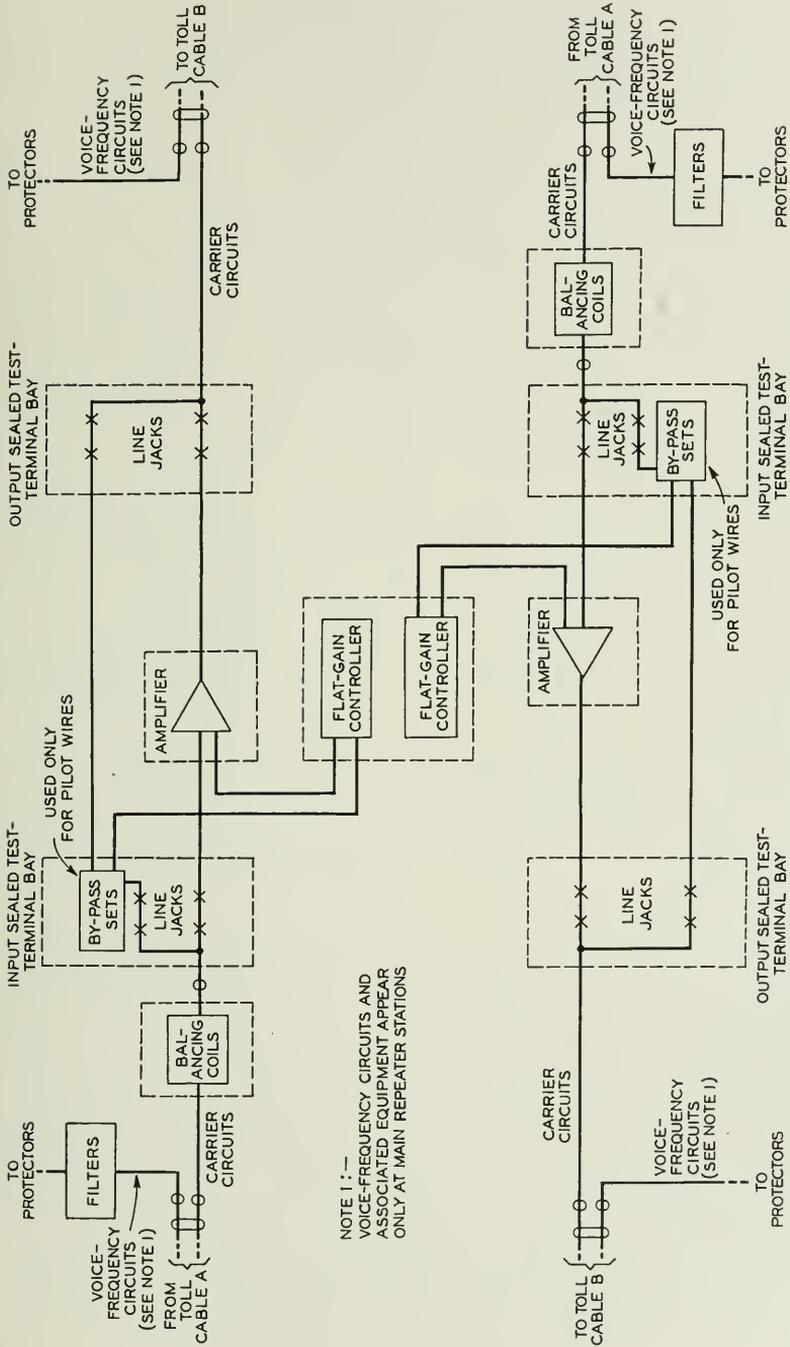


Fig. 15—Schematic showing the order of equipment at a flat gain repeater office.

rangement shown in Fig. 11. Each group of three bays mounts equipment for one direction of transmission for 17 systems. A schematic arrangement of equipment in a twist and flat gain regulating repeater office is shown in Fig. 16.

AUXILIARY REPEATER STATION EQUIPMENT

Each auxiliary repeater station houses crosstalk balancing equipment, sealed test terminals, line amplifiers, pilot wire regulators, and a power plant. A typical floor plan arrangement of the equipment required in one of these stations for a maximum of 100 systems is shown in Fig. 17. The equipment arrangement for a 60-system route is practically the same, except that provision has been made for a smaller number of amplifiers and crosstalk balancing bays. A schematic arrangement of equipment circuits is shown in Fig. 15.

Four bays of sealed test terminals have been installed, one input and one output for each direction of transmission. Initially each unit contains carrier line and equipment jacks for testing or patching purposes for 40 carrier and eight miscellaneous circuits. In addition, miscellaneous auxiliary equipment is mounted in these bays.

Twenty-one amplifier panels for one direction of transmission may be mounted in a bay. Two bays are required for the flat gain master controllers and the associated controller power supply equipment. Figure 18-B shows amplifier, controller, and testing bays.

High-frequency testing apparatus consisting of a variable test oscillator and a portable transmission measuring set mounted in a mobile relay rack bay has been provided at each auxiliary station. This unit may be connected to the jacks in the sealed test terminals as required by means of patch cords.

Since auxiliary stations are designed to operate for considerable periods of time without attention, the power plant is of the automatic type.⁷ It consists of a 70-cell, 152-volt storage battery which is continuously floated across regulated tube rectifiers fed from a commercial power supply. Figure 18-A shows a typical installation. Two rectifiers are provided initially, one which floats the battery, and the other which is connected automatically into the charging circuit in case of failure of the first unit or to increase the charging rate after a prolonged failure of the outside power. Arrangements are available in the power service cabinet to terminate leads from a portable emergency engine driven alternator set which may be set up outside the building. It is expected that the battery installed initially will be of sufficient size to provide a minimum of 24 hours reserve throughout its life, taking

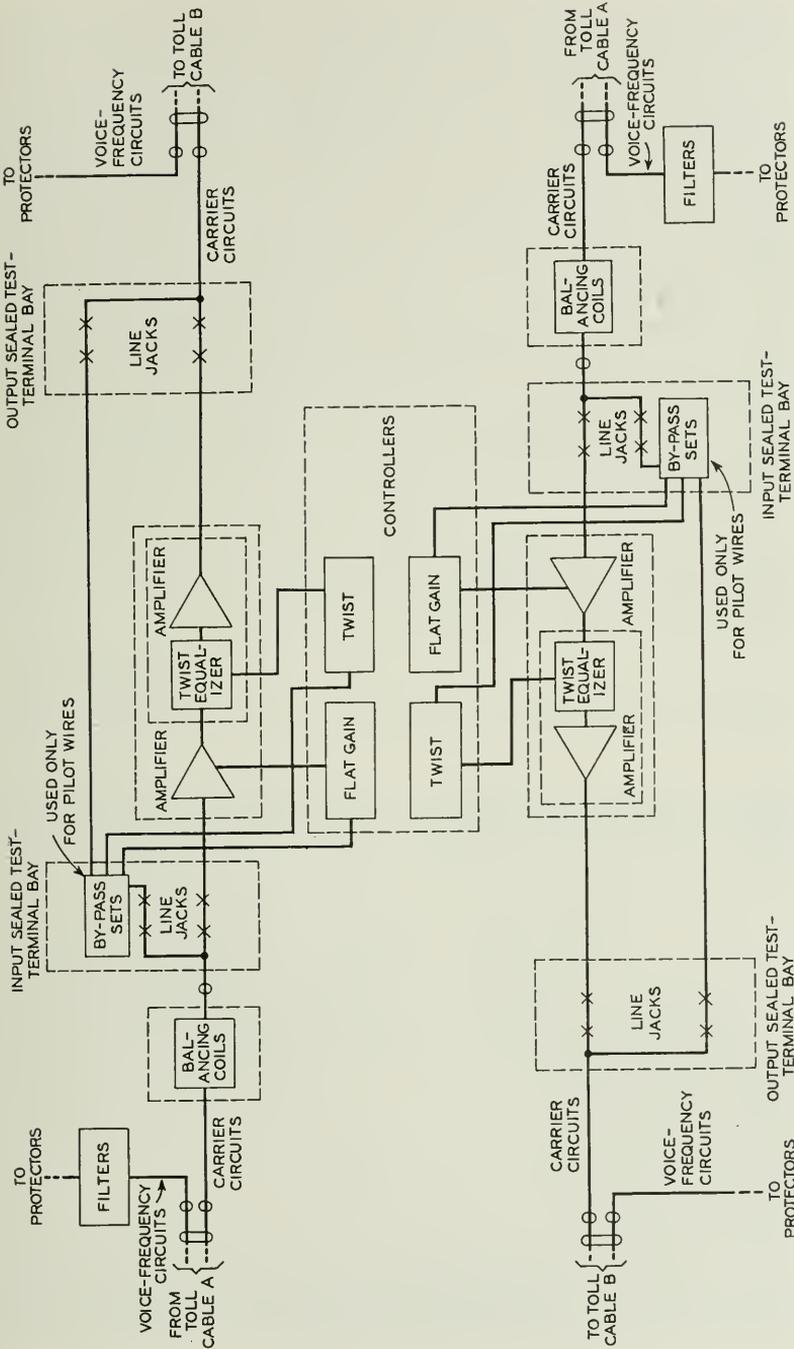


Fig. 16—Schematic showing the order of equipment at a twist and flat gain office.

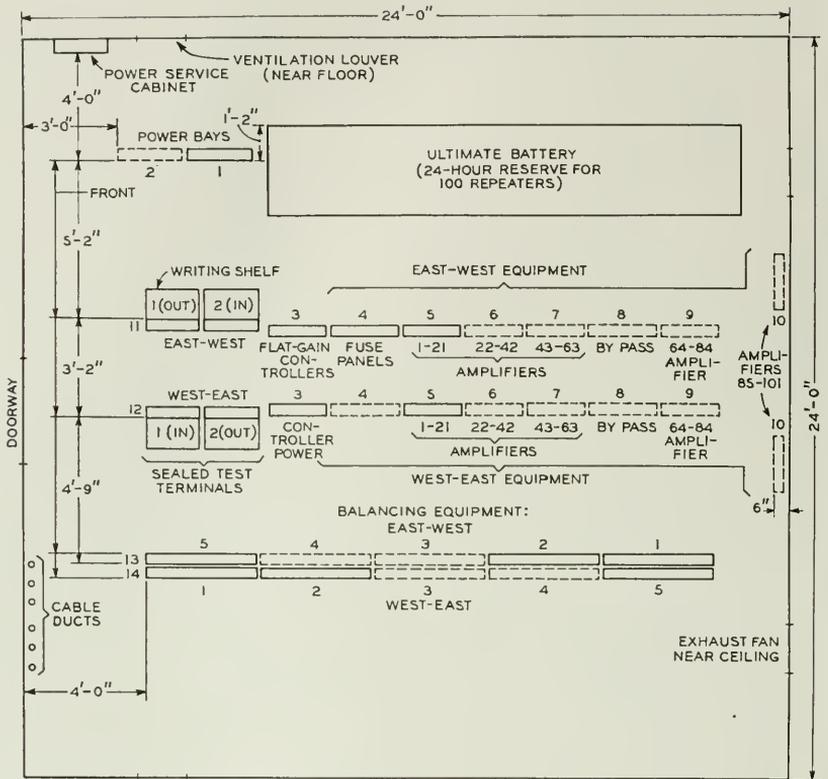


Fig. 17—Floor plan arrangement of equipment for 100 systems at an auxiliary repeater station.

into account the estimated growth of carrier amplifier equipment requirements for that period.

The entire voltage of the battery is used to furnish plate supply for the amplifier tubes. The 70-cell battery is arranged in seven groups of 10 cells each and taps are taken from each group to supply current for the heaters of the tubes of each amplifier. To prevent an uneven drain, amplifiers are connected across the battery in multiples of seven. In case the number of amplifiers installed is not an even multiple of seven, dummy load resistances are connected as required, in lieu of amplifiers to fill out the unequipped multiple.

The controller power supply bay contains apparatus for the 140-volt d-c pilot wire bridge supply and 55-volt, 60-cycle a-c supply. An emergency rotary converter, which automatically provides 110-volt, 60-cycle a-c supply during outside power failure, and operates intermittently from the battery, also is mounted in this bay.

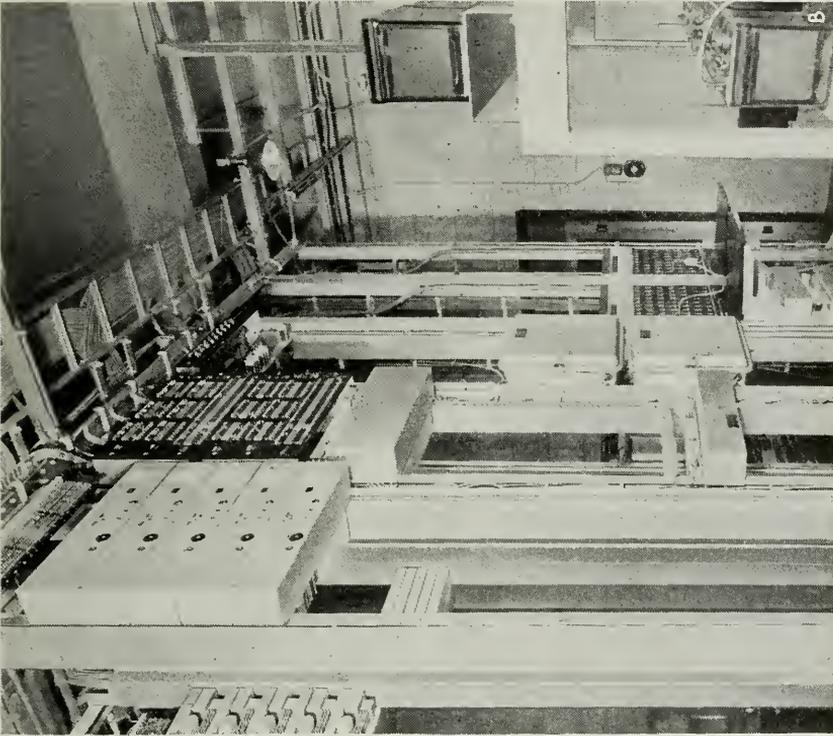


Fig. 18B—Amplifier, controller, and testboard equipment arrangement at an auxiliary repeater station.

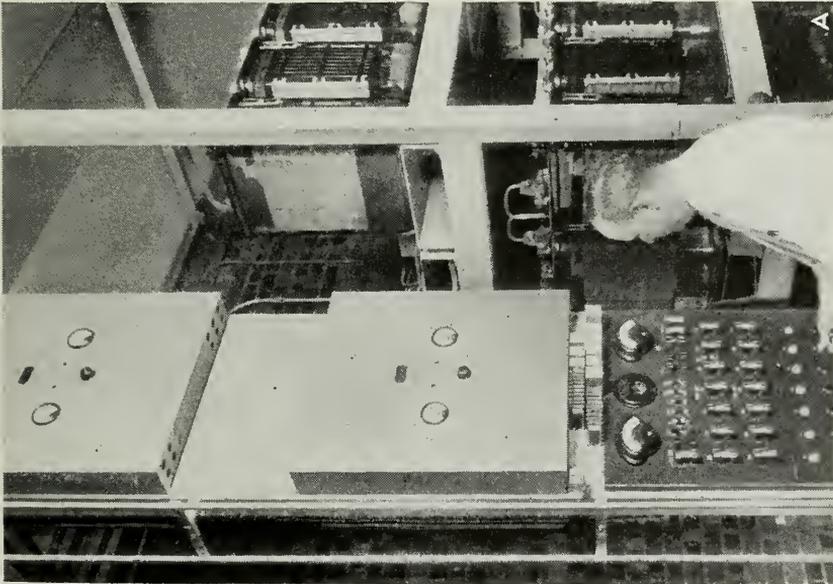


Fig. 18A—Auxiliary repeater station power plant.

CABLING PROBLEMS AND FLOOR PLAN LAYOUTS

The floor plan arrangements for type K carrier equipment have been controlled to a considerable extent by transmission requirements, with consideration also being given to satisfactory operating and maintenance layouts both for the initial installation and for the future. The first consideration of any space which is to be used for carrier equipment is that the various units of equipment can be so located, with respect to each other, that the established maximum wiring lengths, as determined by transmission, operating, and economic requirements, will not be exceeded. For example, the length of shielded pair cable between the jacks in the input sealed test terminal bay and the input side of the line amplifier has been limited to about 50 feet and the potentiometer lead between the voice frequency patching bay and the channel modem units has been kept short in order not to limit the adjustment range of the potentiometer and a limit of 150 feet has been set. These limits were set in the design of the type K systems. Two types of cabling were installed in the transmission part of the carrier circuit; i.e., standard lead covered cable and shielded pair cable.

In cabling the carrier equipment the input leads were not run on the same cable racks with voice frequency cables. Due to the difference in transmission level between cabling connected to the input sides of carrier amplifiers and that connected to the output sides, these two groups of cabling have been segregated by running them over separate cable racks. The input leads were spaced not closer than two feet to any leads carrying interrupted direct current or power supply leads which might possibly carry high-frequency noise currents. Output leads were run on the same cable racks with voice frequency cabling where necessary, but were kept six or more inches away from possible disturbing leads such as those just mentioned. Cabling from the voice frequency side of the channel modem equipment was installed without greater precautions than are used when installing other voice frequency cabling. Crosstalk balancing bays were installed in any convenient location without special limitations in the lengths of lead covered cables between these and the input sealed test terminal bays.

All cable racks carrying the rubber covered shielded cables from the amplifier and group modem bays to the sealed test terminal and high-frequency patching bays were arranged so that these leads were run loosely and without sewing. This arrangement provides a ready means for switching cables for circuit layout purposes, particularly at terminal offices or at junctions of carrier cables.

In existing offices the high-frequency jack and testing equipment has been located as close as practicable to the existing toll testboard posi-

tions. Separate testboard lines have been established in several offices opening off, or convenient to, the main operating aisle in front of the toll testboards. The voice frequency patching jack bays at carrier terminals have been located, where practicable, near the secondary toll testboard equipment. These arrangements have been made in order to facilitate operating and maintenance, particularly during light load periods when a small force is on duty.

The amplifiers and group modems have been closely associated with the high-frequency patching bays and sealed test terminals in order to limit cabling lengths. Channel modems and carrier supply have been located convenient to the other equipment but within wiring limitations to the voice frequency patching bays.

The adequacy of all floor plan layouts, in providing for ultimate requirements, particularly at large terminal offices, was studied. This problem was given special consideration where it is expected that routes in addition to the initial one may be developed later for carrier operation. For example; at New York it was necessary to plan for the development of K carrier and other broad band facilities on four separate routes requiring a considerable amount of space for the necessary terminal equipment. The installation of carrier equipment at this office, therefore, has been made in space separate from the existing voice equipment. A floor plan arrangement of the equipment layout at New York is shown in Fig. 19.

ORDER WIRES AND ALARM CIRCUITS

One or more auxiliary stations have been associated with an adjacent main or terminal office for maintenance control. Interoffice trunk and alarm equipment provide talking and signaling facilities between each auxiliary and main repeater station over a loaded cable pair. The various alarm signals are terminated in lamps which are mounted in the sealed test terminal bay at the controlling main office. These alarms are arranged to indicate such happenings as fire, open door, high-low battery voltage, main discharge fuse operation, a-c power failures, etc. When an alarm signal is received at the main station, it is rechecked and upon its reappearance an attendant may be dispatched at once or later to the auxiliary station involved, depending upon whether the signal is of major or minor importance.

The auxiliary stations are not always controlled by the nearest attended station. For example, in several cases it has been thought better to have them controlled by a station located in a small town rather than a city, because in cases of necessity an attendant should be able to drive to the auxiliary station in less time than required to drive

through a city area. In other cases certain main stations are not manned 24 hours per day and control and alarm circuits have not been terminated at such points. In one case a main repeater station has terminated in it the control leads from eight auxiliary stations. Four of these are connected through the partially attended main stations on either side.

COMPLETION TESTS AND OVERALL SYSTEM ADJUSTMENTS

The usual completion tests were made on each unit of equipment after it was installed and on each cable pair between repeater stations after it was unloaded, in order to insure readiness of each item to be connected to form the overall carrier system. The gains of the repeaters were given a final adjustment by connecting each repeater input to the cable pair with which it was installed to work, then sending a predetermined amount of power at 28 kc into the cable pair at the adjacent office and adjusting the gain of the repeater until it delivered the desired output level. The flat gain master regulator was adjusted with respect to its pilot wire so that it would adjust the gain of the amplifier to maintain the desired output level at 28 kc at all cable temperatures.

The repeater sections beginning at one end of each twist regulating section were measured progressively at the output of each repeater. In each case transmission was checked at ten frequencies throughout the range from 12 to 60 kc. In this way a check was obtained to determine whether proper equalization was being provided at each line amplifier. It was necessary in some cases to change the type of equalizer provided because the transmission characteristics of specific cable pairs differed from the average which had been assumed in providing equalizers. Measurements were also made on the overall twist regulating section and transmission checked at ten frequencies throughout the 12 to 60 kc range, since the output of a perfectly corrected twist section would be the same at all frequencies within this band.

Overall measurements similar to these were made on the high-frequency line between terminal points and Fig. 20 shows the results for typical New York-Charlotte and New York-Washington systems. The most desirable characteristic would be a straight line and it will be noted that this curve differs materially from such an ideal. This difference is due to inadequate compensation by means of equalizers for small deviations from linearity in the individual line amplifiers. This lack of linearity in the overall high-frequency line has not materially affected the systems being operated at present but might be-

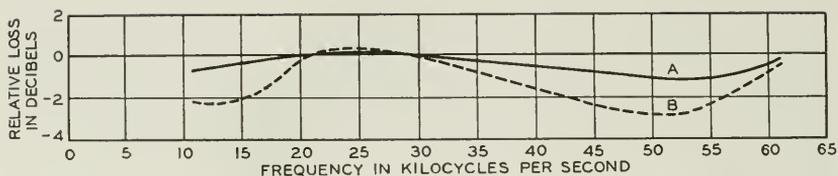


Fig. 20—Typical overall transmission frequency characteristic of high frequency line between New York and Charlotte, N. C. (B), and New York-Washington (A).

come objectionable on future long systems and it is planned to improve this characteristic by means of different equalizers.

The overall transmission frequency characteristic of a channel on a New York-Charlotte type K system is shown by Fig. 21. Measure-

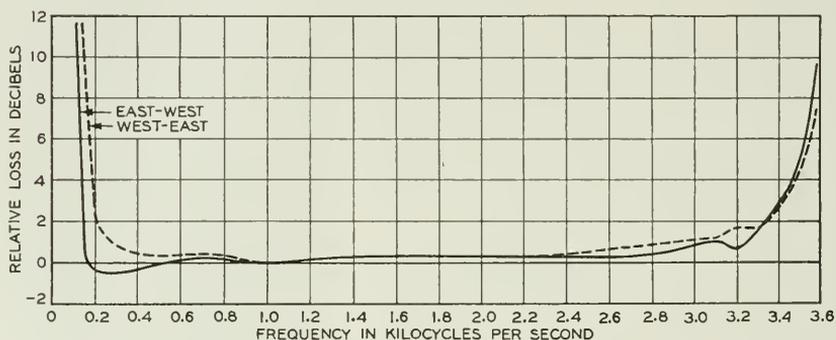


Fig. 21—Overall transmission frequency characteristic of a type K carrier channel between New York and Charlotte, N. C., as measured between two-wire voice frequency lines.

ments for this characteristic were made between the two-wire sides of hybrid coils connected to the two directions of carrier transmission of the channel concerned.

USE OF INITIAL SYSTEMS

Telephone message circuits are being operated over most of the type K channels now available for use. In most cases the channels are used as parts of circuits which are longer than the carrier systems. For example, most of the 60 channels between New York and Charlotte, N. C., are used for circuits between New York and southern cities beyond Charlotte. Some of these circuits are obtained by connecting type K carrier channels at Charlotte to channels of type J open wire carrier systems which operate between Charlotte and West Palm Beach, Fla. Of the 204 channels available for use, only 21 are used as all

carrier message circuits between the system terminals. This is not necessarily typical of what the usage of type K channels will be, but is brought about by their relatively limited application to date and demonstrates the flexibility of the type K carrier circuit in fitting in with the other types of circuit facilities.

CONCLUSIONS*

Although experience with type K systems in service is rather limited, they are providing circuits of excellent quality and performance. The band width of the individual channels slightly exceeds the original estimates. Such instabilities of transmission as have been experienced have, in most cases, been corrected before service was interrupted and have been caused largely by non-recurring troubles inherent to most installations of new equipment. In brief, the operating experience with type K cable carrier systems confirms the view that they are expected to become an important means of providing additional long distance telephone circuits over cable facilities.

REFERENCES

1. C. W. Green and E. I. Green, "A Carrier Telephone System for Toll Cables," *Bell System Technical Journal*, Vol. 17, No. 1, January, 1938.
2. M. A. Weaver, R. S. Tucker and P. S. Darnell, "Crosstalk and Noise Features of Cable Carrier Telephone System," *Bell System Technical Journal*, Vol. 17, No. 1, January, 1938.
3. R. W. Chesnut, L. M. Ilgenfritz and A. Kenner, "Cable Carrier Telephone Terminals," *Bell System Technical Journal*, Vol. 17, No. 1, January, 1938.
4. A. J. Aikens, "Suppressing Noise and Crosstalk on the Type K Carrier System," *Bell Laboratories Record*, Vol. 17, No. 7, March, 1939.
5. F. W. Amberg, "Crosstalk Poling for Cable Carrier System," *Bell Laboratories Record*, Vol. 17, No. 6, February, 1939.
6. P. W. Rounds, "A Longitudinal Noise Filter for the Type K Carrier System," *Bell Laboratories Record*, Vol. 17, No. 9, May, 1939.
7. H. H. Spencer, "Power Plant for Broadband Repeater Stations," *Bell Laboratories Record*, Vol. 17, No. 8, April, 1939.
8. L. Hochgraf, "Crosstalk Balancing for the Type K Carrier System," *Bell Laboratories Record*, Vol. 17, No. 6, February, 1939.

The Toronto-Barrie Toll Cable *

By M. J. AYKROYD and D. G. GEIGER

GENERAL

DURING 1937 a 60-mile toll cable was completed between Toronto and Barrie which, in several respects, is unique. Among the interesting features in the design and construction of this toll cable were the use of non-quaddled exchange cable and loading, a 60-mile repeater spacing, planning for future carrier operation, and extended pole spacings.

Prior to the installation of this toll cable, the territory to the north and northwest of Toronto was served by three open-wire pole lines. Figure 1 shows these lines and the territory served by them. The Toronto-Owen Sound lead entering Toronto through a 7-mile entrance cable was poorly located in towns and on highways, and was paralleled by power lines which caused considerable noise on the longer circuits. The Toronto-Collingwood and Toronto-Barrie lines, which were common for some distance north of Toronto on a 6- and 5-arm lead, entered Toronto through an 11-mile entrance cable which had been in place about four years, and contained a number of spare conductors due to its having been designed for two additional lines.

It was realized that, if open wire were to be continued, circuit growth would require a new line arranged for carrier operation and a general rebuilding, rerouting and retransposing for carrier operation of the existing lines. In addition, carrier operation would necessitate expensive carrier loading of the entrance cables at Toronto, and the length of these entrance cables would limit the length of the carrier circuits for operation without intermediate repeater stations.

STUDIES PRIOR TO CONSTRUCTION

With large expenditures foreseen for the continuance of open wire, it was only natural that a study should be made of the possibility of the use of a toll cable on a basic route and the use of as much as possible of the existing lines as feeders to the cable. Cost studies on an annual charge basis for a twenty-year period of open wire with superimposed 3-channel carrier systems, and for a 2-wire 19-gauge quaddled cable

* The unusual solution of a difficult toll cable problem which is described in this paper will be of interest because of its novelty rather than because of any expected general application of this type of construction to toll cable routes.

with H88-50 loading with open wire or cable feeders, depending on the length and numbers of feeder circuits, indicated the cable plan to be best. In addition to the indicated money savings of the cable plan over the period of the cost studies, other indicated advantages in the toll cable plan were improved service continuity (the southerly section of the territory under study is one of heavy sleet conditions) and reduced noise from power induction.

The quadded cable plan, however, had one disadvantage in that it required a repeater station approximately 45 miles north of Toronto, in a territory remote from any town or village with unfavorable living conditions and subject to isolation during winter snow storms. The nearest feasible location to the ideal, at Cookstown, involved such an increase in length of cable and added expenditure that the cost advantage changed to the open-wire plan. Also, the use of B88-50 loading with a repeater spacing of 50 miles appeared to offer no advantage in that the additional cost of loading became an important factor.

These difficulties in the use of the standardized type of toll cable led to a review of the possibility of employing some combination of conductor and loading which would permit a 60-mile repeater spacing, thus eliminating any need for an intermediate repeater station between Toronto and Barrie. If such a cable were to have the same unit attenuation as 19-gauge H88-50 cable, then it must have considerably improved crosstalk and return loss characteristics. On the other hand, if a cable could be obtained with crosstalk and return loss characteristics about equal to that of 19H88-50 cable, it must have an attenuation of about $\frac{3}{4}$ that of 19H88-50 cable.

Of the standard types of cable and loading, 19-gauge non-quadded exchange cable having a capacity of about 0.083 mf. per mile with B-135 loading appeared to have an attenuation of about the value required to meet the second of the two requirements noted above. It was estimated that such a cable would have the following transmission characteristics:

1000-cycle attenuation at 68° F.	0.26 db per mile
Passive singing point at repeater exceeded by 72% of circuits	25 db
Maximum crosstalk gain	14 db
Overall active balance ¹	6.0 db
Overall circuit loss 8 db (PO-TC) with 4 db pad at PO.	

These assumed limits required a 72 per cent return loss of 26 db or better at the critical frequency which was expected to be about 2600 cycles, and a 1 per cent maximum near-end crosstalk of 74.5 db. Based

¹ Computed by summation of the 72 per cent singing points at individual repeaters with a 5 db end path.

on these values and limits, and assuming that Toronto would be the only gain switching center directly involved, a study was made of the transmission possibilities for each group of circuits that was expected to be routed through the cable. This study indicated that, provided the return loss and crosstalk values required of the cable by the assumed singing points and crosstalk gain could be met, all circuits could be 2-wire between Toronto and Barrie with some transmission margin and also that this type of cable could be extended at least another 20 miles to Orillia.

A cost study, assuming a 101-pair cable, of this type, indicated that while, due to the additional loading costs of the closer loading spacings, the cable costs were very nearly the same as for a quadded 19-gauge H88-50 cable, the considerably reduced repeater and repeater station costs made this plan appreciably less costly than any open wire plan. The elimination of any intermediate repeater station removed the repeater station difficulties of the quadded cable plans.

As no installation of such a length of this type of cable had been made, some confirmation of the estimated values for the transmission study, and particularly of the return loss and crosstalk, was considered necessary. An 8-mile H-44 loaded 19-gauge exchange cable, which had just been erected near Toronto, was chosen for study. Near-end crosstalk measured on 286 combinations of pairs indicated 99 per cent better than 81 db with an average of 91.7 db which, when modified for impedance and length differences, indicated 99 per cent better than 72.5 db, and an average of 83.2 db for the proposed cable. While these values were somewhat poorer than required, the size of the sample and one or two other factors indicated that the proposed cable could be erected to meet the crosstalk requirements. However, to obtain as much crosstalk margin as possible, arrangements were made for the manufacturer to use 6 lengths of twist, alternating 3 in each layer, rather than the 4 twists which had previously been used for this type of cable. It is felt that the excellent crosstalk results obtained as outlined in more detail later are in large part due to this feature.

Singing measurements on 10 pairs averaged 19.6 db. It was evident from impedance frequency measurements that these singing points could be raised to the desired value of 25 db by some modification in the networks. Accordingly an adjustable precision type network was developed.

Also, four 1500-foot lengths of the proposed type of cable were obtained from the manufacturer and tested for mutual capacitance of pairs and capacitance unbalance between pairs. On statistical analysis, these tests indicated a probable average near-end crosstalk of 79 db

and that for return loss 63 per cent of the circuits would be better than 27.0 db or 72 per cent would be better than 26 db at 2600 cycles, provided the following features were incorporated:

- (a) Manufacture of complete length of cable in one continuous production with reasonably careful control of variables.
- (b) Capacity equalization splicing at the mid-point of each 3000-foot loading section.
- (c) Reel lengths be assigned as to location on the basis of average reel length capacity.

On the basis of these preliminary studies, it was decided to proceed with the cable plan, using the B-135 standard 19-gauge exchange cable. Figure 2 shows the plant layout finally adopted for the cable and its feeders.

At the Toronto end it was essential to use pairs in a recently placed 19- and 16-gauge quadded toll entrance cable (mutual capacity .062 mf. per mile) about eleven miles long in order to keep the cost of the cable to a minimum. This appeared feasible, using the same type loading coils as in the main cable, if the loading spacings were extended to provide the same loading section capacity as in the main non-quadded cable, and if the cable were sufficiently well respliced to break up the side-to-side (within-quad) adjacencies so that the crosstalk coupling would be comparable to that obtained in the main (non-quadded) cable.

ROUTE

It was necessary to select the shortest practicable route passing as close as possible to the places to be served (see Fig. 2). The route selected is, for the most part, on a road which lies about midway between the main highways serving the territory north of Toronto. It is expected that the location chosen will be reasonably free from highway changes. Also, for the portion of the route south of Aurora, an existing open wire pole line was suitable for supporting the cable on long span construction.

At one point three miles of swamp covered with bush intervened on the direct route, the avoidance of which meant an increase in expenditure for right-of-way, as well as lengthening of the cable. It was decided to go straight through the swamp, using swamp fixtures, as shown in Fig. 3. An interesting sidelight on securing the route through the swamp was the fact that an original road right-of-way was shown on the map. On searching the records the surveyor found the original survey notes made in 1860 and eventually confirmed the location by finding some old pottery which, according to the records, had been

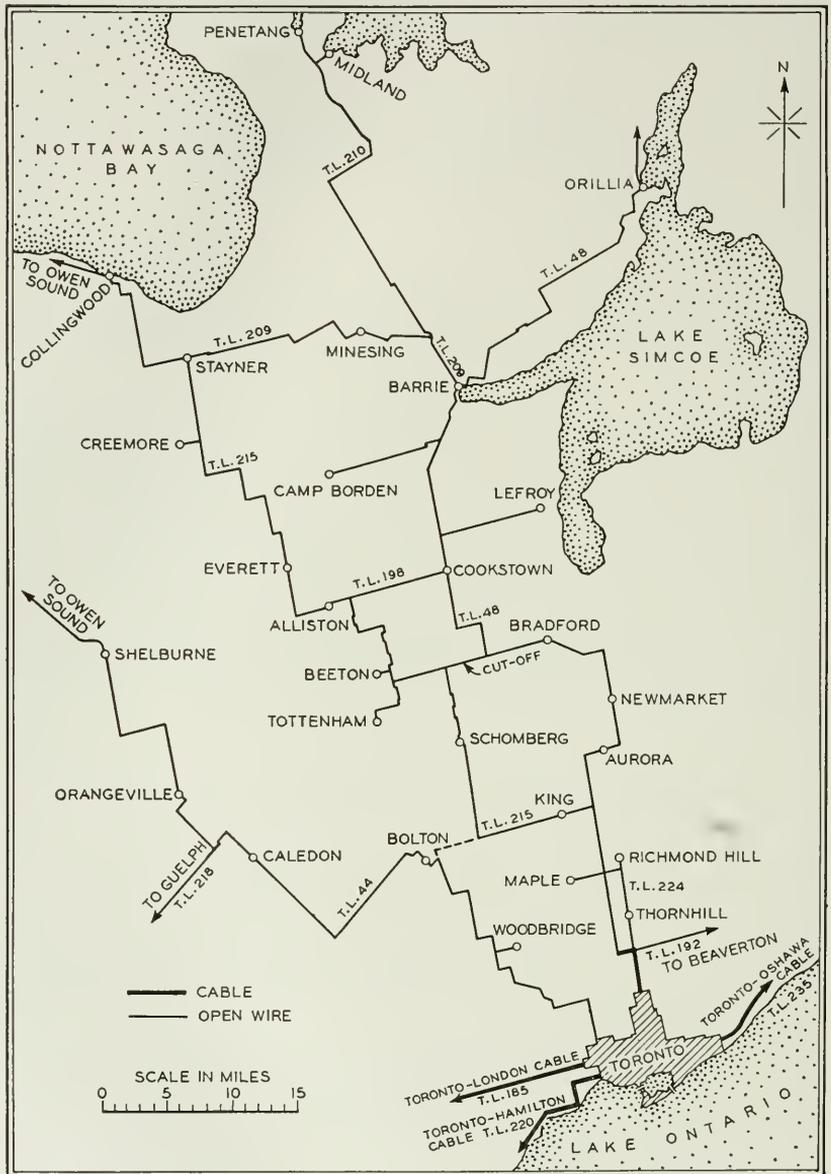


Fig. 1—Toll lines before construction of the Toronto-Barrie cable.

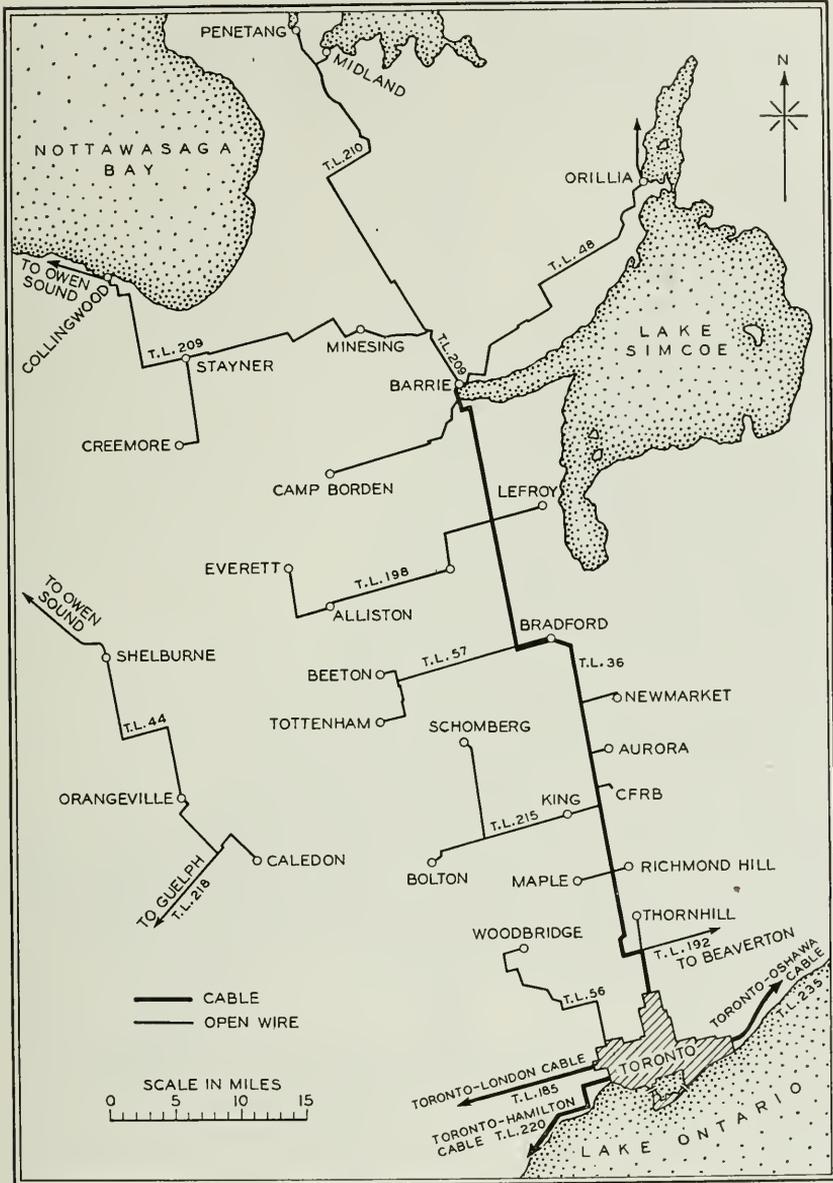


Fig. 2—Toronto-Barrie cable.



Fig. 3—Swamp construction.

buried at the road intersections. During the late summer and the autumn, when the swamp had dried out, a 60-foot right-of-way was cleared. As the soil was still moist and soft, the brush and small trees were uprooted by a tractor, a method of clearing which proved quick and economical. The swamp fixtures were placed before the ground froze, and the cable during the winter.

POLE LINE

The cable was erected on the existing pole line between Toronto and Aurora. The size of the cable erected on 10 M. strand permitted the removal of every other pole in the old line, with a resultant 185-foot average spacing. As is shown in Fig. 4, where it was necessary to change an existing pole, a new pole was placed and fastened to the old pole by means of stub reinforcing bands, thus eliminating the expense of transferring the open wire. Upon the release of the open wire by transfer of circuits to the cable, the wire and old poles were removed.

The new section of pole line was erected on a 200-foot spacing, with occasional spans up to 250 feet, as shown in Fig. 5. This increased pole spacing was also expected to reduce ring cutting and bowing.

CONSTRUCTION DETAILS AND TESTS

At a number of points open wire loops connected directly to the cable. At these junctions there were installed open space protectors having a lower breakdown than the cable pairs, and connected between the open wires and the cable sheath; also a few spans from the junction, 1000-volt protectors were connected between the open wires and driven grounds. This arrangement was more economical than the use of protection cable. The cable has gone through two complete lightning seasons without any failures or even permanent protector operations due to lightning.

In so far as manufacturing and storage facilities permitted, the reel lengths of the cable were assigned to their locations on the basis of obtaining as close an average loading section capacitance to the nominal value of 0.085 mfd per mile as was feasible. All reel lengths for the aerial sections were manufactured 1508 feet long, this length being sufficient to permit the assignment of a reel at any point in the line. Particular care was taken in this respect towards the ends of the cable where departures from the average would have the greatest effect on the return loss. To ensure proper assignment of reels, a route map was made up to scale with the manufacturer's reel number shown in its proper location.



Fig. 4—Replacement of old pole with new creosoted pine pole.

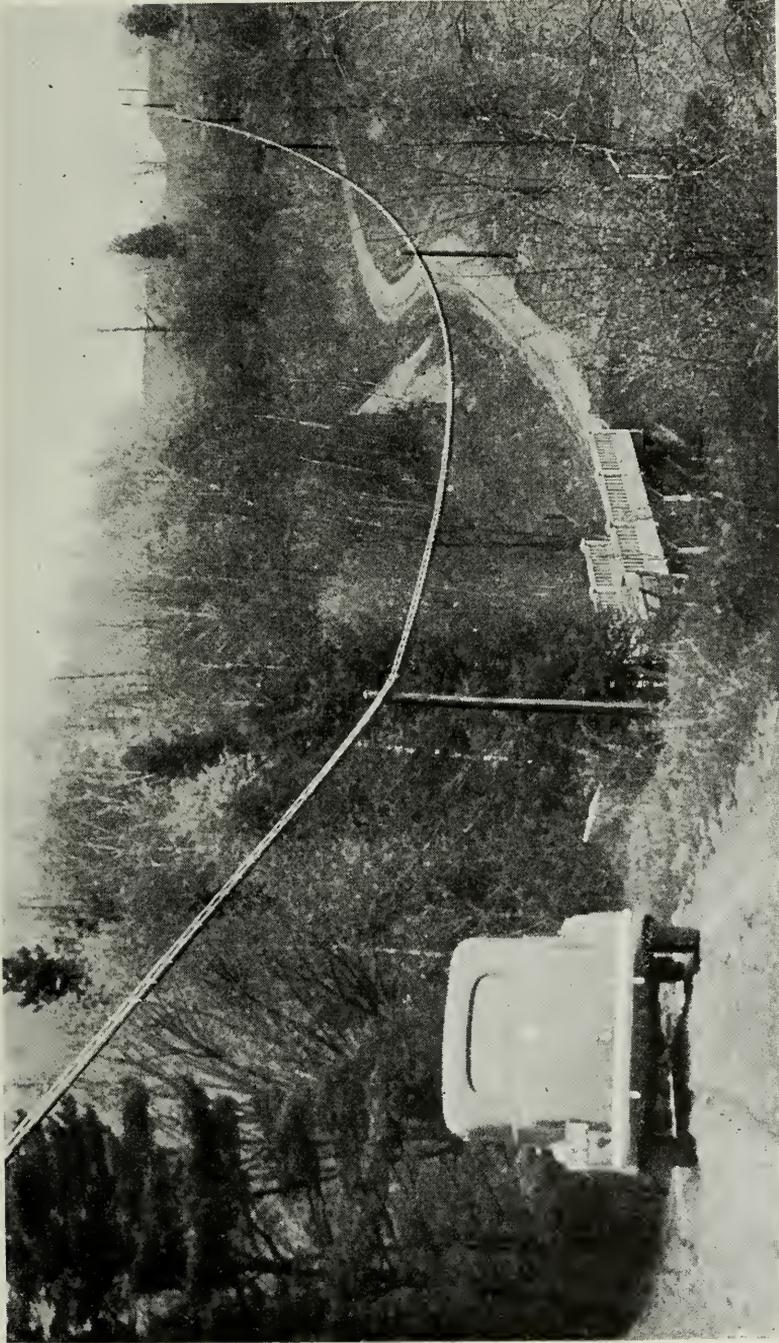


Fig. 5—Long span cable construction.

In addition, at the mid splice of each loading section, a test splice was made to equalize the capacity deviations. For these splices special linen boarding strips were used, each with 40 holes designated by a capacity ranging from about 15 per cent below to about 15 per cent above the expected average capacity of 1500 feet of cable. Small inexpensive capacity meters were used, and each pair was placed in the hole in the boarding strip corresponding to its capacity. The pairs were then spliced high to low capacity. This method did not require special testers, and substantially reduced splicing manhours.

Upon the completion of the splicing in each section, the section capacity of each pair was measured and recorded, and from this was determined the root mean square of the capacity deviations from the average capacity. These deviations combined with the deviation of the loading section average capacitances, loading coil spacings, and loading coil inductances, gave an irregularity function of 2 per cent which is almost identical with that for 19-gauge B88-50 cable. From this irregularity² function a 63 per cent return loss frequency curve was obtained, which is shown as curve 'A' in Fig. 8.

When 15 miles of the cable had been completed south from Barrie, a 100-pair cross-connecting box was temporarily spliced in so that data could be obtained as a further check on the design estimates.

For crosstalk tests each pair was terminated at the box in a 1700 ohm resistance, and measurements were made of all pair combinations (approximately 5000). For these measurements a 15A oscillator and 2A Noise Measuring Sets were used, thereby very materially reducing the manhours required as compared with the labour that would have been required had crosstalk measuring sets been used. Analysis of these tests indicated 99 per cent of combinations better than 76.0 db, an average of 86.6 db and 99.5 per cent meeting the required 74.5 db of the preliminary studies.

For attenuation measurements, the pairs were looped back at the cross-connecting box. In order to obtain a value of the attenuation at a known temperature, a complete set of measurements was made at about 6 o'clock in the morning after the resistance of one of the pairs had been found to have ceased dropping due to temperature change and the outside temperature at the Barrie office had been very nearly constant for about one-half hour. During the time the attenuation measurements were being taken, air temperatures were measured at four places along the 15-mile length of cable. From these tests the average 1000 cycle attenuation at 62° F. was found to be 0.26 db per

² See "Irregularities in Loaded Telephone Circuits," George Crisson, *Bell System Technical Journal*, October 1925.

mile. In Fig. 6 is shown the mean of the attenuations of three pairs plotted against frequency. On one pair the measurements were made at frequencies up to 6500 cycles, from which cut-off was determined to take place at 4000 cycles. Assuming a 60-mile circuit, the frequency at which the attenuation is about 10 db greater than 1000 cycles, is about 3500 cycles.

Before the return loss tests were made, impedance-frequency measurements were taken on two of the balancing networks for each of the three adjustments provided, and on a representative number of cable pairs, to determine the optimum network adjustment. The resistance component of the impedance for one of the networks and one cable

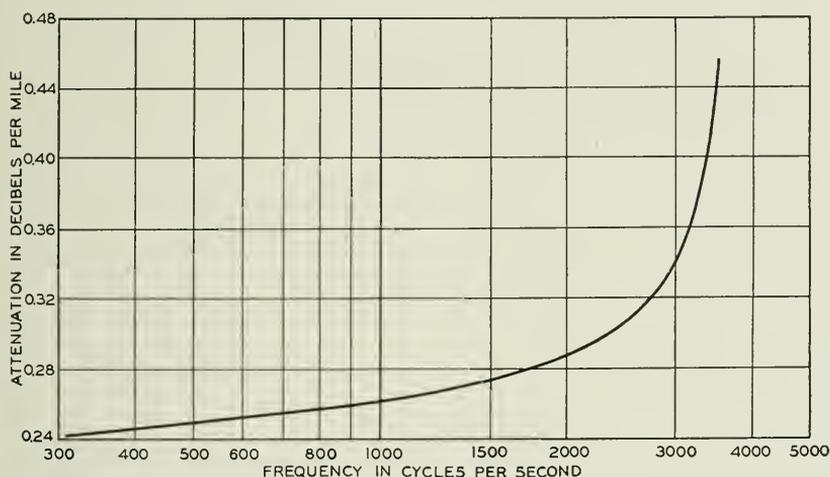


Fig. 6—Attenuation-frequency characteristic; mean of measurements on three pairs.

pair is shown in Fig. 7 (the two networks were found to be identical). From these tests the optimum network adjustment was determined to be that corresponding to a cable capacitance of 0.088 mfd per mile, which adjustment was then used for the return loss measurements.

As it was desired to obtain the singing point to be expected under operating conditions, the return loss measurements were made with the building-out condenser on the return loss set adjusted for optimum return loss at 2600, 2700 and 2800 cycles. The results of these measurements are given in Fig. 8 for comparison with the computed curve 'A' mentioned previously. The improvement at the higher frequencies of the actual over the computed values is due almost entirely to the method employed in making the tests, and indicates the advantage to be derived from individual adjustment of each circuit.

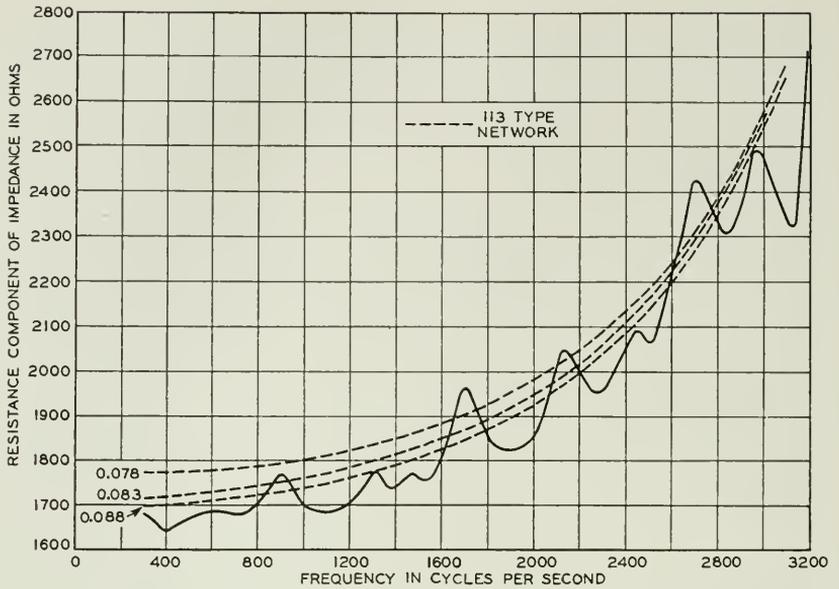


Fig. 7—Resistance component of impedance. 30.8 mile circuit, 19 CNB-B135. Termination, 113 type network.

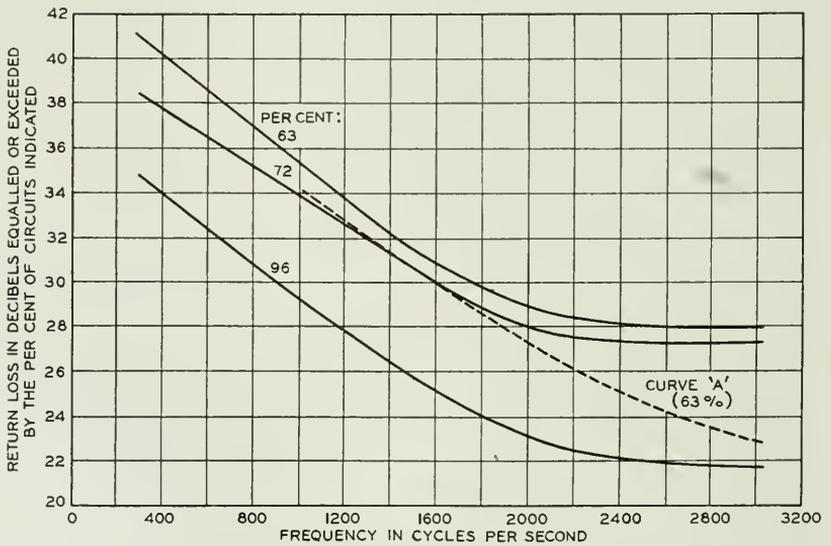


Fig. 8—Return loss—frequency characteristics. Measured on 61 circuits 30.8 miles long. Building-out condenser adjusted on each circuit for optimum return loss in the frequency range 2600, 2700 and 2800 cycles. Curve 'A' is the 63 per cent return loss computed from attenuation and irregularity function measured on cable.

These return loss measurements were made on pairs looped back at the cross-connecting box and terminated at Barrie in one of the networks.

COMPLETION TESTS

Upon completion of the cable, further overall tests were made. Particular attention was paid to those tests made from the Toronto end, to determine the effects of the use of the reloaded and respliced toll entrance cable.

Attenuation measurements at 1000-cycles were found to agree closely but, due to the effects of the toll entrance cable at Toronto,

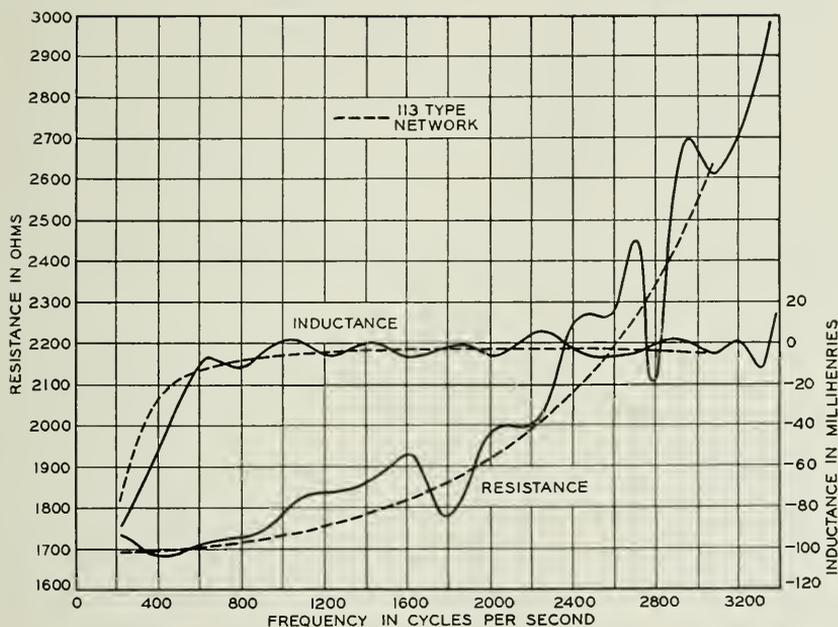


FIG. 9—Impedance measured at Toronto. Termination at Barrie, 113 type network. Sending-end, half section. Makeup from Toronto, 10.8 miles 16-ga., 0.062 mf. per mile, 0.0483 mf., 32.8^w per load section; 48.7 miles 19-ga., 0.085 mf. per mile, 0.0483 mf., 48.9^w per load section.

not to lend themselves to such rigorous analysis as those previously made.

To show one of the effects of the Toronto toll entrance cable, Figs. 9, 10, and 11, showing the resistance and inductance components of the impedance measured at Toronto, are included. These indicate that the important departure from the network characteristic for these pairs occurs in the inductance component at the lower frequencies. This departure is probably due to the difference in the loading section

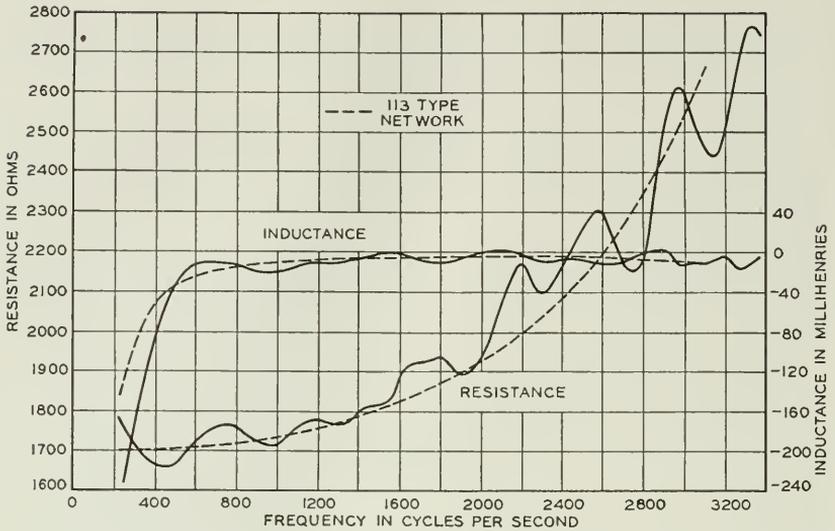


Fig. 10—Impedance measured at Toronto. Termination at Barrie, 113 type network. Sending-end, half section. Makeup from Toronto, 9.5 miles 19-ga., 0.062 mf. per mile, 0.0483 mf., 66.8^w per load section; 1.3 miles 16-ga., 0.062 mf. per mile, 0.0483 mf., 32.8^w per load section; 48.7 miles 19-ga., 0.085 mf. per mile, 0.0483 mf., 48.9^w per load section.

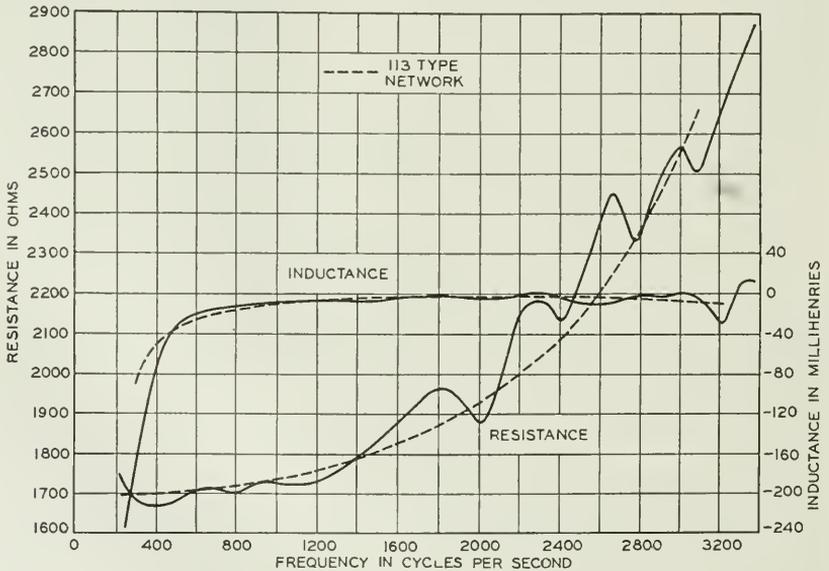


Fig. 11—Impedance measured at Toronto. Termination at Barrie, 113 type network. Sending-end, half section. Makeup from Toronto, 9.5 miles 19-ga., 0.062 mf. per mile, 0.0483 mf., 66.8^w per load section; 50.2 miles 19-ga., 0.085 mf. per mile, 0.0483 mf., 48.9^w per load section.

resistance from that of the main cable. (The geographical spacing on the quadded 0.062 mf. cable was 4100 feet as compared to 3000 feet on the non-quadded cable.)

Since representative return loss data had already been obtained for circuits under working conditions (Fig. 8), the completion return loss measurements were made for the network building-out capacity conditions assumed for the theoretical return loss characteristic (curve 'A,' Fig. 8). The results thus obtained are shown in Fig. 12 for Barrie and Fig. 13 for Toronto. In Fig. 12, the theoretical curve is shown for

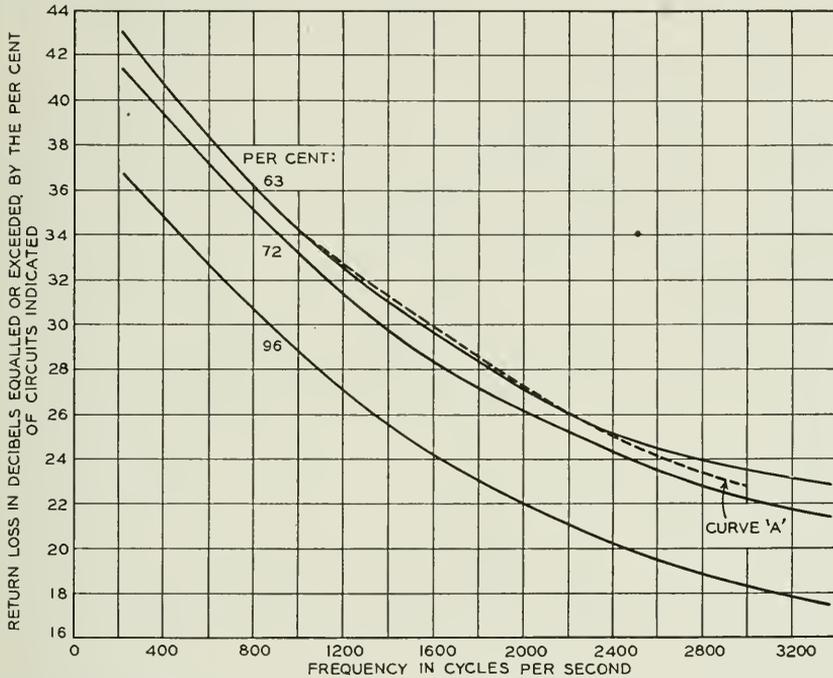


Fig. 12—Return loss—frequency characteristics. Measured from Barrie to Toronto on 53 pairs; building-out condenser adjusted to theoretical value; curve 'A' is the 63 per cent return loss computed from attenuation and irregularity function measured on cable.

comparison, and it is to be noted that the agreement with actual results is remarkably good. The results obtained at Toronto are better than those at Barrie, except below 600 cycles, which is the frequency range of the impedance departures discussed in connection with Figs. 9, 10, and 11.

Analysis of the near-end crosstalk measurements indicated that at Barrie 98.8 per cent, and at Toronto 96.3 per cent of the combinations were equal to or better than the 74.5 db assumed for the preliminary

calculations. Investigation of the combinations poorer than 74.5 db indicated that most of the pairs involved could be assigned either to non-repeated short circuits or to repeated short circuits on which the repeater gains were considerably lower, with consequent lower crosstalk gains, than on the full length circuits assumed for the limit of 74.5 db crosstalk. This required the opening of one splice near Toronto for pair rearrangement.

It was decided to place the cable under permanent gas pressure in order to control service interruption as far as practicable. As there was no previous experience available for cables of this size, an investiga-

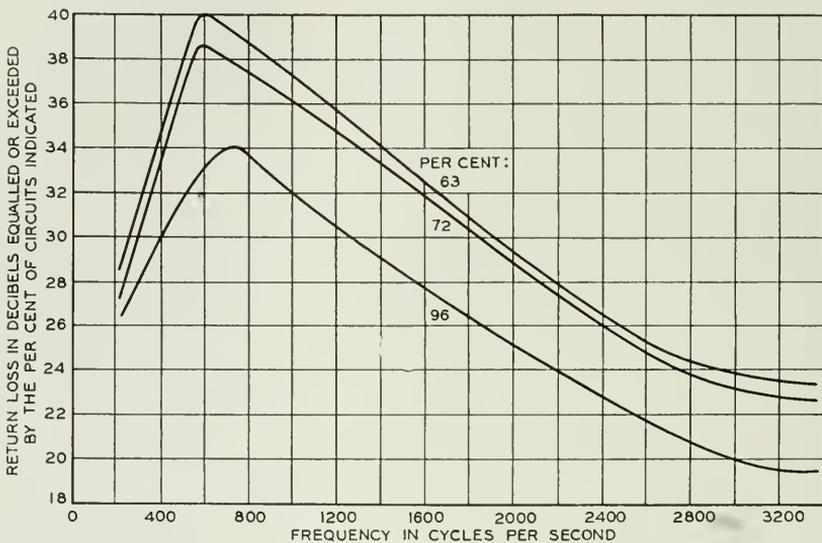


Fig. 13—Return loss—frequency characteristics; measured from Toronto to Barrie on 87 pairs; building-out condenser adjusted to theoretical value.

tion on the job was undertaken to obtain the information necessary for successful application of the gas pressure installation. Based on the results obtained, the installation of gas pressure was completed satisfactorily.

ACKNOWLEDGMENT

The design and installation of this cable represent the coordinated efforts of many people—members of the organizations of the Bell Telephone Laboratories Inc., the American Telephone & Telegraph Company, Northern Electric Company, Western Electric Company and the Bell Telephone Company of Canada—too many for anyone to be specifically mentioned. To all of these credit is due and is here given.

The Computation of the Composite Noise Resulting from Random Variable Sources

By E. DIETZE and W. D. GOODALE, Jr.

A statistical method is described for computing the meter reading which would be obtained on a sound level meter when used to measure room noise resulting from the random concurrent operation of a number of intermittent or continuous noise sources. The application of the method in the solution of practical problems is illustrated.

IT is generally recognized that the effects of noise upon the individual exposed to it are, to a large extent, dependent on the loudness of the noise. Various tests have been made of the relation between loudness and the different effects of noise, such as interference with hearing, reaction on the nervous system, disturbance of rest, reduction of working efficiency,¹ etc.

It has also been recognized that the ear itself, in general, is not a convenient means for the accurate measurement of loudness, especially in absolute terms. To overcome this difficulty sound level meters² have been made available for the measurement of acoustic noises or sound in general.

This paper is concerned with the application of such sound level meters in the study of noise problems and, in particular, with the question of determining the contribution of individual noise sources to the general "composite noise" including noise sources whose outputs are random, discontinuous variables. The paper does not concern itself with the attributes of loudness or the effects of noise, but merely with the computation of a meter reading of the total noise from available measurements of the noise components. It is recognized, of course, that not only are sound level meter readings an incomplete description of the effect of a change in noise but that considerable experience is required to appreciate properly the significance of the decibel unit employed.

TYPES OF PROBLEM

The method described in this paper has been developed to meet a very practical need experienced in the solution of a large variety of noise problems. To illustrate, consideration may be given to reducing

¹ For reference see Bibliography.

the noise in a room by excluding street noise, using quieter office equipment or sound absorbing material. Each of these measures will involve a certain expense and will reduce room noise a certain amount. Which of these measures will be of greatest benefit and most economical, i.e., give the greatest noise reduction per dollar expenditure?

In another assumed case, the noise from a certain noise producer, a piece of machinery, a ventilating system, etc., is known. Will this apparatus be objectionable in the particular location for which it is considered?

These and similar questions can be answered by computation while the project is still in the planning stage, whereas measurements can be made only after the change has been made, i.e., after the money for the project has been spent.

The computation method, as these illustrations show, is useful in specifying apparatus and in planning working or living quarters from the noise standpoint, in studying the comparative effectiveness of various noise reducing means, etc. The method has been used in many practical problems in this way, with satisfactory results. In a number of applications covering noise from 55 to 75 db sound level the computed and measured absolute values agreed, on the average within 1.0 db, and in the worst case within 2.0 db. Computations of the effect resulting from modifications of the noise sources were checked within closer limits. A few illustrations of applications are given at the end of this paper.

THE IMPULSIVE CHARACTER OF NOISES

Acoustical noise frequently is composed of sounds from a large number of sources each of which produces a relatively small proportion of the total noise. Usually these individual noise sources are discontinuous, consisting of a series of individual impulses. Consider, for instance, noise from a busy street. The hearers' first impression is that of a general roar. After a period of listening, however, a variety of individual sources may be distinguished, such as: The movement of automobiles, squeaking of brakes, whistles, street car wheels and bells, hammering and riveting from building operations, footsteps and conversations of people, etc. Each of these sources has a distinct time pattern and even those that appear most steady can frequently be broken up into impulses. For instance, the noise from an automobile passing down the street is composed of a series of impact noises which depend on unevenness of the pavement, the driving gears, number of cylinders in the engine, etc.; the hum of conversation of people in the street is composed of individual syllabic speech sounds from the different talkers.

These impulses occur at a rate which usually is not uniform. Provided, however, that the general conditions do not change, the rate approaches uniformity if the time interval considered is sufficiently large. Some of the impulses from the different sources are superimposed upon each other while others fall in the intervals between the impulses from other sources. As the amount of noise increases, two general phenomena are observed: First, the loudness of the noise increases due to the superposition of impulses; secondly, the noise becomes steadier due to the more complete filling in of relatively silent intervals (20 db or more below the average).

DEFINITION OF TERMS

The solution of the problem of computing the total noise from its component parts requires the definition of a number of terms and a study of the characteristics of the implied measuring instrument.

Definition 1

Each individual producer of noise is referred to as a "noise source."

Illustrations of noise sources are: For the case of room noise—the conversation of one person, the noise from a typewriter or from a fan in the room. A number of sources of street noise have been mentioned above in illustrating the impulsive character of common noises.

Definition 2

The deflections on the measuring device (sound level meter) produced by the impulses of a single source are called "source peaks."

A peak is obtained by passing a noise impulse into a sound level meter. Depending on this measuring device, the characteristics of a peak differ from those of an impulse. The characteristics of the sound level meter, therefore, are important in connection with this computation method. Three of these, the frequency response, the rule of combination of the frequency components of a complex wave, and the dynamic characteristic of the indicating meter, are here considered in detail. These are defined in the "American Tentative Standards for Sound Level Meters" approved by the American Standards Association² from which the following abstracts are made:

1. The free field frequency response of a sound level meter, provided only one response is available, shall be the 40 decibel equal loudness contour modified by differences between random and normal free field thresholds. Methods are given in the ASA specification for correcting the reading when the microphone of the sound level meter responds differently to sound waves arriving with different angles of incidence.

2. The rule of combination is specified so that the power indicated for a complex wave shall be the sum of the powers which would be indicated for each of the single frequency components of the complex wave acting alone.
3. The dynamic characteristic of the indicating instrument is to be such that the deflection of the indicating instrument for a constant 1000-cycle sinusoidal input shall be equalled by the maximum deflection of the indicating instrument for a pulse of 1000-cycle power which has the same magnitude as the constant input and a time of duration lying between 0.2 and 0.25 second.

In addition, the method of reading the sound level meter is important. Where the noise is steady, it is fairly obvious how the meter should be read. When, however, the noise fluctuates, a certain amount of judgment is involved in obtaining an average. A satisfactory procedure in this event is to take a series of instantaneous readings of the noise peaks at approximately 5-second intervals for a period of time sufficient to include all noise sources. One or more of these series of measurements may be made depending on the regularity of occurrence of the noises of interest. The average and standard deviation of the fluctuating noise may then be determined from these measurements.

Using simplifying approximations based on these specified characteristics a peak may be defined as follows:

A peak is an impulse integrated by the measuring device. Its frequency components are weighted in accordance with the loudness weighting incorporated in the meter and combined by direct power addition.

It will be seen from the foregoing that the duration of the source peaks depends on the period of the indicating meter. It has been found that 0.2 second gives satisfactory correlation between computed values and actual sound level meter readings, and is in reasonable agreement with the above specified characteristics. Due to the meter characteristics, full magnitude is not indicated for impulses shorter than 0.2 second. Several impulses in the same integration period appear as a single peak on the meter. Impulses lasting longer may be regarded as producing a number of consecutive peaks. A steady noise, for instance, would be considered as consisting of a series of consecutive peaks of equal magnitude.

On the assumption of discrete integration intervals the average reading on a single source is the arithmetic mean of the intensities of the source peaks. Hence for a source, j , producing on the average m_j peaks per minute of intensities, I_1, I_2, I_m , the average reading on the meter is given by

$$I_j = \left(\frac{1}{m} \sum_{i=1}^m I_i \right)_j \quad (1)$$

On the assumption of discrete integration intervals, furthermore, a source can produce no more than one peak every 0.2 second. The maximum number of peaks per minute that can be obtained from a single source, consequently, is 300.

Definition 3

The noise from all sources as measured by the indicating meter is called composite noise.

Room noise measured at a given observing position in the room is an illustration of composite noise. The peaks of a composite noise are called "composite peaks." Composite peaks have similar characteristics to source peaks as regards duration, frequency weighting, etc.

STATISTICAL METHOD OF COMBINING NOISE SOURCES

In developing this computation method the principal aim of the authors has been to provide a practical, working method which is easy to handle yet is sufficiently reliable for engineering purposes. In accordance with this objective, a number of simplifying assumptions have been made. Some of these have been indicated in connection with the discussion of the assumed characteristics of noise peaks. The division of the time into discrete 0.2 second intervals is another approximation which has been made. The statistical treatment, in addition, includes approximations which are usual in probability mathematics of this type. Practical experience has shown that these approximations do not lead to errors which affect the usefulness of the method.

In the following an expression is derived for computing the average intensity of the composite noise from the average intensities of the source peaks and their number. Consideration is first given to the case when only a single source peak may occur in each 0.2 second interval. The consideration is then extended to cover the general case when more than one source peak may occur in a 0.2 second interval.

When only one source peak may occur in a 0.2 second interval, the average intensity \bar{I} of the composite noise for these intervals is the arithmetic mean of the intensities of the source peaks, weighted by their frequency of occurrence.

$$\bar{I} = \frac{m_1}{N} I_1 + \frac{m_2}{N} I_2 + \cdots + \frac{m_n}{N} I_n, \quad (2)$$

where

I_1, I_2, \dots, I_n = average intensities of the sources 1, 2, \dots , n ,
 m_1, m_2, \dots, m_n = number of peaks of each source per minute,

$N = \sum_{j=1}^n m_j$ = total number of source peaks per minute.

If several source peaks occur in the same 0.2 second interval, they will appear as a single composite peak on the meter. On the assumption of discrete 0.2 second intervals, these source peaks coincide. Their intensities, consequently, add up directly. For instance, if two source peaks occur during each integration period, the average intensity of the composite noise will be twice the arithmetic mean of the intensities. Similarly, the average intensity of the composite noise, when the number of source peaks per 0.2 second interval averages α , will be

$$I = \bar{I} = \alpha \left(\frac{m_1}{N} I_1 + \frac{m_2}{N} I_2 + \dots + \frac{m_n}{N} I_n \right). \quad (3)$$

Let M = the total number of composite noise peaks per minute. The maximum value that M can have is 300, the number of integration periods per minute. Unless the composite noise is continuous, however, there will be a certain proportion of time, t_0 , in which no composite peaks occur. M then can be determined from the relation:

$$M = (1 - t_0) 300. \quad (4)$$

If, on the average, α source peaks per 0.2 second occur, the following relation holds between the total number of source peaks, N , and the number of composite peaks:

$$M = \frac{N}{\alpha}. \quad (4a)$$

Introducing this expression in equation (3) gives

$$I = \frac{m_1}{M} I_1 + \frac{m_2}{M} I_2 + \dots + \frac{m_n}{M} I_n. \quad (5)$$

As shown by equation (4), M is a function of t_0 , the proportion of time in which no composite noise peaks occur. The value for t_0 can be found, as follows: The proportion of time when source j has a peak is equal to the probability $p_j = m_j/300$, and the proportion of time when source j has no peak is $q_j = 1 - p_j = 1 - (m_j/300)$. The proportion of time, t_0 , when there are no peaks from any source then is equal to

the product

$$t_0 = q_1 q_2 \cdots q_i \cdots q_n.$$

This expression can be simplified, when the number of sources is large and none is particularly outstanding, by considering, instead of the individual sources, an average source having $m = N/n$ peaks.

The average probability then is

$$p = \frac{m}{300} = \frac{N}{300n},$$

and

$$q = 1 - \frac{N}{300n},$$

which leads to the approximation:

$$t_0 = q^n = \left(1 - \frac{N}{300n}\right)^n.$$

This expression can be further simplified when the number of sources is large and $p = N/300n$ is small by using the Poisson exponential limit:

$$t_0 = \left(1 - \frac{N}{300n}\right)^n \cong e^{-N/300},$$

where

$$e = 2.718 \dots$$

so that for this case

$$M = (1 - e^{-N/300})300. \quad (4b)$$

MEASUREMENT OF SOURCE DISTRIBUTIONS

The method outlined in this paper for computing the composite noise assumes that information on the noise sources is available. Such data, therefore, must be obtained before the method can be applied. Representative measurements for a particular type of noise source, however, when once obtained, can be used in any future noise computation involving such a source.

It is necessary to consider carefully the acoustic conditions under which the sources are measured. For greatest accuracy the ambient noise level at the point of measurement should be 20 db or more below the average level of the source. Errors due to reflections can be minimized by making the measurements at a relatively short distance from the source out of doors or in a room that contains a large amount of absorbing material. A distance of 2 feet is a convenient value for most cases, and will be used as a reference value throughout the rest of this paper.

Readings should be obtained on all the noise peaks while the source is being operated in a normal manner. If the scale of the particular sound level meter used is limited, it may not be possible to read the highest as well as the lowest peaks with a single potentiometer setting. In such cases, the distribution of peaks may be measured in two or more groups.

Figure 1, Curve A, illustrates the measurement of source peaks in the laboratory and represents a cumulative distribution of the peaks from a typewriter as measured at a horizontal distance of about 2 feet from the type bar guide. The machine was operated by an experienced typist at an average rate.

When it is not possible to simulate actual conditions of use of a device sufficiently well in the laboratory, measurements on the source

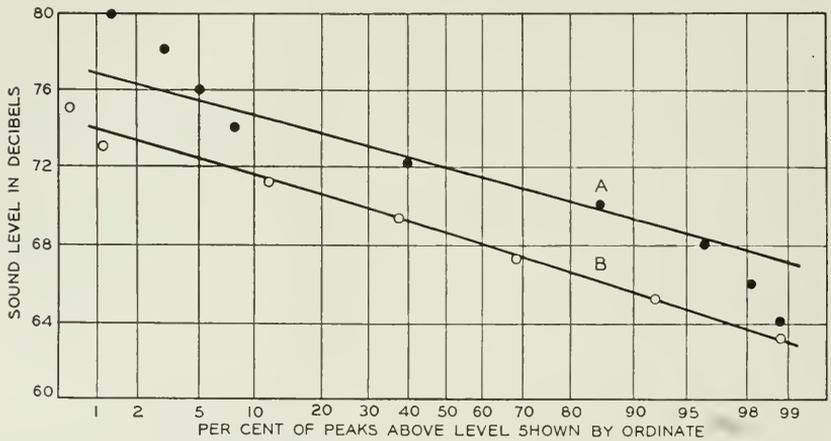


Fig. 1.—Distribution of noise peaks in a typing room.

will have to be taken in the field. This may involve measuring in the presence of considerable noise from other sources. In general, it is feasible only to measure source peaks which are above the ambient noise level. If, however, an appreciable number of peaks is above this noise level, the rest of the distribution can be estimated and the average value determined. Statistical methods for doing this have been worked out for the case of normal distribution curves.³ Experience has indicated that the distributions of noise in db frequently are approximately normal, so that these methods are applicable.

Figure 2 is an illustration of a distribution of a group of sources measured under adverse noise conditions. This curve shows the noise which came from the metal trays in a cafeteria. The distribution had to be obtained in the field because it was not feasible to estimate in

the laboratory how the customers would handle the trays. The points on the curve indicate the peaks that could be measured in the cafeteria which had an average composite noise of 66.5 db. It will be seen that the lowest peaks that could be measured satisfactorily were at 74 db sound level. The rest of the curve was estimated using the statistical methods referred to above.

The curves in Figs. 1 and 2 are plotted on "arithmetic probability paper." On this paper, cumulative normal distributions appear as straight lines.

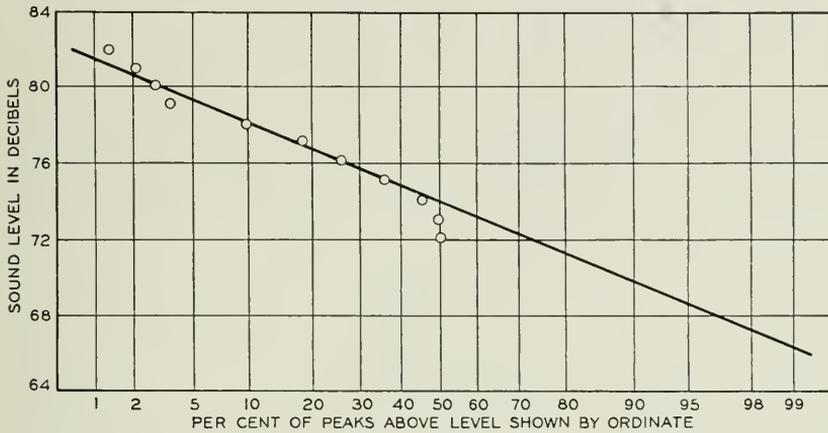


Fig. 2—Distribution of noise peaks from metal trays in a cafeteria.

EFFECT OF ROOM CHARACTERISTICS

Generally the noise sources are at various locations so that it is necessary to determine how much the noise from each is reduced by its distance from the observing point assumed for the computation.

Since it is not the primary concern of this paper to discuss the distribution and decay of sounds in rooms, only a very simple approximate method of computing distance losses, based on the classical theory of the steady-state distribution of sound in a room, is given here. This method has been found adequate for practical purposes in rooms having relatively simple geometric shape and large enough dimensions so that the sound is diffused. For a more complete treatment of room acoustics the reader should refer to the literature on this subject.⁴

The total steady-state intensity, I_T , at an assumed observing position in a room consists of two parts: I_R , the reflected sound intensity and I_D , the direct sound intensity, so that:

$$I_T = I_R + I_D.$$

Assuming the reflected sound to be uniformly distributed in the room, it can be shown that:⁴

$$I_R = \frac{.0038E}{\frac{aS}{S-a}},$$

where E = power emitted by source, in ergs per second,
 S = total surface area of the room in square feet,
 a = absorption in square feet of equivalent open window.

Introducing $F = aS/(S - a)$, the above becomes:

$$I_R = \frac{.0038E}{F}.$$

Assuming the sound source to radiate hemispherically, as is frequently the case because it is associated with a large surface acting as a baffle, the direct sound intensity is:

$$I_D = \frac{E}{2\pi r^2 v},$$

where r = distance from source, in feet,
 v = velocity of sound, in feet per second.

In the above expression the direct sound intensity decreases inversely as the square of the distance from the source. This shows that room absorption is effective mainly in reducing the noise from sources at a considerable distance from the observing point, but has relatively little effect on nearby sources.

The curves in Fig. 3 give the variation in the total sound intensity, I_T , with distance from the source for different values of $F = aS/(S - a)$, as computed by means of the above expressions.

COMPUTATION OF COMPOSITE NOISE

In the following, the application of the statistical method outlined above is discussed. Since noise measurements are usually expressed in db sound level, it is necessary to change the form of the equations given in the preceding sections. For this purpose equation (5) is rewritten as follows:

$$\frac{I}{I_0} = \frac{m_1 I_1}{M I_0} + \frac{m_2 I_2}{M I_0} + \dots + \frac{m_n I_n}{M I_0}, \quad (5a)$$

where I_0 = reference sound intensity.

This equation can also be written

$$\frac{I}{I_0} \frac{M}{300} = \frac{m_1}{300} \frac{I_1}{I_0} + \frac{m_2}{300} \frac{I_2}{I_0} + \dots + \frac{m_n}{300} \frac{I_n}{I_0}. \tag{5b}$$

Equation (5b) is somewhat more convenient in computing than (5a). In this equation a weight factor is associated with each intensity ratio, which is in each case the actual number of peaks divided by the maximum possible number of peaks.

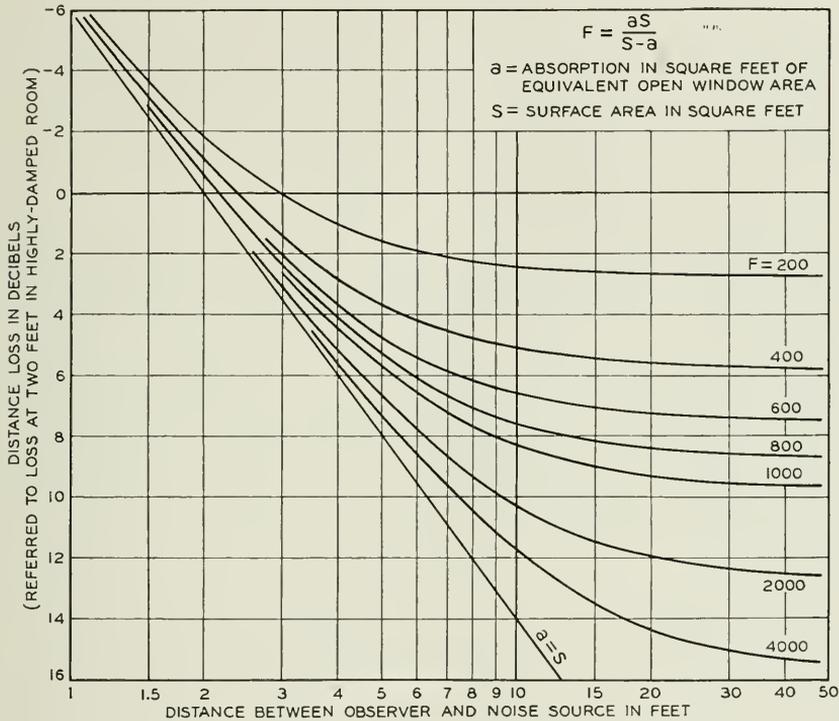


Fig. 3—Loss of intensity with distance from noise source for various amounts of room absorption.

Assuming that the intensity corresponding to the average of the db distribution of each source may be used, the following relations exist for the source noises in db sound level:

$$\left. \begin{aligned} A_1 &= 10 \log_{10} \frac{I_1}{I_0} \\ A_2 &= 10 \log_{10} \frac{I_2}{I_0} \\ &\dots \\ A_n &= 10 \log_{10} \frac{I_n}{I_0} \end{aligned} \right\} \tag{6}$$

and for the average composite noise in db sound level:

$$A = 10 \log_{10} \frac{I}{I_0}. \quad (7)$$

It is usually convenient to use logarithmic weight factors

$$\left. \begin{aligned} w_1 &= 10 \log_{10} \frac{m_1}{300} \\ w_2 &= 10 \log_{10} \frac{m_2}{300} \\ &\dots \dots \dots \\ w_n &= 10 \log_{10} \frac{m_n}{300} \\ w &= 10 \log_{10} \frac{M}{300} \end{aligned} \right\} \quad (8)$$

Figures 4 and 5 permit ready computation of these logarithmic weight factors. The chart in Fig. 5 is based on the relation between M and N given in equation (4b). It should be recalled that the derivation of this equation involved a number of approximations. This expression especially does not apply when one or more of the noise sources are continuous, in which case the exact expression (eq. 4) gives $M = 300$ (for $t_0 = 0$). Hence $w = 0$ in this case.

The terms of equation (5b) then can be rewritten in logarithmic form by using equations (6), (7) and (8), as follows:

$$\left. \begin{aligned} A_1 + w_1 &= 10 \log_{10} \frac{m_1}{300} \frac{I_1}{I_0} \\ A_2 + w_2 &= 10 \log_{10} \frac{m_2}{300} \frac{I_2}{I_0} \\ &\dots \dots \dots \\ A_n + w_n &= 10 \log_{10} \frac{m_n}{300} \frac{I_n}{I_0} \\ A + w &= 10 \log_{10} \frac{M}{300} \frac{I}{I_0} \end{aligned} \right\} \quad (9)$$

This gives the following formula:

$$10^{\frac{(A+w)}{10}} = 10^{\frac{(A_1+w_1)}{10}} + 10^{\frac{(A_2+w_2)}{10}} + \dots + 10^{\frac{(A_n+w_n)}{10}}, \quad (10)$$

from which the average composite noise A can be found.

The application of this expression is materially simplified by the use of the chart shown in Fig. 6. Power addition of a number of

components may be carried out with this chart by first adding two components, then adding the resultant to a third component and continuing until all components have been summed up. Incidentally,

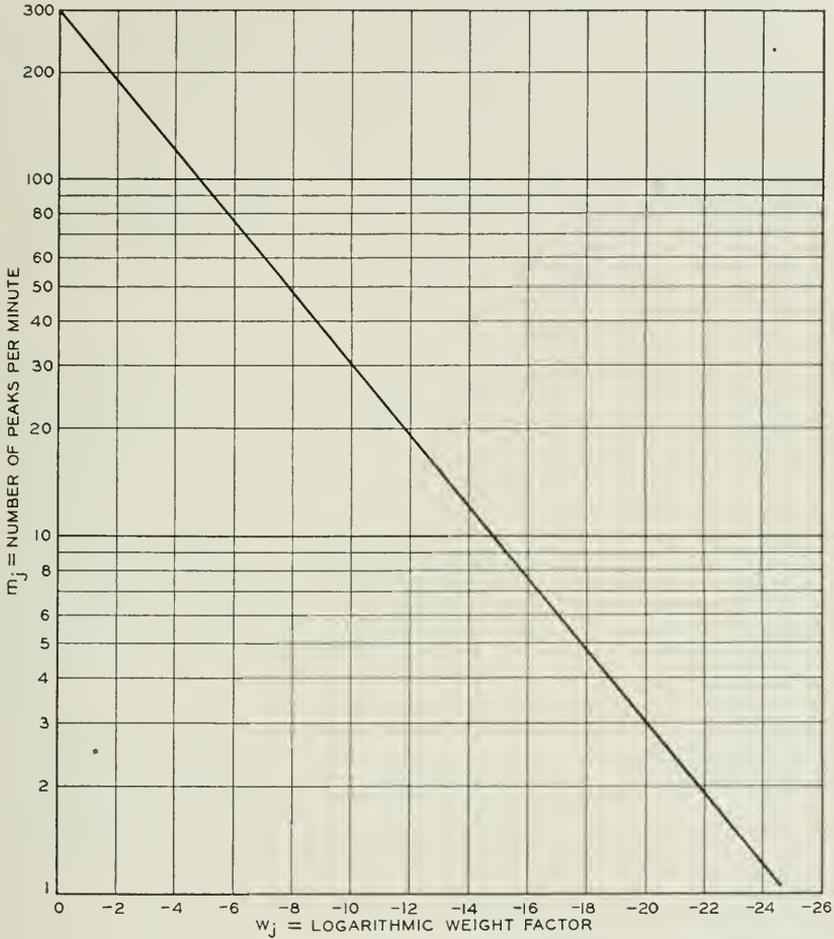


Fig. 4—Relation between peaks per minute (m_j) produced by a noise source and its logarithmic weight factor (w_j).

the chart shows that the contribution to the composite noise from a source whose weighted intensity ($A_1 + w_1$) is 20 db or further below that of another noise source ($A_2 + w_2$) is negligible.

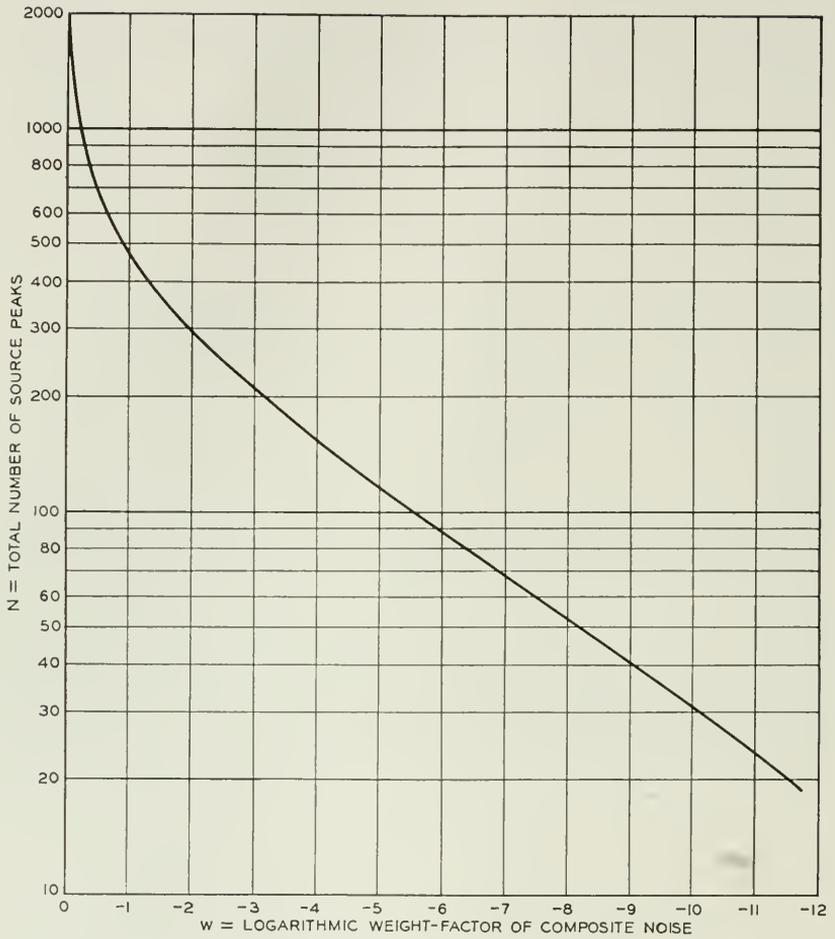


Fig. 5—Relation between total number of peaks per minute (N) and the logarithmic weight factor of composite noise (w).

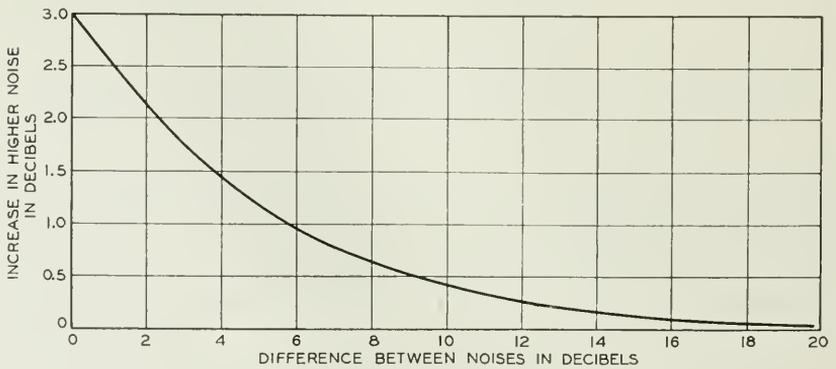


Fig. 6—Power addition of noises.

APPLICATIONS

In the following, several applications of the theory are made to illustrate its practical usefulness.

I. The composite noise in a typing room is computed. This computed noise is compared with actual measurements. The effect of increased room absorption is discussed.

II. The effect on the composite noise of installing additional office equipment is computed for the two cases where this equipment produces a continuous noise and where its noise is intermittent.

III. The maximum permissible noise from added equipment is determined on the basis that the composite noise level shall not be increased by more than 0.5 db.

Problem I

For the purpose of computing the noise at a given location in the typing room, the following information was obtained:

A distribution of the noise from a typical typewriter was measured at a distance of 2 feet from the type bar guide while the machine was being operated at a normal rate. This is shown by Curve *A* of Fig. 1.

A location was chosen as a point of observation, and the distances between it and the typing desks were measured.

Estimates of the time spent in typing at each desk were obtained which, taken together with data on average typing speeds, gave information on the number of typing peaks produced per minute at each desk.

Computation of the absorption of the room using the usual values of absorbing coefficients⁵ gave a value of 650 units for F .

Noise due to other sources, such as conversation and street noise, was negligible in this room.

The table shown below was then prepared.

In this table Column 2 gives the average noise A_j' produced by each of the sources at 2 feet distance. This value is the median point of Curve *A* in Fig. 1. Column 3 is the average number of source peaks per minute m_j produced at each desk. Column 4 is obtained from Column 3 by using Fig. 4. The total number of source peaks is $N = 750$, and from Fig. 5 the composite noise weight factor $w = -0.4$ db. The distances between the observing position and each source are given in Column 5 and the losses in db due to these distances are given in Column 6. These values were obtained from Fig. 3 for a value of $F = 650$. Column 7 is obtained by subtracting the losses of Columns 4 and 6 from the values of Column 2.

Adding the values of Column 7 successively on a power basis by means of the curve in Fig. 6 gives $A + w$ from which is obtained the total composite noise $A = 69.7$ db sound level. This differs by approximately 1 db from the average of the measured composite noise distribution shown by Curve *B* of Fig. 1.

The effect of sound treatment on the walls and ceiling in reducing the typing room noise may readily be calculated by means of the curves in Fig. 3. Supposing that the added absorption raises the value of F from 650 to 2000 units, this figure shows that noise produced by sources 20 feet or more away from the observing point would be reduced by approximately 5 db. At shorter distances the reduction would be less. For the observing position here considered, a computation similar to that carried out above indicates that the composite noise level would be reduced about 3 db by the added absorption in the room.

TABLE

(1)	(2)	(3)	(4)	(5)	(6)	(7)
Source No.	$A_j' =$ Average Source Noise at 2 ft.	$m_j =$ Source Peaks per Minute	$w_j =$ Freq. Weight Factor	Distance from Source	Intensity Loss vs. 2 Feet	$A_j + w_j =$ Weighted Source Noise at Observing Position
	db Sound Level		db	Feet	db	db Sound Level
1	71.9	105	-4.5	18	-7.4	60.0
2	71.9	90	-5.2	13	-7.2	59.5
3	71.9	90	-5.2	9	-6.7	60.0
4	71.9	120	-4.0	9	-6.7	61.2
5	71.9	120	-4.0	7	-6.0	61.9
6	71.9	40	-8.8	7	-6.0	57.1
7	71.9	65	-6.6	7	-6.0	59.3
8	71.9	120	-4.0	7	-6.0	61.9

Total Source Peaks: $N = 750$

Power Addition $A + w = 69.3$ db
(Fig. 6)

$w = -0.4$ db
(Fig. 5)

$A = 69.7$ db sound level.

Problem II

A piece of office machinery, such as an addressing or copying machine, which produces an average sound level of 75 db at 2 feet distance, is to be installed in the typing room considered in Problem I, 20 feet away from the observing position. How much will the composite noise level be raised?

(a) *The machine produces a steady noise.* The new value for the number of composite noise peaks is then 300 and the weight factor of this noise is zero. The distance loss of the machine noise (for $F = 650$) is -7.5 db from Fig. 3. Hence, the weighted value of

the noise from the new machine at the observing position is:

$$75 - 0 - 7.5 = 67.5 \text{ db sound level.}$$

Adding this figure to the weighted value of the existing composite noise ($A + w = 69.3$ db) on a power basis gives the new composite noise value of 71.5 db sound level (the new weight factor being zero). Hence, the composite noise at the listening position is increased 1.8 db by the machine.

(b) *The machine produces noise intermittently.* The increase in the composite noise level will not be as great in this case as in the preceding case. For example, assuming the rate to be 100 peaks per minute, the new value of N will be 850 peaks per minute and the corresponding weight factor from Fig. 5 will be -0.3 db. For the noise from the new machine, the weight factor (by Fig. 4) is -5.0 db. The distance loss as before will be -7.5 db. The weighted value of the machine noise at the observing position is then:

$$75 - 5.0 - 7.5 = 62.5 \text{ db sound level.}$$

Adding this figure on a power basis to the weighted value of the existing composite noise, 69.3 db, results in a new weighted composite sound level, $A + w = 70.1$ db. Since $w = -0.3$ db, this gives $A = 70.4$ db. Hence, the composite noise is increased 0.7 db by the intermittent machine noise.

Problem III

What is the maximum permissible noise, measured at 2 feet, which the machine considered in Problem II, may produce without raising the composite noise in the typing room by more than 0.5 db?

(a) *The machine produces a steady noise.* In this case, the composite noise has 300 peaks and its weight factor is zero. The existing composite noise was 69.7 db sound level (see Problem I). The maximum permissible value of the new composite noise level is consequently

$$A = 69.7 + 0.5 = 70.2 \text{ db sound level.}$$

Let the unknown machine noise be A_9 (its weight factor is zero), and since for the existing composite noise $A + w = 69.3$, equation (10) gives:

$$10^{\frac{70.2}{10}} = 10^{\frac{69.3}{10}} + 10^{\frac{A_9}{10}}.$$

Entering the ordinate of Fig. 6 at the value of $70.2 - 69.3 = 0.9$ db, the chart indicates that A_9 must be 6.3 db below 69.3. Hence $A_9 = 63$ db sound level at the observing position.

The distance loss for 20 feet is -7.5 db. The machine then could produce a noise of:

$$63.0 + 7.5 = 70.5 \text{ db sound level}$$

at 2 feet distance without raising the composite noise level by more than 0.5 db at the observing position.

(b) *The machine produces noise intermittently.* The solution of the problem in this case follows the same lines as in Part (a) except that the weight factors are changed. The rate for the new machine noise, A_9 , is assumed to be 100 peaks per minute, so that $w_9 = -5.0$ db, as in part (b) of Problem II. The maximum permissible value of the new composite noise is 70.2 db sound level (as before) but its weight factor now is -0.3 db as in part (b) of Problem II. The weighted value of the existing composite noise as before is 69.3 db sound level. Equation (10) then gives:

$$10^{\frac{(69.9)}{10}} = 10^{\frac{69.3}{10}} + 10^{\frac{(A_9-5.0)}{10}}.$$

From Fig. 6 it is found that for a value of $69.9 - 69.3 = 0.6$ on the ordinate, the abscissa is 8.3 db. Hence, $A_9 - 5.0$ must be 8.3 db below 69.3 or $A_9 = 66.0$ db sound level at the observing position. Applying the same distance loss as before, the machine could produce a noise of 73.5 db sound level at 2 feet without increasing the composite noise level by more than 0.5 db at the observing position.

From the computations, then, it may be expected that adding a steady noise will increase the general noise level more than adding an intermittent noise having the same average value, when there are a number of sources operating. That this is actually so can readily be verified by sound level measurements. As has been stated, sound level measurements under most conditions are directly related to the effects of noise upon the individual exposed to it, and the method described provides a convenient and reasonably reliable way of computing such readings and thereby makes possible the engineering analysis of noise problems.

BIBLIOGRAPHY

1. "Environment and Employee Efficiency" by Harold Berlin and others—Office Management Series No. 81—Copyright 1937, American Management Association.
"City Noise"—Report of Noise Abatement Commission 1930, Dept. of Health, New York City.

2. "American Tentative Standards for Sound Level Meters for Measurement of Noise and Other Sounds"—Z24.3—Approved American Standards Association, Feb. 17, 1936.
"Indicating Meter for Measurement and Analysis of Noise" by T. G. Castner, E. Dietze, G. T. Stanton and R. S. Tucker—District Meeting A. I. E. E., April 1931, Rochester, N. Y.
3. "Determination of Normal Curve from Tail"—Table XI of "Tables for Statisticians and Biometricians" by Karl Pearson—Part I—Second Edition, 1924.
"Graduation by a Truncated Normal" by Nathan Keyfitz—*Annals of Mathematical Statistics*—Vol. 9, March 1938.
4. "Collected Papers on Acoustics" by W. C. Sabine,
"Reverberation Time in 'Dead' Rooms" by Carl F. Eyring—*The Journal of the Acoustical Society of America*, Vol. I, No. 2, January 1930,
"Architectural Acoustics" by V. O. Knudsen.
5. *Official Bulletin of the Acoustical Materials Association*, No. 6, March 1938.

Load Rating Theory for Multi-Channel Amplifiers*

By B. D. HOLBROOK and J. T. DIXON

The amplifiers of multi-channel telephone systems must be so designed with regard to output capacity that interchannel interference caused by amplifier overloading will not be serious. Probability theory is applied to this problem to determine the maximum single frequency output power which a multi-channel amplifier should be designed to transmit as a function of N , the number of channels in the system. The theory is developed to include the effects of statistical variations in the number of simultaneous talkers, in the talking volumes, and in the instantaneous voltages from speech at constant volume.

INTRODUCTION

IN A perfect multi-channel carrier telephone system, each channel would be entirely free from interference produced by the energy present in the other channels. Since all the channels are amplified by the same repeaters, which as a practical matter cannot have perfectly linear characteristics, this is an ideal that may be approached but not completely realized. The interchannel interference must be kept down to a value which will be satisfactory for the grade of transmission concerned, further reduction being uneconomic. To do this the repeaters must meet definite load capacity requirements and modulation (non-linearity) requirements. The load capacity requirement is most conveniently specified in terms of the maximum single frequency sine wave power which a multi-channel amplifier must transmit without appreciable overloading. The modulation requirement pertains to the performance of the amplifier for impressed loads equal to or smaller than the load capacity, and specifies the allowable power in the modulation products resulting from such loads. Because of the numerous factors which affect these requirements, their determination is a rather complicated matter and the present discussion will be restricted solely to a determination of the load capacity requirement. The object is to determine this quantity as a function of N , the number of channels in the system.

The criteria ordinarily used for determining the load capacity of single-channel amplifiers are of little use here because of two funda-

* Presented at Great Lakes District Meeting of A.I.E.E., Minneapolis, Minn., September 27-29, 1939.

mental differences between single-channel and multi-channel systems. In the first place, the modulation produced in a single-channel amplifier depends only upon the input to that channel and occurs only when the channel is energized. In addition, the most important frequencies resulting from modulation fall directly back upon frequencies already impressed and the net effect appears as a distortion of the original input, rather than as noise. The situation is entirely different in a multi-channel system. In this case, the modulation products falling into one particular channel are in the main unrelated either to the impressed frequencies or to the volume of impressed speech in that channel. Thus it is no longer possible to think of the interference as distortion; the effect must rather be considered as that of a particular kind of noise whose level depends upon the load on the other channels of the system. For a given grade of service, the ratio of signal to noise must be much larger than the ratio of signal to modulation products resulting in distortion; thus it is to be expected that the non-linearity requirements will be more stringent for multi-channel operation than for single-channel operation.

The second fundamental difference between single-channel and multi-channel systems arises from the character of the load which each system must be designed to handle. A single-channel amplifier must be capable of handling one channel at the maximum volume normally expected. Inasmuch as the amplifier will be loaded only about one-fourth of the time, even in the busiest hour, and as the average impressed volume will be some 15 db below the maximum that must be provided for, the ratio of maximum to average load of such an amplifier is inherently very high. In a multi-channel system, however, the several channels will very rarely be heavily loaded simultaneously. There is thus a favorable diversity factor, increasing with the number of channels, and multi-channel amplifiers may accordingly be worked successfully at lower ratios of maximum to average load.

Occasionally, of course, there will be short periods of excessive loading during which the interchannel interference in multi-channel systems will rise above the value normally permitted. This sort of thing often occurs when it is desired to make economical use of facilities of any kind in common. In machine switching systems, for example, it is common practice to associate a large number of lines with a smaller number of switches and trunks. The number of switches and trunks provided is sufficient to ensure a satisfactory service, with a very small probability of requiring more facilities than are available. The multi-channel amplifier problem presents a situation identical in principle, though the methods of solution are necessarily very different.

The application of probability theory is evidently indicated as the method of attack.

Those characteristics of multi-channel amplifiers which are important to the problem will be described first. Then a description will be given of the variables which must be taken into account in computing load capacity. Finally, the combined effects of these variables will be determined on a statistical basis to establish the required load capacity as a function of the number of channels in the system.

CHARACTERISTICS OF THE MULTI-CHANNEL AMPLIFIER

At the present time, multi-channel systems of primary interest employ single sideband transmission; the carrier frequencies are largely suppressed and different amplifiers are used for the two directions of transmission. For such systems negative feedback amplifiers have outstanding advantages, particularly with respect to stability of gain and reduction of modulation effects, and are thus being used almost exclusively in present day multi-channel systems. The following discussion is related particularly to such systems, although many of the calculations are also applicable to less common types.

At light loads the principal modulation products in a negative feedback amplifier increase approximately as the square or the cube of the fundamental output power. Beyond a certain critical point, however, the modulation increases very rapidly and the total output of the amplifier soon becomes practically worthless for communication purposes. This critical point will be called the "overload" point. For most tube circuits it is either the point at which grid current begins to flow, or that at which plate current cutoff occurs. This point obviously defines the instantaneous load capacity.

Below the overload point the higher order modulation products are negligible in comparison with second and third order products, and the interference may be regarded as due to the latter sources alone. Beyond the overload point, however, the higher order products become important very rapidly and the resultant disturbances appear in most, if not all, of the channels. With given tubes, the interference below the overload point may be altered by changing the amount of feedback. The interference above the overload point, however, may be little changed in this way because of the rapid loss of feedback as the amplifier overloads. Accordingly, in designing an amplifier, the necessary load capacity may be determined solely by insuring that the output will rarely rise above the overload point, afterwards adjusting the amount of feedback so that the interference below the overload point will be tolerable. There are thus two problems which may be

handled separately, at least for negative feedback amplifiers, it being understood that the results are combined in the final design. As previously stated, only the load capacity problem will be considered in detail here but many of the methods used have been applied successfully to the interchannel modulation problem.

THE LOAD ON A SINGLE CHANNEL

The total load applied to a multi-channel amplifier varies rapidly between widely separated limits. A complete knowledge of the variations in the load applied to a single channel is necessary first; these variations arise from several recognizable causes which may be discussed separately.

Number of Active Channels

First of all, a single channel at a given instant may or may not be carrying speech; if not, it contributes nothing to the multi-channel load. A channel will be called "active" whenever continuous speech is being introduced into it; i.e., a channel is active during the time it is actually carrying speech power, and also during the short pauses that occur between words and syllables of ordinary connected speech. A channel is said to be "busy" when it is not available to the operator for completing a new call. Busy time is by no means all active time, for a busy channel is inactive during much of the time the connection is being completed, during pauses in the conversation, and finally during the time the other party is talking. The fraction of time during the busiest hour that a channel may be busy depends on the size of the group of circuits of which it is a member and on the methods of traffic operation. Measurements on circuits in large groups, made by Mr. M. S. Burgess, indicate that the largest fraction of the busiest hour that a channel may be active is about $\frac{1}{4}$. For channels in small circuit groups, this figure may become considerably smaller but it is unlikely that any probable increase in group size or improvement in operating practices will increase it appreciably. This figure, which will be represented by τ , may accordingly be taken as a conservative estimate of the limiting probability that a channel will be active in the busiest hour.

The number of channels that are active at a given instant in an N -channel system may be anything from zero to N . Inasmuch as the channels are independent, it is possible to write down at once the probability that exactly n of them are simultaneously active. This probability is

$$p(n) = \frac{N!}{n!(N-n)!} \tau^n (1-\tau)^{N-n}. \quad (1)$$

Talking Volumes

A second source of variation in the load on a given channel is that the impressed volume may have any value within rather wide limits when a channel is active. By "volume" is meant the reading of a volume indicator of a standard type. Its importance in the present problem arises from the fact that the volume is an approximate measure of the average speech power being introduced into the channel. Although some other instrument might give a better measurement of the latter quantity, only the volume indicator has been used sufficiently widely in the plant to give data on the distribution of average speech power per call under commercial conditions. The average speech power is dependent on the type of instruments, the character of the speech, and the time interval over which the average is determined. From an analysis of phonograph records of continuous speech it is found that the average speech power of a reference volume talker may be taken as 1.66 milliwatts, and the relationship between volume¹ and average power may be expressed by the following equation:

$$\text{Volume (db)} = \frac{10 \log_{10} \text{Average Speech Power in Milliwatts}}{1.66}. \quad (2)$$

This equation is based on the long average speech power. It will be understood that for purposes other than load rating computations, a different relation might be found more suitable.

The use of equation (2) to relate volume to average speech power is applicable to speech in a single channel. It is convenient to refer to a quantity related in the same way to the total average power contributed by a number of channels as the "equivalent volume."

The single-channel volumes on commercial circuits are conveniently measured at the transmitting toll test board, which will be taken as a point of "zero transmission level." Henceforth it is assumed that there is no gain or loss between this point and the output of the amplifier, so that the latter is also a point of zero transmission level. While this will seldom be the case in an actual system, the necessary change in the load capacity is easily computed. The volumes at this point are found to be distributed approximately according to the "normal" law; that is, the probability that the volume will be between V and $V + dV$ is given by

$$p(V)dV = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(V-V_0)^2}{2\sigma^2}} dV. \quad (3)$$

¹ Subsequent to the preparation of this paper, a new volume indicator was standardized for use in the Bell System. With the new volume indicator, volume is expressed in vu , $+8vu$ being approximately equal to reference volume (0 db) as used herein.

For calls on typical toll circuits, the best present values for the parameters are $V_0 = -16.0$ db and $\sigma = 5.8$ db. These parameters depend, of course, upon the character of the local plant and upon the habits of telephone users, and changes in either will affect their values. Curve *A* of Fig. 1 shows this distribution of talker volumes at a point of zero transmission level. Curve *B* of Fig. 1 is the talker volume distribution used for load rating computations when a particular amount of peak amplitude limiting occurs in the terminal equipment. This will be discussed later. Although the mean volume is V_0 , the

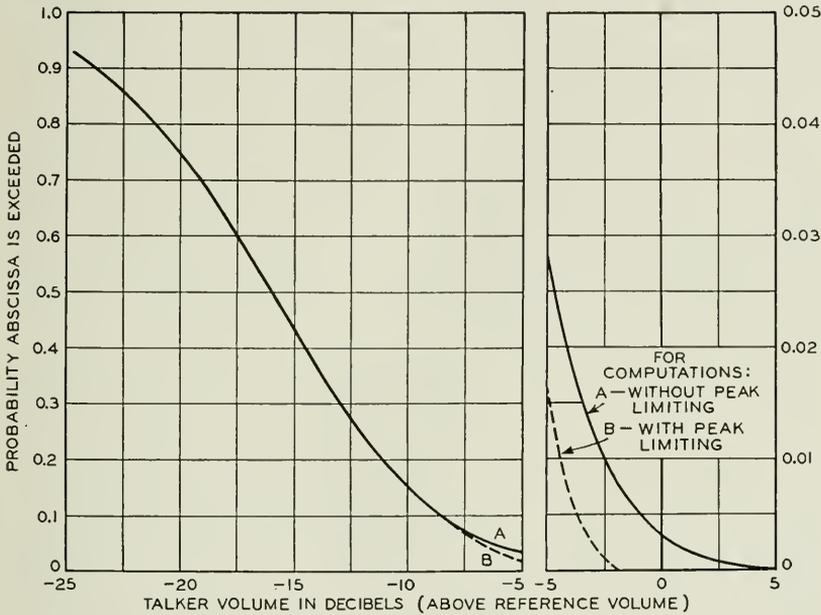


Fig. 1—Talker volume distribution.

volume corresponding to the mean power of the distribution (3) is equal to $V_0 + .115\sigma^2$, as may be seen by converting the volume scale of the distribution to power ratios, averaging, and reconvertng the average to volume in db. For the values of the parameters given above, $V_0 + .115\sigma^2 = -12.1$ db.

Instantaneous Voltage Distribution

Finally, the voltage in an active channel fluctuates widely even at constant volume. Not only the differences between successive syllables and the differences between vowel sounds and consonants, but also the fine structure of single sounds, are important in this connec-

tion. The total voltage impressed on the amplifier is the quantity which determines whether or not it will overload, and the phases as well as the amplitudes of the frequency components in the several channels must be considered in determining this. It is most convenient for analysis to work directly with instantaneous voltages of speech, the frequency of occurrence of the magnitudes being expressed in the form of a distribution function.

This distribution function has been measured by Dr. H. K. Dunn, using apparatus which measures 4 samples per second of the instantaneous voltage out of a commercial subset and typical loop. By operating the apparatus until about 1000 successive samples have been measured, usable distribution curves of instantaneous voltage are obtained; this is readily checked by making repeated runs comprising the same number of samples on speech recorded on high quality phonograph records. It is, of course, known that commercial transmitters have considerable asymmetry as regards positive and negative voltages but the poling referred to the toll board is expected to be random. As the measurements were considerably simplified by doing so, it appeared desirable to average out this asymmetry by arranging a linear rectifier ahead of the sampling apparatus to obtain equal samples of positive and negative voltages.

Such measurements have been made for a number of different talkers, different commercial subsets, and different volumes, with the speech input held at substantially constant volume in each test. The various subsets now in commercial use all give essentially the same distribution curve. The resulting distributions, if they are considered as functions of the ratio of instantaneous to rms voltage, are also nearly independent of the speech volume at the subset. Specifically, the only important effect of volume is that which may be ascribed to amplitude limiting in the transmitter; i.e., to the fact that the transmitter itself has a limited load capacity. However, this effect does not appear until the volume is 10 db or more above the mean of the volume distribution curve, and is only of importance for talkers at still higher volumes. For all lower volume talkers, the instantaneous voltage distribution may be considered as the same for all volumes when expressed as a ratio of instantaneous to rms voltage. The cumulative distribution curve of the quantity E/U , where E is the rectified instantaneous voltage and U the rms voltage, is shown by the curve $n = 1$ of Fig. 2.

Voltage Limiting

While this curve of Fig. 2 is accurate for the bulk of the talkers, it changes for the high volume talkers who overload the subset trans-

mitters. It is also the custom to provide a certain amount of amplitude limiting in each channel by suitable circuit design of the channel terminal equipment. This limiting alters the shape of the instantaneous voltage distribution curve for a range of voltages below the maximum, the extent of the modification depending on the talker volume and the characteristics of the limiting device. Its effect on the load capacity will be considered later.

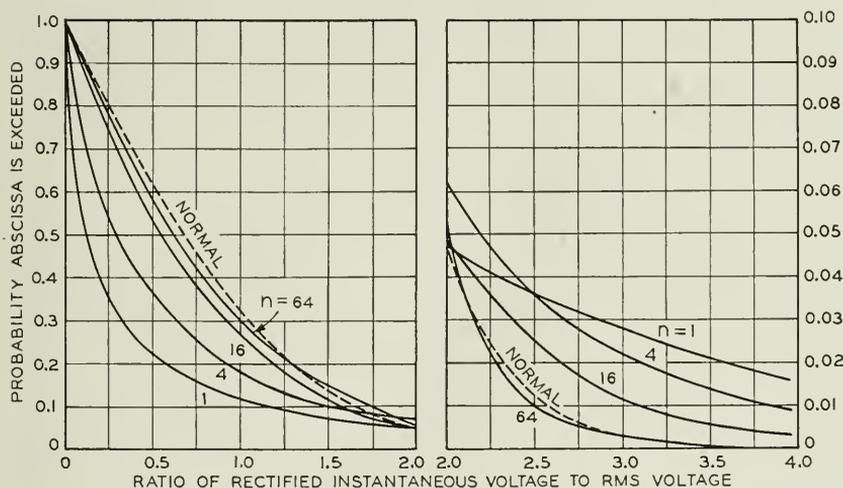


Fig. 2—Instantaneous voltage distributions for n talkers.

MULTI-CHANNEL INSTANTANEOUS VOLTAGE DISTRIBUTIONS

The number of variables with which it is necessary to deal makes the general load capacity problem rather a complicated one. The analysis will be easier to follow if the effects of the different variables are taken up one at a time, thus building up a complete theory in successive steps. To do this, it is advantageous to start with a case so simplified that it rarely, if ever, occurs in ordinary practice; i.e., that in which the volumes in all the channels are regulated to a common constant value, and in which the number of *active* channels is also kept constant. For this condition, it is necessary to consider only the effects of the distribution of instantaneous voltages in each channel. This distribution curve is the same for all of the channels, since all are at the same volume, but the voltage in any channel at a particular instant is entirely independent of the condition of the other channels.

Overload Expectation

The total voltage impressed on the amplifier by a number of channels at a given instant is the sum of the instantaneous voltages in the

separate channels. Since disturbances will be produced in many of the channels when the applied voltage goes beyond the overload point, it will be useful to know the fraction of the time that this may be expected to occur; this fraction will be called the overload expectation and denoted by ϵ . It is important to notice that this quantity ϵ is not necessarily the fraction of the time during which the performance of the amplifier will be unsatisfactory. This might perhaps be the case for a device having an instantaneous cutoff characteristic, but for an ordinary amplifier the time constants (among other things) affect the results of overloading. The interpretation of the overload expectation will be discussed further later; consideration must be given first to how it is obtained.

The n -Channel Voltage Distribution

The load in each channel is applied at voice frequency to the input side of a modulator, the voice frequency instantaneous voltage distribution being as shown by the curve $n = 1$ of Fig. 2. The overloading of the amplifier is determined, however, by the distribution of the sum of n such voltages after each has been shifted by the modulator to the appropriate carrier frequency, one side-band being suppressed. It may be shown that if the phases of the various components of the voice frequency input were random, the distribution of instantaneous voltage at side-band frequency would be identical with that measured at voice frequency. It is known, however, that the phases at voice frequency are not entirely random, and there may thus be differences between the two distributions. The results of a number of tests bearing upon this point indicate that any error resulting from the use of the distribution measured at voice frequency will be small for systems of few channels, and will rapidly disappear as the number of channels is increased.

Theoretically, the resultant n -channel voltage distribution can be derived from the single-channel distribution by straightforward analytical methods; in the present case, however, expression of the result in useful form is very difficult because of the form of the single-channel curve. This difficulty might be resolved by using graphical or numerical methods, as applied later to the volume distribution curves; fortunately, the fact that the voice frequency voltage distributions may be used throughout permitted the resultant n -channel distributions to be obtained much more easily. Since the addition of voltages from the several carrier channels does not depend materially upon the frequencies at which the channels appear in the system, the addition of n channels at voice frequency will give the desired n -

channel distribution directly. Mr. M. E. Campbell effected this addition by the use of phonograph records, the n -channel distributions being determined by means of the instantaneous voltage sampling apparatus previously mentioned.

As material for this process, 16 high-quality phonograph records were made of the outputs of commercial subsets through representative subscriber loops. Both male and female voices were used. The speech was furnished by reading magazine stories containing considerable conversational material, due precautions being taken that the volume on each record was substantially constant throughout. A calibrating tone was cut on each record to enable it to be played at any desired volume and most of the volumes recorded were well below the point at which the transmitter began to act as a voltage limiter.

These individual records were then combined in groups of four, with all records adjusted to the same volume by means of the calibrating tones, and re-recorded. Several such 4-voice records were made; by combining them again in the same way, 16-voice records and finally 64-voice records were obtained. The instantaneous voltage distributions were measured before and after each re-recording to insure that the recording process introduced no errors. A few minor discrepancies were found, but all were small enough to be disregarded. Each single-voice record appeared several times in a 64-voice record, but since the phases of its different appearances were random, this had no appreciable effect on the resultant voltage distribution. This was verified by comparing the voltage distributions of the various possible 16-voice combinations. By this process n -channel voltage distributions were obtained for $n = 1, 4, 16$ and 64 .

These distributions, together with a normal curve, are shown in Fig. 2 in cumulative form. To show the curves conveniently to the same scale, it has been necessary to plot for each case not the distribution of E , the rectified instantaneous voltage, but that of E/U , where U is the rms voltage. The rms voltage, it will be remembered, is directly related to the equivalent volume by equation (2). The figure shows clearly the gradual transition from the single-channel distribution to the normal one for large n , and also indicates that for 64 active channels the curve is normal within the precision of the measuring apparatus. Hence, the normal distribution may justifiably be used for any value of $n > 64$.

Further significance is accorded the above data by plotting the ratio of the voltage exceeded a fraction ϵ of the time to the single-channel rms voltage, as a function of the number n of active channels, for several fixed values of ϵ . From the data given, points on such

curves can be obtained for $n = 1, 4, 16, 64$; furthermore, the fact that the distribution for $n > 64$ is normal permits drawing the asymptote for large values of n . The points read from Fig. 2 and replotted in this way give the full lines shown in Fig. 3.

In order to make practical use of these curves, it is necessary to know what value of ϵ corresponds to satisfactory performance of the amplifier. Experiments have been conducted on a number of different multi-channel amplifiers, each loaded by various numbers of active channels all at the same volume. It has been found that for low enough values of ϵ , no audible disturbance is produced but that as ϵ is increased by increasing the load on the amplifier, the disturbance falling into a channel not energized increases rapidly to a large value. Two different amplifiers having the same computed load capacity may show noticeable differences in performance in this respect when subject to identical fixed loads of the type being considered, thus indicating the influence of circuit design on the value of ϵ . In general, however, the allowable values of ϵ measured for all of the amplifiers that have been tested lie in a relatively narrow band on either side of the curve for $\epsilon = 0.001$. The broken curve of Fig. 3 represents the

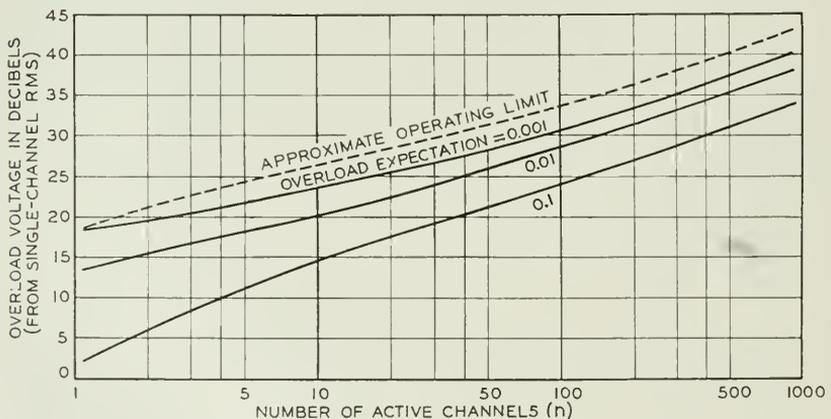


Fig. 3—Overload voltage for n active channels.

approximate upper limit of the observations, extrapolated parallel to the $\epsilon = 0.001$ curve above $n = 14$. It is possible that some amplifiers would overload even if operated in accordance with this curve, but for the great majority of amplifiers of types thus far tested the operation would be satisfactory, with perhaps a small margin.

Multi-Channel Peak Factor

It is useful at this point to introduce the concept of "multi-channel peak factor," which is defined as the limiting ratio of the overload

voltage to the rms voltage for a given number of active channels at constant volume. The ratio of the overload voltage for n active channels to the rms voltage of one active channel is given directly by the broken curve of Fig. 3, and the rms voltage for n channels is simply \sqrt{n} times that for one channel. A simple computation then gives the multi-channel peak factor. This is plotted in Fig. 4 as a function of the number of active channels n . The reduction in multi-channel peak factor as the number of active channels increases reflects the transition from the single-channel distribution curve to the normal curve, as depicted in Fig. 2.

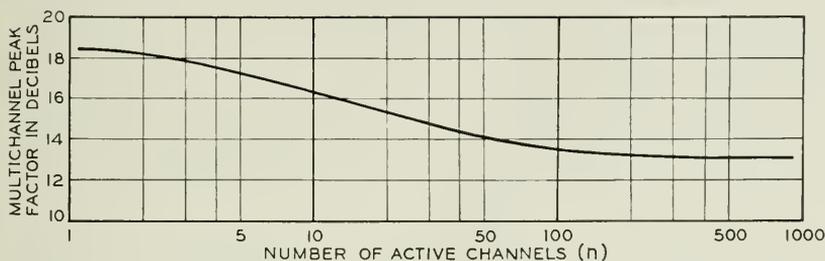


Fig. 4—Multi-channel peak factor for n active channels.

THE DISTRIBUTION OF EQUIVALENT VOLUME

The multi-channel peak factor deals only with the effects of changes in the instantaneous voltages of the channels, all other variables being fixed. It is next necessary to extend the treatment to include the effects of the other load variations that occur in practice—those in number of active channels and in channel volumes. It is important, first of all, to notice that the instantaneous-voltage variations occur very rapidly, while changes in the other two quantities are, in comparison, very slow. In the experiments described above, the loads were so fixed that the equivalent volumes could be changed only by changing the operating transmission level of the amplifier; in practical cases the amplifier transmission level is kept fixed, but the equivalent volume is constantly changing because of changes in number of active channels and in channel volumes.

The amplifier is thus loaded with a constantly changing equivalent volume but because of the great difference in the time-scales of the two classes of variations the load may be regarded as a succession of equivalent volumes, each constant for a small interval of time that nevertheless is long enough to include a representative sample of the resultant instantaneous voltage distribution. If the distribution function for equivalent volume is computed, and then corrected by the

multi-channel peak factor, the fraction of such intervals during which the amplifier will be unsatisfactory from the standpoint of overloading may be determined. For a particular amplifier, the operating transmission level must be so chosen that this fraction will be small enough to make any adverse effects on transmission unimportant. For systems of very many channels the proper value of this fraction is probably about 1 per cent. During the busiest hour, this corresponds to 36 seconds during which audible interference *may* occur and as this will be broken up into many very short intervals, the total effect should be slight. For systems of very few channels, the equivalent volume may reach objectionably high values during these intervals and it might be necessary to make this fraction smaller than 1 per cent to secure good performance. For illustrative purposes, the 1 per cent figure will be used in what follows without implying that it may not need alteration in some cases. The methods used are applicable no matter what value is chosen for the fraction of time overloading is permitted.

Controlled Volumes

As the simplest case to which the above procedure may be applied, and one that may occasionally be of practical interest, consider a commercial system with all the channels controlled to the same volume. If there are N channels in the system, the probability that exactly n channels will be active at any given time is given by equation (1), with $\tau = 0.25$. By computing the value of $p(n)$ for all values of n , and taking the cumulative sum, the value of n which makes the sum 0.99 (or the next greater n) is readily determined. This determines the number n of active channels that is exceeded 1 per cent of the time. A plot of these values of n is given by the curve of Fig. 5, as a function of N , the number of channels in the system. For small values of N this curve has been drawn in a manner to smooth out the steps introduced because n must of necessity be an integer and when the value of n read from the curve is not an integer, the next higher integral value is to be used. It is of interest to compare this curve with the two straight lines of the figure. The lower straight line represents the asymptote for sufficiently large N and the upper straight line is for the condition where all channels are active simultaneously ($n = N$).

The average power for n channels is n times that of one channel, and the equivalent volume expressed in db is $10 \log_{10} n$ above that in one channel. The equivalent volume may thus be computed as a function of n , and by means of Fig. 5, as a function of N . Curve *A* of Fig. 6 shows the values of equivalent volume so determined as a function of N , the number of channels in the system; it applies specifically to the

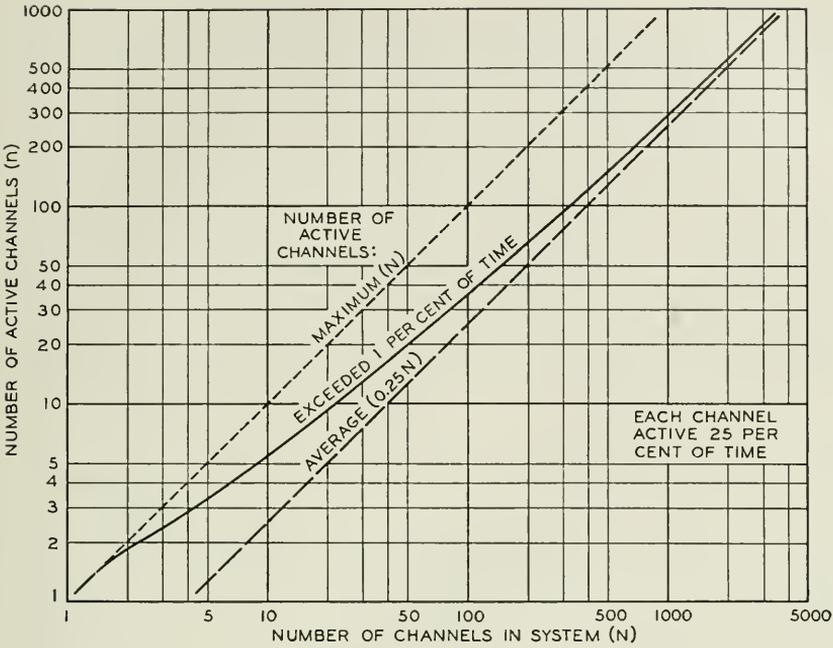


Fig. 5—Number of active channels as a function of the number of channels in the system.

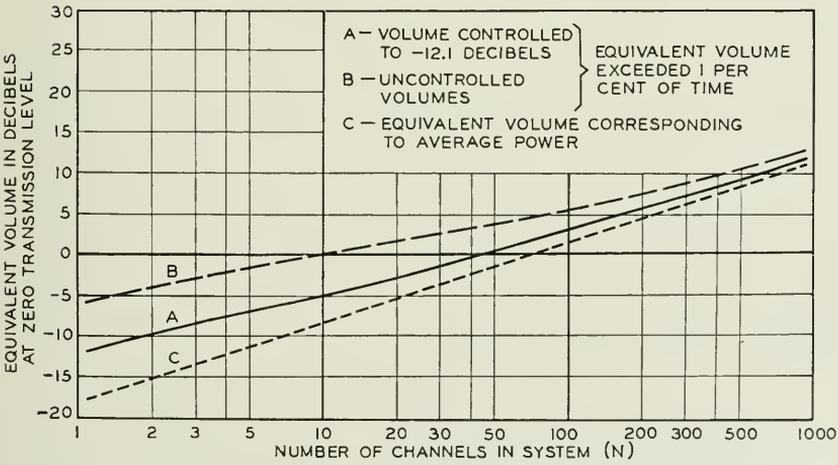


Fig. 6—Equivalent volume for systems of N channels.

case where the volume of each of the active channels is controlled so as to be 12.1 db below reference volume. The choice of this particular volume is purely arbitrary, but it corresponds to the average power of the single talker volume distribution.

The equivalent volumes given by curve *A* of Fig. 6 are a measure of the average power of the N channels, as computed by means of equation (2). To determine the required instantaneous load capacity of the system, the average power must be corrected by the multi-channel peak factor which is read directly from Fig. 4, using for the number of active channels the values read from Fig. 5.

For design purposes, it is more convenient to use the rms power of the single frequency test tone whose peak value represents the

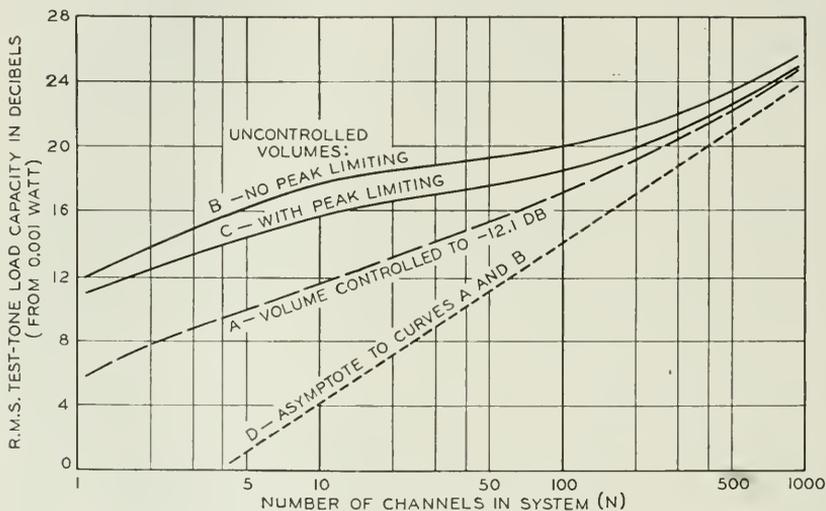


Fig. 7—Load capacity for systems of N channels.

instantaneous load capacity. As the ratio of the peak to rms power of a single frequency tone is 3 db, this test power is obtained by subtracting 3 db from the instantaneous load capacity. This required test-tone capacity is plotted as a function of N in curve *A* of Fig. 7, which gives the output capacity required for an N -channel system with volume control as specified above.

Uncontrolled Volumes

For systems in which volume control is not used, the application of this procedure becomes more involved. To study this more general case, it is convenient first to interchange the conditions of the preceding section, letting the number of active channels be fixed at any value n

and examining how the distribution curve of equivalent volume may be obtained for this fixed number of channels. The relation between volume and average speech power given in equation (2) may be rewritten for this case in the form

$$V_i = 10 \log_{10} \frac{W_i}{W_0} \text{ db,}$$

where $W_0 = 1.66$ milliwatts, W_i is the average speech power in milliwatts, and V_i is the volume in db for any one of the active channels, all at a point of zero transmission level.

Likewise, the relation between equivalent volume and average speech power for n active channels is given by the expression

$$V = n\text{-channel equivalent volume} = 10 \log_{10} \frac{\sum W_i}{W_0} \text{ db.}$$

Since the distribution of the channel volumes V_i is known and the volumes of the various channels are independent, the straightforward procedure to obtain the distribution of the n -channel equivalent volume V would involve the following steps: (1) the obtaining of the distribution function of W_i by a transformation of that of V_i ; (2) the calculation of the distribution function for the quantity $Y(n) = \sum_1^n W_i$; (3) the transformation of the $Y(n)$ distribution to that of V by inverting the process used in step (1).

The difficulties in this process are all in the second step, where, having given $p_1(W)$, the distribution of average powers for a single channel, it is required to obtain $p_n(Y)$, the distribution for n active channels, with Y defined in terms of W by the relation given immediately above. The formal solution requires the evaluation of integrals of the following type:

$$p_n(Y) = \int_0^Y p_{n-k}(W) p_k(Y - W) dW.$$

By successive calculation of such integrals for $n = 2, 4, 8 \dots$, taking k each time equal to $n/2$, the required distributions may be obtained for the necessary range of values of n .

As in the case of the instantaneous voltage distributions, it has not proved feasible to perform the integrations analytically. It was necessary to resort to numerical evaluation of these integrals; by combining the transformations in steps (1) and (3) with the process

of evaluating the integral, the process was somewhat shortened. In this way equivalent volume distributions have been obtained for $n = 2, 4, 8, 16 \dots$; needed points on the distribution curves for intermediate values of n are obtainable by interpolation.

The accuracy of such a process depends upon the number of division points used in the numerical integration and this as a practical matter must be kept fairly small. When the process must be repeated many times, the errors introduced at each step may accumulate and lead to inaccuracies for large n . It is thus desirable to have some control over the accuracy other than by repeating the calculation with a larger number of division points. This is provided by calculating the moments of $p_n(Y)$ from those of $p_1(W)$ without the use of numerical integration.

The moments S_k of $p_1(W)$ are defined by

$$S_k = \int_0^{\infty} W^k p_1(W) dW,$$

and the moments $T_k^{(n)}$ of $p_n(Y)$ similarly. By the use of the semi-invariants of Thiele,² it may be shown that

$$\begin{aligned} T_1^{(n)} &= nS_1, \\ T_2^{(n)} &= nS_2 + n(n-1)S_1^2, \text{ etc.} \end{aligned}$$

By comparing the moments of the distributions obtained by numerical integration with those calculated in this way, and making occasional minor alterations in the curves to bring the first and second moments into agreement, assurance was obtained that all the distributions used are reasonably accurate, with no accumulation of error as n becomes large.

Examples of the cumulative distribution curves of equivalent volume for 1, 4, 16 and 64 active channels are given in Fig. 8. The decrease in standard deviation which occurs as n increases is of interest for it indicates how the fluctuations in load due to talker volume variations are reduced by combining a large number of channels in one system.

Having now n -channel equivalent volume curves for a range of values of n , the resultant equivalent volume curves may be calculated when the restriction to a fixed number n of active channels is removed. Let $p_n(V)$ denote the probability that, with n channels active, the equivalent volume lies between V and $V + dV$ and let $p(n)$ denote the probability that just n channels *will* be active. Then the total proba-

² T. N. Thiele, "The Theory of Observations," 1903; reprinted in *Annals of Mathematical Statistics*, Vol. 2, 1931. See especially Sections 22, 29.

bility that the equivalent volume will be between V and $V + dV$ is given, for an N -channel system, by

$$p(V) = p(1)p_1(V) + p(2)p_2(V) + \dots + p(N)p_N(V).$$

The $p_n(V)$ are given by equivalent volume curves such as those in Fig. 8 and the $p(n)$ by equation (1). Examples of curves thus computed are given in Fig. 9, which shows the equivalent volume distributions at a point of zero transmission level for 3, 12 and 240 channel systems. The equivalent volume that is exceeded 1 per cent of the time, read from such curves, is plotted as curve B of Fig. 6.

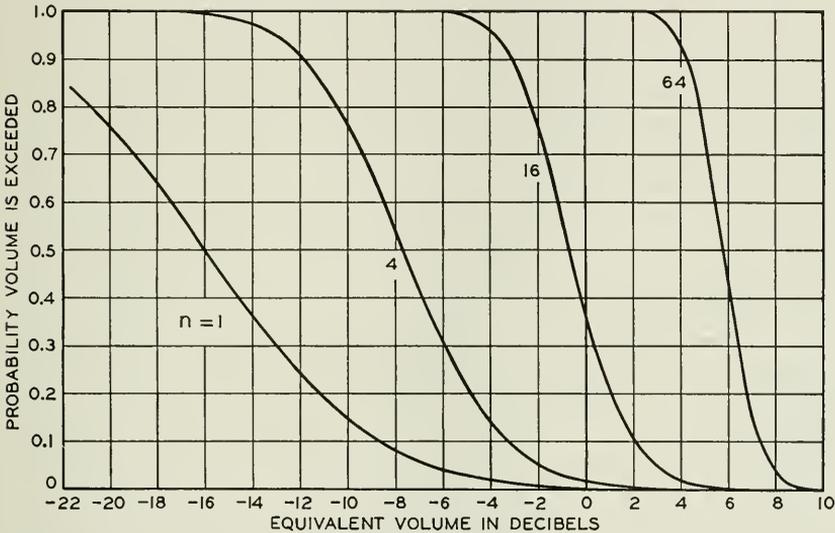


Fig. 8—Equivalent volume distributions for n active channels.

This curve gives, for any number of channels having uncontrolled volumes, the equivalent volume which will be exceeded just 1 per cent of the busy hour. To obtain the necessary load capacity, this must be corrected for the multi-channel peak factor. In the controlled volume case, for a given number N of channels in the system, there was no difficulty in deciding the value of n , the number of simultaneously active channels, for which the multi-channel peak factor should be taken. Now, however, there is no unique relation between equivalent volume and the number n ; in addition, the multi-channel peak factors were measured with all n channels at the same volume, which represents a condition rarely holding on a system without volume control. It is apparent, however, that in the majority of cases in which the equivalent volume approaches values on curve B of Fig. 6, the number

of simultaneously active channels will be greater than the average number $N\tau$ of active channels. Since the multi-channel peak factor decreases as n increases, the peak factors for $n = N\tau$ active channels may be safely used. A more detailed analysis, feasible only for very small systems but avoiding the use of this approximation, shows that its effect is small and tends to give load capacities slightly higher than actually required, but the difference diminishes rapidly as the size of the system is increased.

For the uncontrolled volume condition, therefore, the multi-channel peak factors are read from Fig. 5 for values of $n = N\tau$. They are added to the equivalent volumes obtained from curve *B* of Fig. 6, and

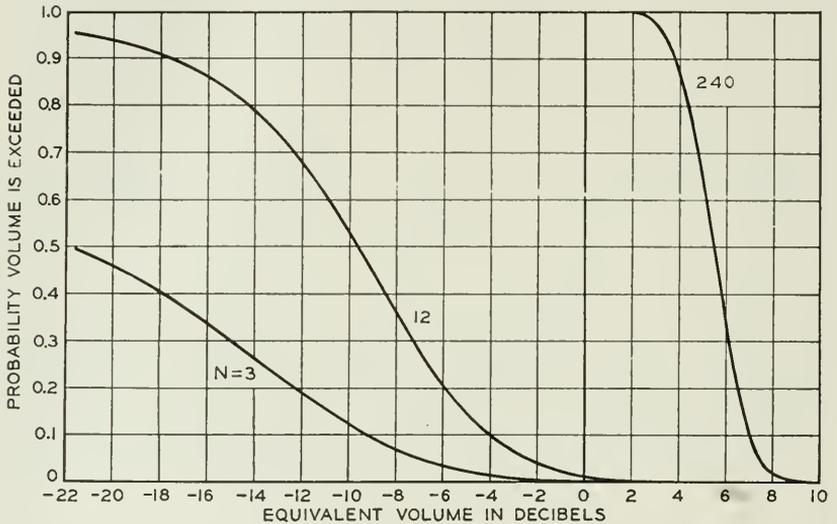


Fig. 9—Equivalent volume distributions for systems of N channels.

reduced to single frequency power as previously described for the volume controlled case. Curve *B* of Fig. 7 is obtained in this manner and shows the load capacity required in an amplifier for an N -channel system in which the volumes of each channel are distributed in accordance with curve *A* of Fig. 1. The load capacity which is approached asymptotically as N increases indefinitely is represented by curve *D* of Fig. 7.

The load capacities given by Fig. 7 are valid only for systems for which the basic single-channel data apply. As these may not hold in specific cases, and may be subject to modification in the future, estimates of the effects of small changes in these data are useful. These effects cannot be described simply for moderate numbers of

channels but for large numbers of channels the effects are readily estimated from the change in the location of the asymptote shown on Fig. 7. The equation of this asymptote is as follows:

$$L = 10 \log_{10} N\tau + (V_0 + .115\sigma^2) + MPF + P_0 - 3 \text{ db,}$$

where L = test tone load capacity,

MPF = asymptotic multi-channel peak factor,

P_0 = long average power of a reference volume talker in db
above .001 watt.

The other quantities are as defined before.

PEAK VOLTAGE LIMITING

The curves referred to in the preceding discussion have so far neglected the effects of peak voltage limiting in the transmitters and in the channel terminal equipment. Fundamentally, the effect of such limiting is to modify the distribution of instantaneous voltages in the individual channels. The extent of the modification, however, depends on the volume. For single-channel systems it is obvious that the improvement in load capacity due to limiting will be substantially equal to the reduction in the maximum peak voltage. For a large number of channels the improvement will approach the reduction in the rms voltage per channel. An approximate method of accounting for these complicated reactions is to consider that peak voltage limiting modifies the upper end of the single-channel volume distribution. Strictly the amount of such modification is a function of the number of channels as well as of the characteristics of the limiters. Curve *B* of Fig. 1 represents a compromise between the different effects which is believed to give reasonably accurate results for both small and large numbers of channels for the limiting characteristic of present terminals.

With the talker volume distribution modified in accordance with curve *B* of Fig. 1, computations of the load capacity with voltage limiting present may be made in a manner identical with that previously described. Curve *C* of Fig. 7 shows the results obtained for this amount of limiting.

All of the load capacity curves of Fig. 7 are based on the equivalent volume which would be exceeded 1 per cent of the time, irrespective of the number of channels in the system. Where voltage limiting is used, it appears reasonable to consider this percentage as fixed because the action of the limiters serves to restrict the range of voltages above the overload point, thus reducing the severity of any overloading effects. When there is no limiting, and particularly

for a small number of channels, the range of overloading voltage is not so restricted and overloading effects may become undesirably severe during the 1 per cent of time when the overload voltage is exceeded. If voltage limiting is not provided in some form, it may be important to reduce the percentage of time during which overloading may occur for small numbers of channels. This is a matter to be determined by experience and, if necessary, would require modification of curve *B* of Fig. 7 in the direction of requiring more load capacity for a small number of channels, thus increasing the spread between curves *B* and *C*.

OPERATING MARGINS, ETC.

The curves which have been given for output capacity versus number of channels apply to a single amplifier, or to a system in which all amplifiers are identical and work at the same output level without appreciable impairment of overall performance. In practice, the number of amplifiers in tandem in a long system may be very large and problems of equalization and regulation may make it difficult to maintain exactly the same level conditions at all amplifiers. In addition, aging of tubes, and other effects will introduce some impairment. It is important, therefore, to allow a margin for these effects in the design of an amplifier for a multi-channel system. The proper margin is essentially a matter of system design and it is often economical to build a liberal margin into the amplifiers in order to allow greater latitude and economy in the design of equalizing and regulating arrangements.

In addition to the speech loads, there are also impressed on the amplifiers various signaling and pilot frequencies, carrier leaks, etc. It is not always possible in practice to make these negligibly small and the load capacity requirements must be corrected to allow for their presence. Multi-channel telephone systems are also required to transmit other types of communication circuits, such as program channels and voice-frequency telegraph systems, superposed on one or more telephone channels. Modifications of the methods applied to speech loads may readily be made to determine the effect of these on the amplifier load capacity.

ACKNOWLEDGMENT

Many members of the Bell Telephone Laboratories, in addition to those mentioned in the text, have contributed to various phases of this work. The authors take this opportunity to acknowledge their indebtedness to these colleagues, and in particular to Dr. G. R. Stibitz, who first developed the theoretical approach here used.

The Quantum Physics of Solids, I

The Energies of Electrons in Crystals

By W. SHOCKLEY

It is proposed to make this paper the first of a series of three dealing with the quantum physics of solids. This one will be concerned with the quantum states of electrons in crystals. The discussion will commence with an introductory section devoted to the failure of classical physics to account for phenomena of an atomic scale. Next, the quantum theory of electrons in atoms will be discussed, together with the resultant explanation of the structure of the periodic table; this is designed to illustrate the meaning of various quantum mechanical ideas which are important in understanding solids. Furthermore, much of the detailed information about atomic quantum states of particular atoms will be needed in the later discussion of the properties of certain solids. As an introduction to the modification of the quantum states occurring when atoms are put together to form a crystal, a short section will be devoted to structure of diatomic molecules. The next section will be concerned with quantum states for electrons in crystals. Whereas in an atom there are a series of isolated energies possible for an electron (corresponding to the various quantum states), in a crystal there are bands of allowed energies. This concept of energy bands is essential to the theory of crystals in much the same way that the concept of energy levels is essential to that of atoms. In terms of energy bands, the energy holding crystals together can be interpreted on a common basis for a wide variety of crystal types. This will be followed by a brief description of various crystal types and by a discussion of thermal properties in which the smallness of the electronic specific heat will be shown. The last section will be devoted to a discussion of para and ferromagnetism on the basis of the energy band picture.

In the second paper, problems connected with electric currents and the motion of electrons through crystals will be discussed. This leads to the concept of the Brillouin zone which is complementary to that of the energy band, the two together forming the basis for discussing the quantum states of electrons in crystals.

The third paper of the series will contain a comparison between theory and experiment for the alkali metals, the principal emphasis being placed upon the physical picture of the state of affairs in these simple metals.

INTRODUCTION

“THE parts of all homogeneal hard Bodies which fully touch one another, stick together very strongly . . . I . . . infer from their Cohesion, that their Particles attract one another by some Force, which in immediate Contact is exceeding strong, at small distances

performs the chymical Operations above mention'd, and reaches not far from the Particles with any sensible Effect. . . . There are therefore Agents in Nature able to make the Particles of Bodies stick together by very strong Attractions. And it is the Business of experimental Philosophy to find them out." But it was not destined for experimental philosophy to finish the business which Sir Isaac Newton set for it in the above words¹ until two centuries had elapsed. Only since the advent of quantum mechanics have scientists had laws capable of explaining the cohesive forces of solid bodies and predicting their numerical magnitudes. The new laws were developed first in order to explain the behaviors of independently acting atoms but, as we shall see, they are laws capable of extension to systems containing large numbers of atoms and thus to solid bodies. The fact that a solid body remains a solid body, resists being pulled apart, and exerts the cohesive forces of which Newton wrote, is explained by showing from theory that atoms packed together in a solid are in a state of low energy, and to change the state requires the expenditure of work. In this paper we shall describe how the quantum mechanical concepts developed for isolated atoms are applied to interacting atoms and lead to methods of calculating the energies and forces binding atoms together in crystals.

A crystal is a regular array of atoms. The regularity of this atomic array is frequently exhibited in the macroscopic appearance of the crystal. A crystal of potassium chloride—sylvine—is a good example (Fig. 1A). The natural growth faces of the crystal are parallel to planes passing through the atoms, which are arranged in the microscopic array pictured in Fig. 1B. It is evident that the microscopic arrangement of the atoms in the crystal is one of its most basic features. In sylvine the atoms are arranged on the corners of cubes in an alternating fashion. The arrangement of the atoms in the crystal is called a "lattice." Sodium chloride—rock salt—has the same arrangement as sylvine and the type lattice pictured in Fig. 1B is known as a "sodium chloride lattice." The distance between atoms in a given lattice is specified by giving the value of the "lattice constant," which for a cubic crystal is defined as the distance between like atoms along a line parallel to a cube edge. Lattice constants are usually expressed in angstroms; 1 angstrom $\equiv 1\text{A} = 10^{-8}$ cm. The lattice constant of sylvine, designated by "*a*" in Fig. 1B, is 6.28A. Figure 1C suggests how a large number of atoms, arranged as in Fig. 1B, produce the shape of the crystal photographed for Fig. 1A. Studies of the directions of the natural growth faces and cleavage faces of crystals are

¹ "Opticks" 3rd ed., 1721, p. 363.

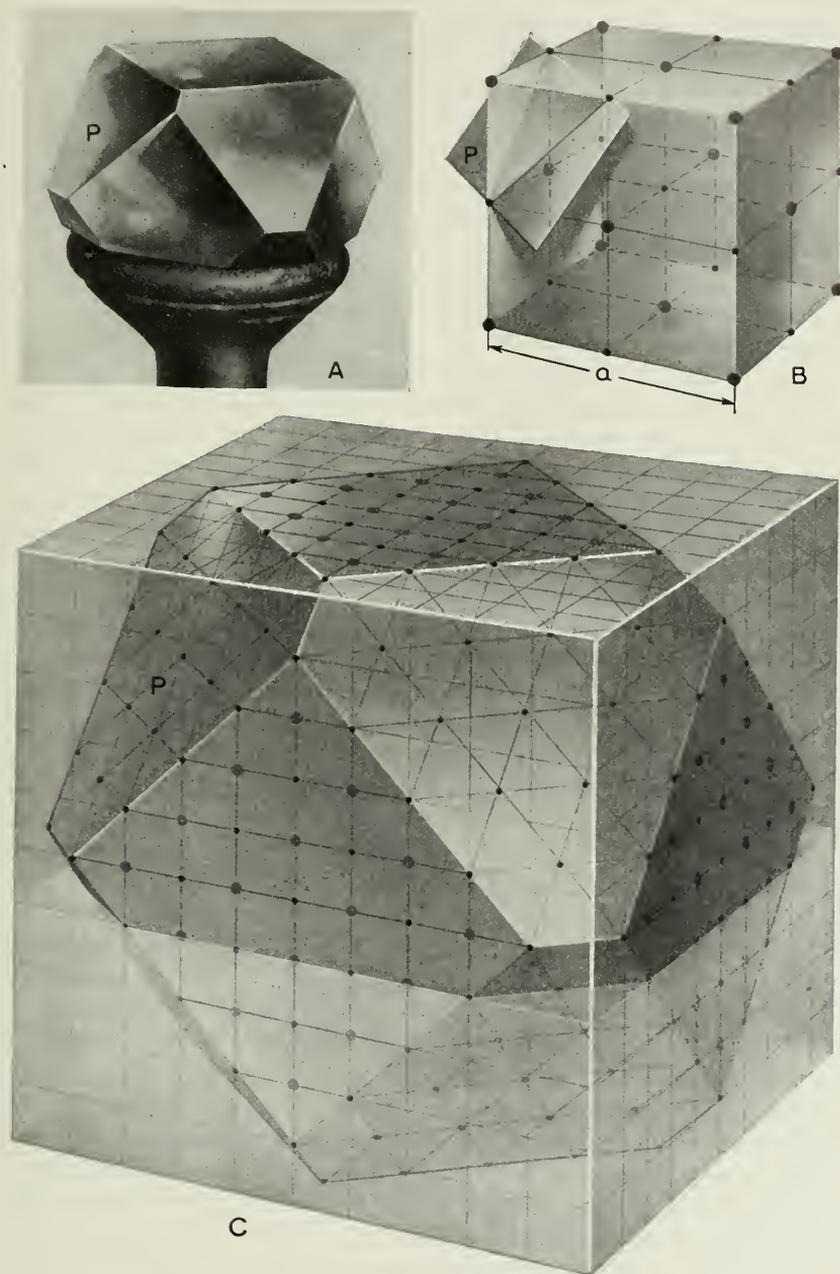


Fig. 1—Crystal structure.

- A. Macroscopic appearance of a crystal; retouched photograph of a sylvine crystal.
B. Microscopic arrangement of atoms in crystal showing natural planes.
C. Large number of atoms arranged as in B to show formation of A.

primarily of importance as an aid in classifying and identifying minerals; and although they do give some information about the arrangement of the atoms in planes within the crystal, the information is too meager to permit a determination of the microscopic structure. The latter can be deduced by the methods of x-ray diffraction. X-rays are light waves of very short wave-length and they are diffracted from crystals in much the same way as light is diffracted from a ruled grating. From studies of x-ray diffraction patterns, the arrangements of atoms in a large number of crystals have been determined. Exceedingly strong forces act to hold the atoms in these arrangements and, by application of the laws of quantum mechanics, we shall try to find them out.

There is now no question that the elementary building blocks of the material world are primarily electrical and of two sorts.² The negative particles, electrons, are all alike and have the same charge $-e$ and the same mass m ; the positive particles, atomic nuclei, are not all alike and may differ in charge and mass from one another. The positive charge is always some integral multiple, Z , of the fundamental charge e and we shall not be concerned with the mass except to say that it varies upwards from about 2,000 times the electron mass. An atom of a chemical element consists of one nucleus surrounded by enough electrons to neutralize its charge; all atoms of a given chemical element have the same nuclear charge, Z , which is appropriately known as the "atomic number"; atoms having the same nuclear charge but different masses are called "isotopes"; their chemical behaviors are slightly different, so that it is possible by chemical processes to separate one isotope of a chemical element from the others, but this difference is so slight that we can neglect it here. An atom, then, consists of a number of electrons circulating about and attracted by the nucleus, which, by virtue of its relatively great mass, is effectively an immobile center for their motions. A simple molecule consists of an assemblage of a few such atomic systems and a crystal of an immense number. The fundamental problem of atomic mechanics—which is now solved quite satisfactorily but not yet perfectly—is to find the laws governing the motion of these particles.

The necessity of finding such laws is made most apparent by considering the failure of the older laws of "classical mechanics," Newton's laws. These laws were satisfactory for dealing with large bodies—but not perfect; for, as is well known, they are approximations to the more adequate laws of relativity—and they were successfully applied

² Since we are here concerned with problems of a chemical nature, we may disregard those particles such as positrons, mesotrons, neutrons, etc., which are concerned with cosmic rays and nuclear processes but not with ordinary atomic behavior.

even to single atoms so long as no attempt was made to investigate the internal structure of the atom. Considering the atom to be a perfectly elastic miniature billiard ball having size, mass, and velocity but no internal properties, classical mechanics was able to handle in a statistical fashion the dynamics of large systems of atoms in a gaseous form and to deduce a number of valid conclusions concerning the specific heat, gas laws, viscosity, and diffusion constants of gases. On the other hand, failure attended all endeavors to apply these laws to the swarm of electrons surrounding a nucleus. A system of this sort is unstable classically and can never come to thermal equilibrium. Applying the classical laws of statistical mechanics, one finds that some of the electrons will move very close to the nucleus, the energy lost in this process being acquired by other electrons which move farther out. According to classical mechanics this process will continue without ever reaching equilibrium and during it the atom will be thoroughly torn apart.

Another difficulty in the classical theory arises from the electrodynamics of an accelerated electron. An electron moving in the field of a nucleus is accelerated, and classical electromagnetic theory predicts that under these circumstances electromagnetic energy will be radiated—the atom being in effect a microscopic radio transmitting station in which the charging currents in the antenna are represented by the motions of the electrons. According to this theory an atomic system would continually radiate energy, and it could be proved that no equilibrium like that actually observed between matter and radiation would ever be achieved.

Thus, classical mechanics and electromagnetics were incapable of taking the electrons and nuclei as building blocks and constructing solids or even atoms from them. To put it bluntly, the classical laws were wrong; although adequate for large-scale phenomena, they were inapplicable to phenomena of an atomic scale.

Nevertheless, modified applications of the classical theory had a great number of successes in the atomic theory of solids. Dealing with the atoms as elastic idealized billiard balls led to the correct value for the specific heat of solids, at least at normal temperatures, and the electron theory of conduction in metals was in many respects quite successful. None of the successes of the conduction theory were completely satisfying, however, because the assumptions needed to explain one set of facts were incompatible with other sets of facts and the whole field was greatly lacking in unity. According to this classical theory a metal contained free electrons which could move under the influence of an electric field and thus conduct a current. Their motion was

impeded by collision with the atoms (ions, really, since they are atoms which have given up free electrons) according to some theories, and with the spaces between atoms according to other theories, and this impeding process gave rise to electrical resistance. The free electrons were capable of conducting thermal as well as electrical currents. Although the theory gave reasonable values for the electrical and thermal conductivities of metals at room temperature, the predicted dependence upon temperature was wrong: the resistance of a pure metal is known from experiment to be very nearly proportional to the absolute temperature; the classical theory, unless aided by very unnatural assumptions, predicted proportionality to the square root of the absolute temperature. Another difficulty, the greatest in fact which beset the old theory of free electrons in metals, was concerned with the specific heats of metals. According to the billiard ball theory of gases, the specific heat arose from the kinetic energy of motion of the gas atoms; thus the specific heat at constant volume of one gram atom of a monatomic gas was $(3/2)R$, where R is the gas constant. This was in good agreement with experiment. For solids this specific heat was just doubled, giving $(6/2)R$ because of the addition of potential energy to the kinetic. For a metal the free electrons were regarded as having kinetic energy. In order to explain the observed electrical properties of a metal, the number of electrons was taken as approximately equal to the number of atoms. Hence, as for a monatomic gas, a specific heat of $(3/2)R$ was expected for the electron gas and, therefore, a specific heat of $(9/2)R$ was predicted for a metal. Measurement shows that most crystals, metals included, fit quite well the value of $(6/2)R$ and that $(9/2)R$ is incorrect. Thus classical theory was left with the dilemma that to explain electrical properties one free electron per atom was needed while to explain specific heat one free electron per atom was far too many. This dilemma is very neatly resolved in the new theory; in this paper we shall show why the free electrons are not free for specific heat and in a later paper why they are free for conduction. We shall also show that the new theory leads to quite proper values for the conductivity and also explains facts concerning the resistance of alloys, which the classical theory could not do.

According to the classical theory there was one quantity that should be the same for all metals and this was the ratio of the thermal to the electrical conductivities. This ratio, known as the Wiedemann-Franz ratio, was predicted to be equal to the absolute temperature times a universal constant L called the Lorentz number. This prediction was in reasonable agreement with experiment. The new wave me-

chanical theory predicts the same result, but with a slightly different value for L . According to the old theory $L = 2k^2/e^2 = 1.44 \times 10^{-8}$ volts²/degree² where k is Boltzmann's constant, while the new gives $L = \pi^2 k^2/3e^2 = 2.45 \times 10^{-8}$ volts²/degree², and the experimental values for several elements are Cu 2.23, Ag 2.31, Au 2.35, Mo 2.61, W 3.04, Fe 2.47—all times 10^{-8} volts²/degree². We see that the constancy of the Lorentz number predicted by both theories is in reasonable agreement with experiment, but that in predicting the numerical value of the constant the new theory is better than the old.

The fundamental problem of how the electrons and nuclei form stable atoms and crystals was, as we have said above, inexplicable on the older theory. The newer quantum mechanics of Bohr and later that of Schroedinger, Heisenberg, and Dirac were needed. Bohr postulated that out of the infinity of possible motions for the electrons of an atom, only a certain restricted set was permitted. Each permitted motion corresponded to a definite energy for the atomic system as a whole. This concept of energy levels for the atom gave a natural interpretation to nature of atomic spectra and explained the meaning of the combination principle. In order to restrict the atomic motions to certain energy levels, Bohr supposed that the laws of atomic dynamics were such that only those modes of motion were permitted for which certain dynamical quantities, called phase integrals, had values equal to multiples of Planck's constant h . For the case of the hydrogen atom these laws led to the now well-known Bohr orbits for the electron and to energy levels which were in good agreement with experiment. For atoms with more electrons it was very difficult to apply Bohr's laws except in a very approximate and unsatisfactory way. However, two very valuable concepts came from his theory which are preserved in the newer wave mechanical theory. These were that the individual electrons could be thought of as restricted to certain orbits and that these orbits were specified by giving them certain quantum numbers. It was found that three quantum numbers were needed to specify the orbit. All atoms were found to have the same general scheme of orbits. The number of electrons moving in these orbits varies from atom to atom and for any given atom is equal to the atomic number Z . In order to explain the facts of spectroscopy and the periodic table of the elements, it was necessary to introduce a rule known as Pauli's principle. This principle states simply that no more than two electrons may occupy the same orbit in an atom; that is, no more than two electrons of an atom may have the same three quantum numbers. As we shall discuss in the next section, a complete specification of the state of an electron in an atom requires four quantum numbers; two

electrons in the same orbit have different values for their fourth quantum number. We shall use the term "quantum state" to signify the permitted behavior corresponding to specified values for the four quantum numbers. In this language, Pauli's principle asserts simply that no two electrons in a given atom can be simultaneously in the same quantum state; that is, Pauli's principle is a quantum mechanical analogue of the classical principle that two bodies cannot occupy the same place at the same time. The two ideas—first that the motions of the electrons are quantized so that only certain quantum states are allowed, and second that in an atom only one electron can occupy a given quantum state—form the basis of all quantum mechanical thinking. We shall make use of them continually in the following discussion. We shall use them, however, not in connection with the orbits of Bohr but instead with the wave functions of Schroedinger.

The Bohr theory can be applied only with difficulty to any atom but hydrogen. The difficulty lies in determining the motions of the electrons in the complex interacting fields of the electrons and the nucleus. This problem is even more difficult in the case of a solid where there are many atoms, and it would seem hopeless to try to find out why the electronic orbits in insulating crystals such as rock salt or diamond do not permit electrons to move through the crystal and carry a current, while the orbits in metals do. Indeed not only does the Bohr theory have the foregoing disadvantage but it is probably wrong. Fortunately there is a theory both sounder and easier to apply embodied in the "wave equation of Schroedinger."

One feature, probably not sufficiently stressed, about Schroedinger's equation is its relative convenience. The word "relative" must be used here because it is usually very laborious to obtain solutions for the equation and only in the simplest cases can we obtain exact solutions. Compared to the classical equations and the equations of Bohr, however, it *is* convenient. Quite satisfactory approximate solutions can be obtained for Schroedinger's equation even for the complex case of solids, where it would be prohibitively difficult to obtain as good solutions for the classical and Bohr equations.

ELECTRONS IN ATOMS

According to the Schroedinger theory, a differential equation can be written down for any system consisting of electrons and atomic nuclei. This equation contains an unknown wave function and an unknown energy and the instructions of the theory are to solve the equation for the unknown quantities. Furthermore, the wave function must satisfy a certain mathematical requirement which embodies

in a generalized form the restrictions imposed by Pauli's principle. As is too frequently the case in mathematical physics, it is much easier to state the problem than to solve it; the solutions of Schroedinger's equation are, in fact, so difficult to obtain that exact solutions have been found for atomic systems only of the simplest type, namely those consisting each of a single nucleus and a single electron. For this case, the quantum states and their energies are all exactly known. For other cases approximations of varying degrees of exactness must be used. The difficulty arises from the interactions between the electrons. If it were not for these interactions, one could obtain exact solutions for atoms having many electrons. The difficulty is that the interactions—they are merely electrostatic repulsions—prevent each electron from being independently in a definite quantum state. The interaction of each electron with another is in general small compared to its interaction with the nucleus. To a first approximation, then, the electrons are treated as not interacting and then corrections are applied to this over-simplified picture. (In this first approximation, the generalized mathematical statement of Pauli's principle reduces to the one we gave in the last section—only one electron may occupy a given quantum state.) As a result of this procedure of over-simplification followed by corrections, our exposition will commence with a discussion of the quantum states of an electron in an atom as if these quantum states were private possessions of the electron and not influenced or disturbed in any way by the other electrons. We shall then correct this picture to some extent by considering how the energy of a given electron depends upon the behavior of the other electrons. One correction term which we shall introduce in this way is the important "exchange energy" discussed below. Thus atomic theory represents a field of endeavor in which further progress is made largely by improvements and refinements. It should be emphasized, however, that the corrections and refinements are not additional assumptions, which are added to the theory, but that they represent instead only steps forward in improving the wave mechanical solutions.

The last paragraph mentions that an approximate treatment of Schroedinger's equation leads to a set of possible quantum states for an electron in the atom. We shall discuss Schroedinger's equation and the wave functions corresponding to the quantum states in more detail later and at present be concerned only with a description of the results. In a neutral atom the electrons arrange themselves in the quantum states in such a way as to make the energy of the atomic system a minimum. Consistent always with Pauli's

principle, only one electron can occupy a particular quantum state. When the atom is in the arrangement of lowest energy, we can say that each electron has a definite energy corresponding to whichever quantum state it occupies. This energy is most conveniently defined in terms of the amount of work required to take an electron from its state in the atom and put it in a standard state defined as zero energy. An electron in the zero energy state is to be thought of as at rest and so far removed from the atom that there is no energy of interaction between them. In this way we can define the energy of every occupied quantum state in the atom. Each of these energies must be taken as negative—since potential energy is yielded up when the electron returns to the atom—and by definition represents how tightly the electron is bound to the atom. One of the electrons will be the most loosely bound (it may be that there are several with the same energy) and the energy required to remove it is called the “ionization energy.” From our definition this is obviously the minimum energy required to convert the atom to a positive ion. The definition of the energy of a quantum state given above can be used only when an electron is in the quantum state; we can, however, define the energy of an unoccupied state conveniently in terms of the energy the atom would have if the state were occupied by “exciting” one of the electrons to this state by giving it the proper amount of energy.^{2a}

The Quantum States of the Atom

Using this definition of the energy of a quantum state, we find that for all atoms the arrangement of quantum states in energy is as shown in Fig. 2, where the ordinates represent energies and states of equal energy appear as divisions of the horizontal lines. Figure 2 does not indicate which states are normally occupied, nor could it unless we knew how many electrons there were in the atom. The general scheme of Fig. 2 is applicable, with certain changes discussed below in the energy scales, to any neutral atom in its normal state, and the energy

^{2a} This definition is subject to restrictions because the energy of an electron in the state in question depends upon the arrangement of the other electrons in the atom and this arrangement depends in turn upon which electron was excited to the initially unoccupied state. In constructing the figures we have supposed that the electron (or one of the electrons in case there are several) that is most easily removed from the atom is caused to shift from its normal state to the unoccupied state in question; this shift will change the state of the atom and since the atom was initially supposed to be in the state of lowest energy, the change in energy cannot be negative and will in general be positive but may in certain special cases be zero. The energy of the unoccupied state is defined as the energy of the occupied state from which the electron is taken plus the change of energy caused by shifting the electron. This is equivalent to saying that the energy of the unoccupied state is the ionization potential of the atom after an electron has been shifted from the highest state normally occupied to the normally unoccupied state in question.

levels represented on Fig. 2 apply to unoccupied and occupied states as well.

I have already mentioned that four quantum numbers are required to specify each quantum state. These are indicated by the letters

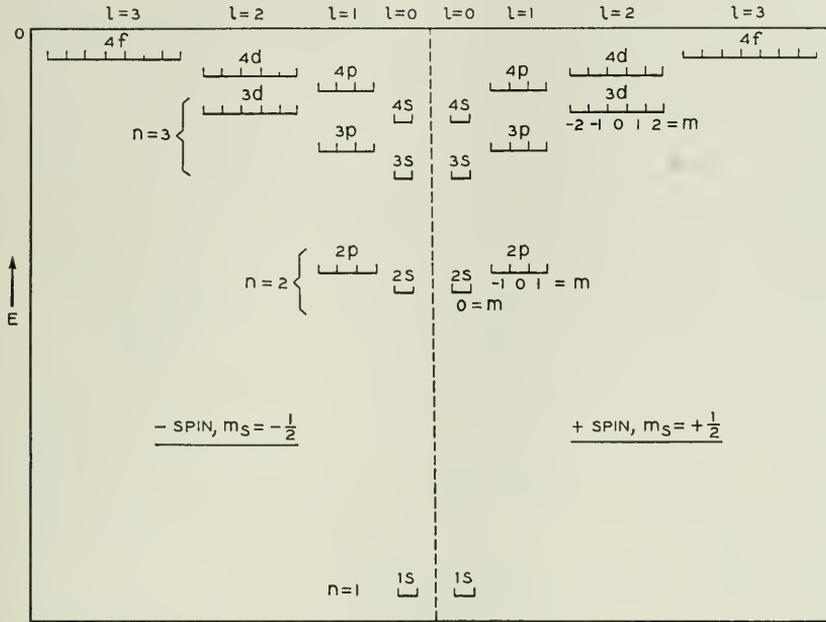


Fig. 2—Quantum states for electrons in atoms.

$n, l, m,$ and m_s . Roughly speaking, the “principal quantum number n ” fixes the “energy level” of the state; however, there is some dependence upon the “angular momentum quantum number l .” The dependence of energy upon the third quantum number, “the magnetic quantum number m ,” is slight and will be neglected in this paper. We shall consider the energy to be specified by giving n and l . A notation borrowed from spectroscopy is applied to this pair of quantum numbers and one uses the apparently quite fortuitous choice of letters $s, p, d, f, g, h,$ etc., to stand for $l = 0, 1, 2, 3, 4, 5,$ etc., and a state with $n = 3$ and $l = 2$ is known as a “ $3d$ state” and an electron occupying such a state is called a “ $3d$ electron.” The quantum laws permit the following values for $n, l,$ and m :

n takes on all positive integral values. (All states with n greater than four have been omitted from the figure; they lie between the highest states shown and zero energy.)

For a given n , l takes on all positive integer values from 0 to $n - 1$ inclusive.

For a given n and l , m takes on all integer values including zero from $-l$ to $+l$ inclusive.

The difference between right and left sides of the figure corresponds to the fourth quantum number: an electron, in addition to its electric charge, possesses angular momentum or "spin" about its axis. The rotating charge resulting from this angular momentum produces a magnetic moment. The angular momentum is quantized and there are two possible values $+1/2$ and $-1/2$ for the "spin quantum number m_s ," corresponding to the right and left halves of Fig. 2. Electrons occupying states on the right half of Fig. 2 have their spins parallel to each other and directly opposite to electrons occupying states on the left half. As already implied, the quantum numbers l and m also correspond to angular momentum and magnetic moments for the electron "orbits" (really wave functions) in the atom.³

For our purpose we need two results of the theory of the spinning electron, first that its *spin introduces a duplicity of quantum states* as indicated by the two halves of Fig. 2, and second that *all the electrons of one spin have their magnetic moments parallel and opposite to those of the other spin*. Later when we consider the question of magnetism, we shall be concerned with the direction in space of the spin vector and the magnetic moment, but not now.

Several units of energy are employed in describing atomic processes. The simplest of these is the electron volt; it is the energy acquired or lost by an electron in traversing a potential difference of one volt. For example, in a vacuum tube operating with one hundred volts between cathode and plate, the electrons strike the plate with a kinetic energy of one hundred electron volts, 100 ev. Another unit is the ionization potential of hydrogen, and as hydrogen has only one electron, which normally occupies the $1s$ state, this is also the energy of the $1s$ state. This energy is called the "atomic unit" of energy or the "Rydberg." Another unit of energy useful in chemical processes is the kilogram calorie per gram atom; this is related to the others as follows: if the energy of each atom in one gram atom is increased by one electron volt then the energy of the whole system is increased by 23.05 kilogram calories. The conversion factors are: $1 \text{ Ry} = 13.5 \text{ ev}$, $1 \text{ ev per atom} = 23.05 \text{ Kg.-cal./gm. atom}$.

³ For a discussion of the quantum states of the electrons from the point of view of angular momentum see "Spinning Atoms and Spinning Electrons" by K. K. Darrow, *Bell System Technical Journal*, XVI, p. 319, or standard texts on spectroscopy.

Variation of the Energy Levels with Atomic Number

All atoms have the same general scheme of quantum states indicated in Fig. 2. Quantitatively the energy scale varies from atom to atom. Thus the 1s state lies at -13.5 eV for hydrogen and at -24 eV for helium. This decrease (i.e., becoming more negative or moving lower down on Fig. 2) is due to the increase in nuclear charge, $Z = 1$ for hydrogen and 2 for helium, which results in greater attraction and tighter binding for electrons in helium. This steady downward motion of the levels continues as one goes from element to element in the periodic table. However, the ionization potential, the energy required to remove the most easily removed electron, does not steadily increase. In Fig. 3 we show the ionization potentials of the first twenty elements.

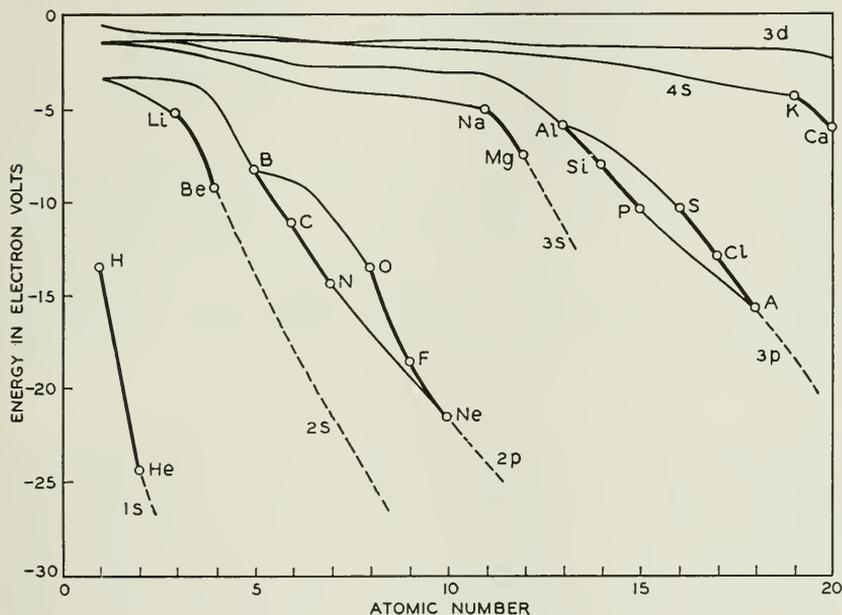


Fig. 3—Ionization potential versus atomic number.

Since we are interested in the energies of electrons rather than in ionization per se, the ionization potentials have been plotted as negative giving in this way the energies of the states in the atom. The main features of this figure can be explained by using Fig. 2 and the Pauli principle.

The Pauli principle, also known as the exclusion principle, permits only one electron to occupy each of the states of Fig. 2. The electrons in a many-electron atom will tend to go to the states of lowest energy.

Thus in helium, since its two electrons can have oppositely directed spins, each fills one of the $1s$ states; we say the "electron configuration" of helium is $1s^2$ (read as "one ess squared"). For lithium, $Z = 3$, the third electron, which cannot go to the completely filled $1s$ states, goes to the next highest, $2s$, giving $1s^2 2s$. In going from helium to lithium, all the states move to lower energies but not so much lower as to make $2s$ for lithium as low as $1s$ for helium. For this reason lithium can be relatively easily ionized, as is seen in Fig. 3.

Before continuing the discussion of particular atoms, we must point out that two changes accompany each advance from one element to the next in the periodic table. In each step the nuclear charge increases by one plus unit and at the same time an electron is added to the atom and the combined effects produce the results of Fig. 2. Quite different results are obtained if one electron alone is added to the atom. Then instead of the general falling of the levels which accompanies the double change, there is a general rising of all the levels. This is due to the unbalanced negative charge on the added electron, whose presence on the atom raises the potential energy of all the electrons and therefore raises their energy levels. For some atoms, the raising of the energy levels produced by an unbalanced electron may be so great that the electron is not bound at all or at least only very slightly, and for these atoms negative ions do not form. On the other hand, when an electron is removed from an atom all the remaining electrons become more tightly bound and the energy levels are lowered.

Exchange Energy

In Fig. 4 we show the electron configurations for the elements from lithium to neon. The decrease in ionization potential in going from beryllium to boron is due to the completed filling of the $2s$ states and the consequent start of filling of the $2p$ states. The decrease in going from nitrogen to oxygen suggests that not only do the $2s$ and $2p$ states lie at different levels but that the $2p$ states themselves lie at two different levels. This is true but in a rather special sense: *the difference in energy between the two sets of $2p$ states depends upon how they are occupied.* This difference is an "exchange energy." We shall discuss the origin of the exchange effect in the next paragraph but one; however, the aspect of it needed for this paper is illustrated in Fig. 4. We there imagine that the quantum states are represented by little trays upon which are placed weights to represent occupancy by electrons. The exchange effect corresponds to hanging the trays on springs; in this way we see that as the electrons fill up the $2p$ states

with one spin (the same effect occurs for either spin; the figure shows + spin), these states are depressed in respect to the $2p$ states with the other spin. The springs must, however, be considered to pull the

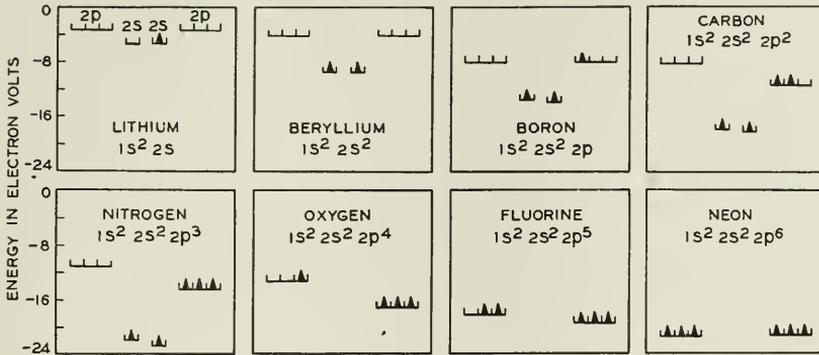


Fig. 4—Electron configurations illustrating the exchange effect.

trays up against stops with a force such that a single weight upon a tray will produce no lowering whereas two or three weights will. This effect seems contradictory to the simple idea that adding electrons raises the potential energy and the energy levels; however, it must be remembered that we are here discussing neutral atoms and that with each added electron there is also an added plus charge on the nucleus. These two charges produce the dominant variation in the energy levels and upon this variation the exchange effect is superimposed.

The reader may verify that so far as the distribution of electrons in the $2p$ states is concerned, the exchange effect will lead to the configurations shown in Fig. 4 for the states of lowest energy for the atoms. Let us consider carbon for example; if the electrons have opposite spins—that is, if there is one weight on each $2p$ tray—there will be no lowering due to the exchange effect; if the electrons have the same spin, however, then each loses energy because of exchange and the energy of the atom is less than for the case of parallel spins. The fact that one electron is not enough and that two or more electrons are required to produce the exchange effect is a natural consequence of the origin of the exchange energy.

The exchange energy is due to the electrostatic repulsion between the electrons and results directly from the application of Pauli's principle to Schrodinger's equation. The exchange effect emerges in a quite straightforward fashion from a consideration of wave functions, but usually no attempt is made to explain it in non-mathematical

terms. It seems to the writer, however, that the explanation given below does contain the mathematical essence in physical language.^{3a} Pauli's principle, we have said, is the quantum mechanical analogue for electrons of the classical law that two bodies may not occupy the same place at the same time; it is, however, more general in the sense that it does not apply alone to location but rather to a combination of location and velocity and spin, and it requires that any two electrons differ essentially in one or more of these. Now a difference in the values for the spin quantum numbers of two electrons is a sufficiently great difference to permit them to have the same velocity and the same location (i.e., be very near together compared to atomic dimensions). If the spin quantum numbers are the same, however, there must be a difference in location or in velocity. Now two electrons having the same values of n and l , as for example two $2p$ electrons, move in similar orbits and have much the same velocities; hence, if their spins are the same they must differ in location—that is, they will satisfy Pauli's principle by keeping away from each other. If, however, their spins are different, then they need not keep away from each other, and in their motion about the nucleus they are, on the average, closer together than for the case of the same spin. Since the energy of repulsion between the two electrons decreases as they move farther apart, the average energy of the electrons is less for the case of parallel spins, for which Pauli's principle requires most difference in location; and this is just the effect shown in Fig. 4. Furthermore, if the electrons differ in their values of n and l , then their velocities are quite different and the restriction upon location is not so important and their electrostatic energy of repulsion for parallel spins is nearly the same as for opposite spins. There is, however, a small exchange effect between electrons of different n and l values as may be appreciated in Fig. 4 for boron, for example, by noting that one $2s$ level is depressed compared to the other owing to the presence of the $2p$ electron.

We see that helium and neon correspond to electron configurations which fill all the levels below $n = 2$ and $n = 3$ respectively. One sometimes refers to the states with $n = 1$ as the K shell, and to those with $n = 2, 3, 4$, etc. as L, M, N, etc., shells. The rare gases helium and neon then correspond to electron configurations consisting of "closed shells"—that is, to shells all of whose states are occupied.

^{3a} As the aspects of exchange energy needed for the exposition are those discussed above in connection with Fig. 4, this explanation is not essential to the later argument of this article and is given in the hope that it may invest the concept of exchange energy with the appearance of a little more physical reality. If it fails in this, the reader is requested to disregard it.

The Transition Elements

Actually argon does not correspond to a complete system of closed shells, since for it none of the $3d$ states in the M shell is occupied. For the elements below copper, $Z = 29$, these $3d$ levels lie above the $4s$ and below the $4p$. They are filled up progressively in the series of elements scandium, titanium, vanadium, chromium, manganese, iron, cobalt, and nickel, which are known as the transition elements of the first long period of the periodic table. The first two elements after argon are potassium (an alkali) and calcium (an alkaline earth); these are similar to sodium and magnesium in having respectively one and two s electrons. The first transition element, scandium, however, is not like beryllium or aluminum, for with it the filling of the $3d$ states begins. The electron configurations for several of the transition elements are shown in Fig. 6. An interesting case occurs at chromium;

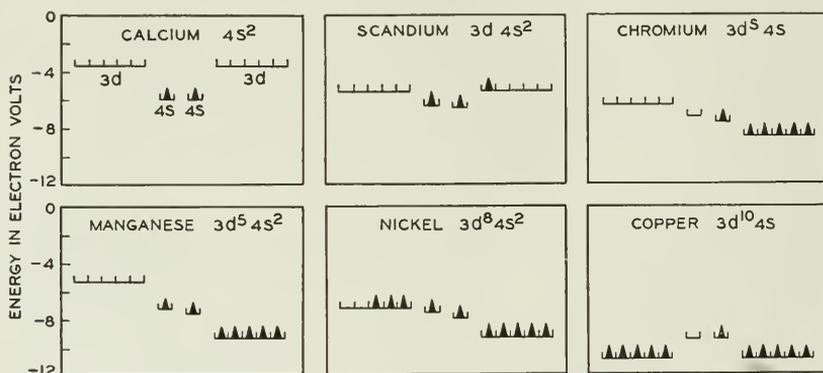


Fig. 6—Electron configurations for transition elements.

for it the exchange effect is so great that the $3d$ levels drop below the $4s$ and one $4s$ electron is transferred. Since there is an exchange effect between all electrons of the same spin, the remaining occupied $4s$ state in chromium has the same spin as the occupied $3d$ states. A similar transfer of a $4s$ electron occurs at copper, which is then left with one $4s$ electron and tends therefore to be monovalent (in the divalent copper ion the $4s$ electron and one $3d$ electron are removed). These transition elements are of particular interest because three of them, iron, nickel, and cobalt, are ferromagnetic in the solid. The atoms themselves are magnetic, as may readily be seen for chromium, for example; in it all the electrons have their spins parallel and hence their magnetic moments add to give a free chromium atom a magnetic

moment six times as large as the spin magnetic moment of the electron.⁵ The same exchange effect which causes the $3d$ quantum states to fill unevenly in the isolated atom causes, in the case of the metal, an uneven filling of the "energy bands" which arise from these $3d$ states. We shall return to this topic in the section on ferromagnetism.

Solving Schroedinger's Equation

The possible quantum states of an atom are obtained by solving Schroedinger's equation for an electron moving in the potential field of the nucleus and the other electrons. In Fig. 7*a* we have represented the potential energy of an electron in an atom. If this potential energy (call it U) is known as a function of the position x, y, z , of the electron, then the Schroedinger equation is

$$\frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2} + \frac{\partial^2\psi}{\partial z^2} + \frac{8\pi^2m}{h^2}(E - U)\psi = 0, \quad (1)$$

where m is the mass of the electron, h is Planck's constant and E is an unknown energy and ψ an unknown wave function, for which a physical interpretation will shortly be given. It is found that this equation possesses proper solutions only for certain values of E ; once these values are known, the equation can be solved for the unknown wave functions. The fact that only certain values of E are possible will probably seem more natural after reading the discussion given below of a mechanical system. The permitted energies and wave functions give the system of quantum states of Fig. 2.

The wave equation of Schroedinger is similar in form to many of the other wave equations of mathematical physics. In Fig. 7*g* to 7*j* we represent a stretched membrane like a rectangular drumhead. If the mass per unit area of the membrane is σ and the surface tension is T , then the wave equation for it is

$$\frac{\partial^2\varphi}{\partial x^2} + \frac{\partial^2\varphi}{\partial z^2} + \frac{4\pi^2f^2\sigma}{T}\varphi = 0, \quad (2)$$

where f is the unknown frequency of vibration and φ is the unknown vertical displacement. Applied to the membrane, this equation has solutions only for certain values of f ; the standing wave patterns corresponding to the four lowest frequencies are shown in Figs. 7*g* to 7*j*.

⁵ For transition elements other than chromium, the motions of the electrons in their wave functions produce magnetic moments that must be considered as well as the spin; for a discussion of this point the reader is again referred to "Spinning Atoms and Spinning Electrons" by K. K. Darrow, *Bell System Technical Journal*, XVI, p. 319 and to texts on atomic physics.

The type of vibration of the system in one of these patterns is called "a normal mode." The patterns are described by two "quantum numbers" p and q which are equal to one plus the number of nodal lines (indicated by arrows) running across the membrane from front to back and from right to left respectively. In Figs. 7b and 7c we show the wave patterns corresponding to the $1s$ and $2s$ states of the atom; the quantum numbers of the ψ waves are also correlated to nodes. In the case of the membrane the frequencies and standing

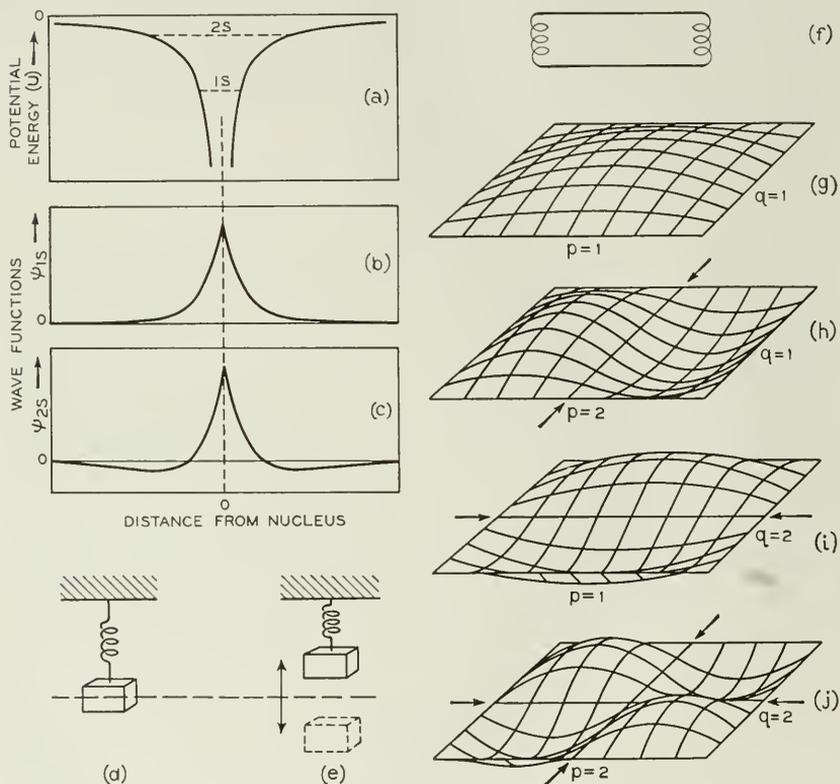


Fig. 7—The atom and some mechanical and electrical analogues.

- (a) Potential energy of an electron in the atom.
- (b) The $1s$ wave function.
- (c) The $2s$ wave function.
- (d) A mechanical analogue and
- (e) its normal mode of vibration.
- (f) An electrical analogue.
- (g) to (j) The first four normal modes of vibration of a stretched drum head. (From "Vibration and Sound" by P. M. Morse, McGraw-Hill, New York, 1937. Courtesy of the McGraw-Hill Book Co.)

wave patterns are determined not only by the values of mass per unit area, σ , and tension, T , of the membrane but also by the "boundary condition" that it be clamped on its rectangular edge; at this edge the vertical displacement φ must vanish. The corresponding boundary condition for the atom is that the wave function ψ vanish at all points infinitely far from the nucleus.

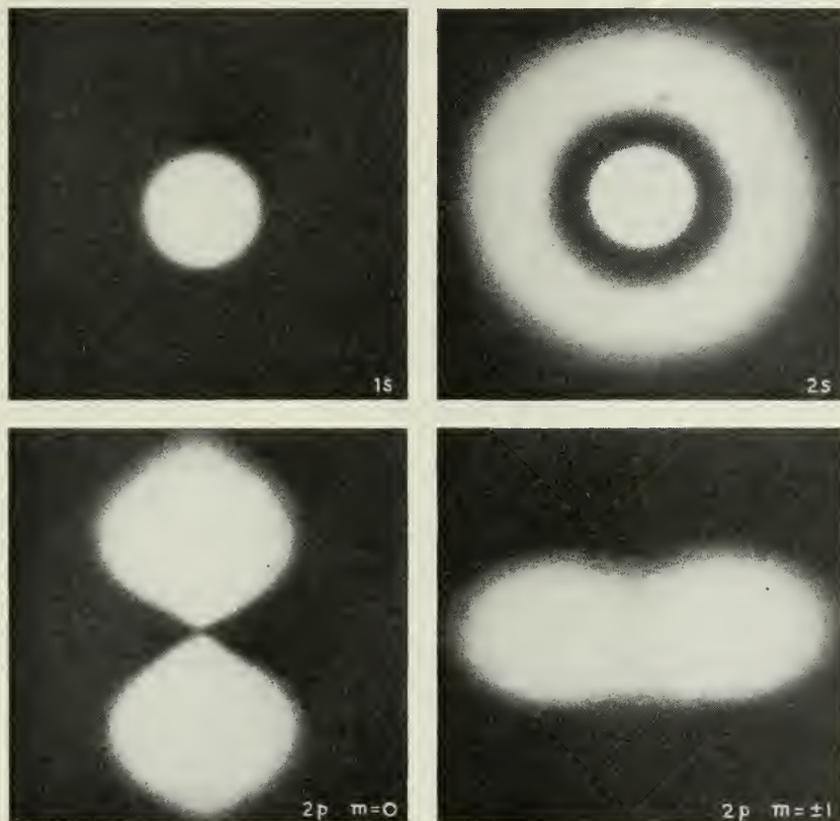


Fig. 8—The electron charge densities for four wave functions. Cross-sections are given for the $1s$ and $2s$ wave functions and perspective views for the $2p$. $1s$ represents a ball of charge; $2s$, a ball surrounded by a shell; $2p\ m = 0$, a dumbbell-like distribution; $2p\ m = \pm 1$, a doughnut-like distribution seen edgewise.

The quantity $|\psi|^2$ has a direct physical interpretation: its value at any point in space gives the probability of finding the electron at that point. If it were possible to take a photograph of the electron's motion with a time exposure so long that a true average of its positions would be obtained, this photograph would represent $|\psi|^2$. In Fig. 8

we show the predicted patterns as obtained by H. E. White,⁶ who photographed a model representing the wave functions. We see that for the $2s$ wave function the electron is much farther from the nucleus on the average than for the $1s$; this accounts for higher energy of the $2s$ state. For a hydrogen atom the $2s$ and $2p$ actually have the same energy. For other elements the $2s$ lies lower as shown in Fig. 2; this is because an electron in the $2s$ state penetrates the K shell and feels the full charge of the nucleus whereas an electron in the $2p$ state stays outside of the K shell and is thus shielded from the nucleus by the two electrons of the K shell.

For purposes of illustration we have considered the rectangular drumhead as a mechanical analogue for the wave equation. Other analogues are represented by sound waves in rooms and in organ pipes and by standing electromagnetic waves in wave guides, tuned cavities, and rhumbatron oscillators. We shall use two simple analogues in our later discussion. One is the mechanical vibrator represented in Fig. 7*d* which we consider to be restricted to vertical motion. It is a system with a single frequency—like an imaginary atom with only one possible state—and its one normal mode of vibration is a simple harmonic motion up and down equally far above and below its equilibrium position as indicated in Fig. 7*e*. The other is an electrical analogue, Fig. 7*f*, consisting of a section of transmission line terminated at each end by a high inductance. This system has a series of normal modes of vibration and a related series of allowed frequencies. The allowed frequencies correspond to the energy levels of the atom.

ELECTRONS IN MOLECULES

We shall next consider what happens when two atoms are brought so close together that their quantum states “interact.” Two similar atoms widely separated have each a distinct set of quantum states and wave functions and the scheme of energy levels for the two atoms is obtained by duplicating the energy level scheme of Fig. 2. However, if the atoms move so near together that the wave functions for the corresponding quantum states of the two atoms overlap, there is an alteration in the energy levels. Figure 9 is intended to illustrate this process. Figure 9*a* shows the potential energy of an electron for points on a line passing through the centers of the two nuclei, and Figs. 9*b* and 9*c* show for points on the same line the values of the correct wave functions in this field. These wave functions are obtained, approximately, by using the $1s$ wave function for the two separate

⁶*Physical Review*, 37, 1416 (1931). I am indebted to Professor White for the photographs used for these illustrations.

atoms; b represents the sum of the wave functions and c the difference. The process involved in getting these molecular wave functions is mathematically similar to that of finding the normal modes for a system of two similar coupled oscillators. In Fig. 9d we represent two

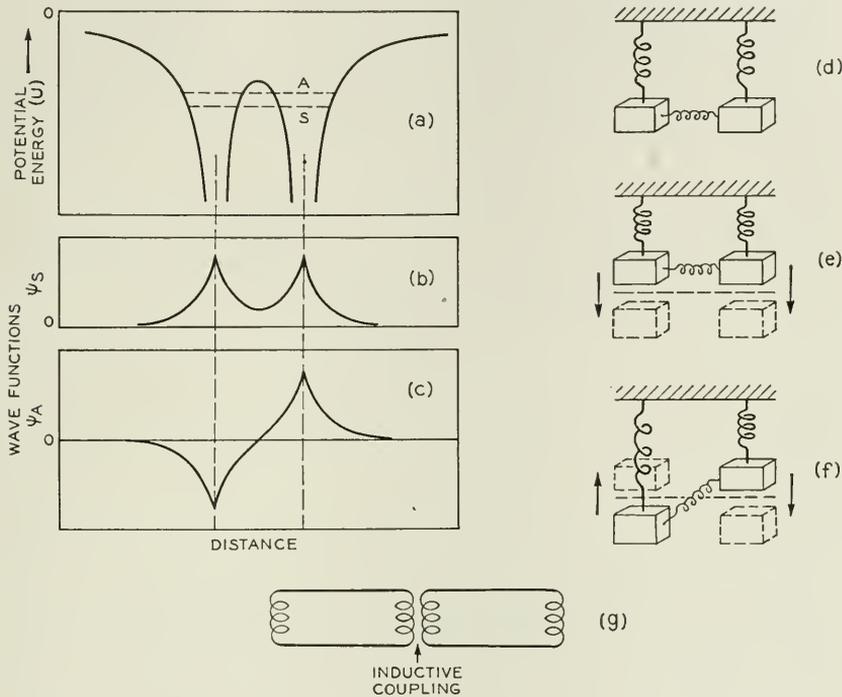


Fig. 9—A diatomic molecule and some mechanical and electrical analogues.

- (a) The potential energy of an electron for points on a line through the two nuclei.
- (b) and (c) Values of two wave functions for points on the same line.
- (d) Two coupled oscillators.
- (e) and (f) Their normal modes of vibration.
- (g) Two coupled circuits.

weakly coupled oscillators. The normal modes of vibration for the coupled system are as indicated in Figs. 9e and 9f. These two modes have different frequencies. Similarly if two electrical circuits are placed so that there is some inductive coupling between them, we find that each frequency is split into a pair. This inductive coupling is similar to the overlapping of the wave functions; thus the coupling between the circuits is large when the electromagnetic field of one reaches over to the other. We may summarize the situation by saying that before coupling each frequency occurred twice, once for each sys-

tem; after coupling two frequencies are present and the corresponding modes of vibration belong not to the individual systems but instead to the pair of systems. For the case of atoms each quantum state occurs twice, once for each atom, before the atoms interact; after interaction there are still two quantum states but now they have different energies and are shared by both atoms.

As the atoms are brought closer together the energies separate more and more. The behavior is indicated qualitatively in Fig. 10. The

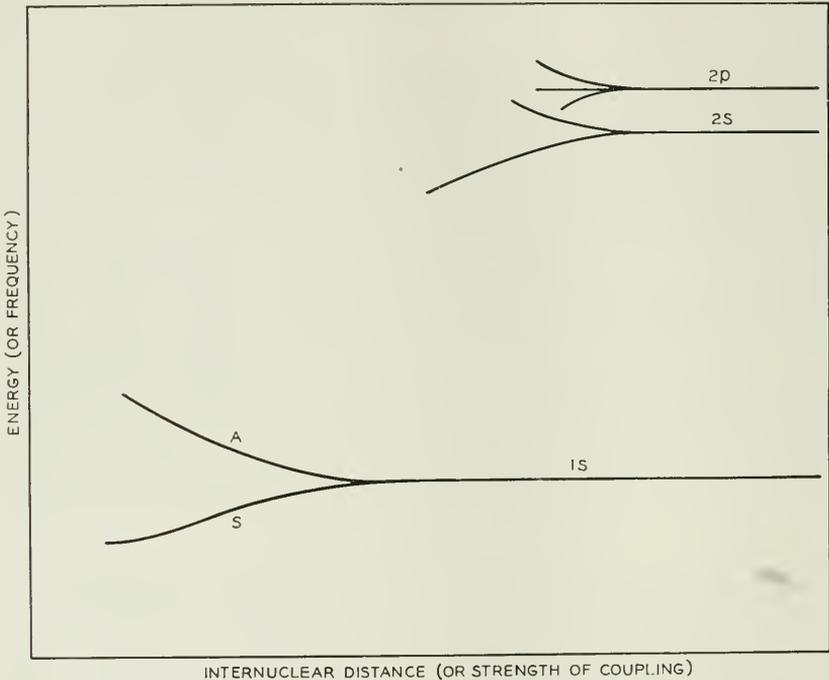


Fig. 10—Energy levels of a diatomic molecule versus internuclear distance.

L levels ($2s$ and $2p$) split at larger distances than the K levels because their wave functions extend farther from the nucleus (see Figs. 7 and 8) and overlap at greater distances. The details of the splitting are somewhat complicated and only the start is shown here. For the mechanical analogues shown in Fig. 8, the coupling raises one frequency and leaves the other unaltered. On the other hand, the quantum mechanical interaction results, at large distances, in equal displacement up and down for the energy levels.

We can use Fig. 10 to describe the formation of a molecule of hydrogen, H_2 . We start initially with the single electron of each atom

in the $1s$ state. When the atoms have come together the $1s$ states have split into two energies with two states for each energy—one with each spin. In the H_2 molecule, the two electrons will both go into the lower $1s$ states, which both have the wave function of Fig. 8*b*. Bringing the atoms closer together decreases the energy of the electrons and results in the binding together of the atoms. This tendency of the electrons to reduce their energies by drawing the atoms together is opposed by the electrostatic repulsion between the nuclei. The repulsion between the nuclei is inoperative when the atoms are sensibly separated because then each nucleus is shielded from the nucleus of the other atom by the electron of that atom. When the atoms are closer together, however, the electrons no longer perform this shielding perfectly and the nuclear repulsions are important. Hence with decreasing interatomic spacing the electronic energy decreases and the energy of repulsion of the nuclei increases, and the equilibrium internuclear distance is the one which makes the total energy of the molecule a minimum.

The situation is quite different for two helium atoms. There being two electrons in each, for them all four of the “ $1s$ molecular orbitals,” as the states of Fig. 9 are called, are occupied. When all the molecular orbitals are occupied, there is no decrease in energy when the two atoms are brought together: in this case the decrease of energy for the electrons in the two lowest states is compensated by the increase for the electrons in the upper states—more than compensated, as a matter of fact, because the upper states rise slightly more rapidly than the lower ones fall. This effect results in a repulsion between two helium atoms. This repulsion is a consequence of the closed shell nature of the helium atom and always occurs between such closed shells even if the atoms are different, as, for example, a neon and an argon atom. We shall refer to this closed shell repulsion, which occurs when the wave functions of the two closed shells encroach upon each other, as an “encroachment energy.” The encroachment energy, as we have said, always corresponds to a repulsive force between the closed shells. We shall find that it plays a very important role both in ionic crystals and in metals.

The encroachment energy occurs not only between rare gas atoms but also between ions of elements which as neutral atoms have partly filled shells but in the ionic form have closed shells. Consider, for example, an alkali halide molecule such as LiF. For this case the $2s$ valence electron of lithium is transferred to the vacant $2p$ level of fluorine (see Fig. 4), thus leaving two ions with closed shell configurations, the Li^+ being He-like, the F^- being Ne-like. These two oppo-

sitely charged ions attract each other and draw together until the encroachment repulsion between their closed shells balances the attraction and holds them apart. Conversely if one of two atoms having closed shells normally is converted to an ion, the closed shell arrangement will be destroyed and an attraction will result. For example, He_2^+ ions, which may be thought of as formed from an atom and an ion, have been observed in the mass spectrograph.⁷ The attraction is explained by noting that in this case there are three electrons and the effect of two of them in the lower 1s molecular orbital overbalances the one in the upper orbital and gives rise to a net attraction.

ELECTRONS IN CRYSTALS

We must now investigate the quantum states and their energy levels for electrons in crystals. As in the case of the diatomic molecule we shall study the dependence of the energies upon the distance between atoms, which in the case of a crystal is called the lattice constant. We shall treat the lattice constant as a variable and shall refer to the values for it found experimentally as "observed" or "experimental lattice constants" and indicate them on the figures by the symbol a_0 . We shall consider the allowed states to be occupied in accordance with Pauli's principle and on this basis find how the energy of the crystal as a whole depends upon the lattice constant. In this section we shall deal with crystals at the absolute zero of temperature and leave the complicating features of thermal effects to a later section. According to theory, the equilibrium state of a system at absolute zero is that one which makes the energy least. Hence, a knowledge of the dependence of energy upon lattice constant can be used to predict the equilibrium lattice constant—that is, the one which should be found experimentally—for according to the theory quoted above, the equilibrium lattice constant is the one which makes the energy of the crystal least.

In Fig. 11 we show the potential energy for an electron in a one-dimensional crystal, the distance being measured along a line passing through the atomic nuclei of the constituent atoms. In the interests of simplicity we imagine that high potential walls through which the electron cannot pass bound the crystal at both extremities. These boundary conditions lead to a simpler set of wave functions than would boundary conditions like those discussed for the free atom. The simplification of problems by arbitrarily choosing certain boundary conditions is a standard device in some branches of quantum mechanics; it introduces an error, but if the crystal is large, the error is negligible;

⁷F. L. Arnot and Marjorie B. M'Ewen, *Proc. Roy. Soc.*, 171, 106, 1939.

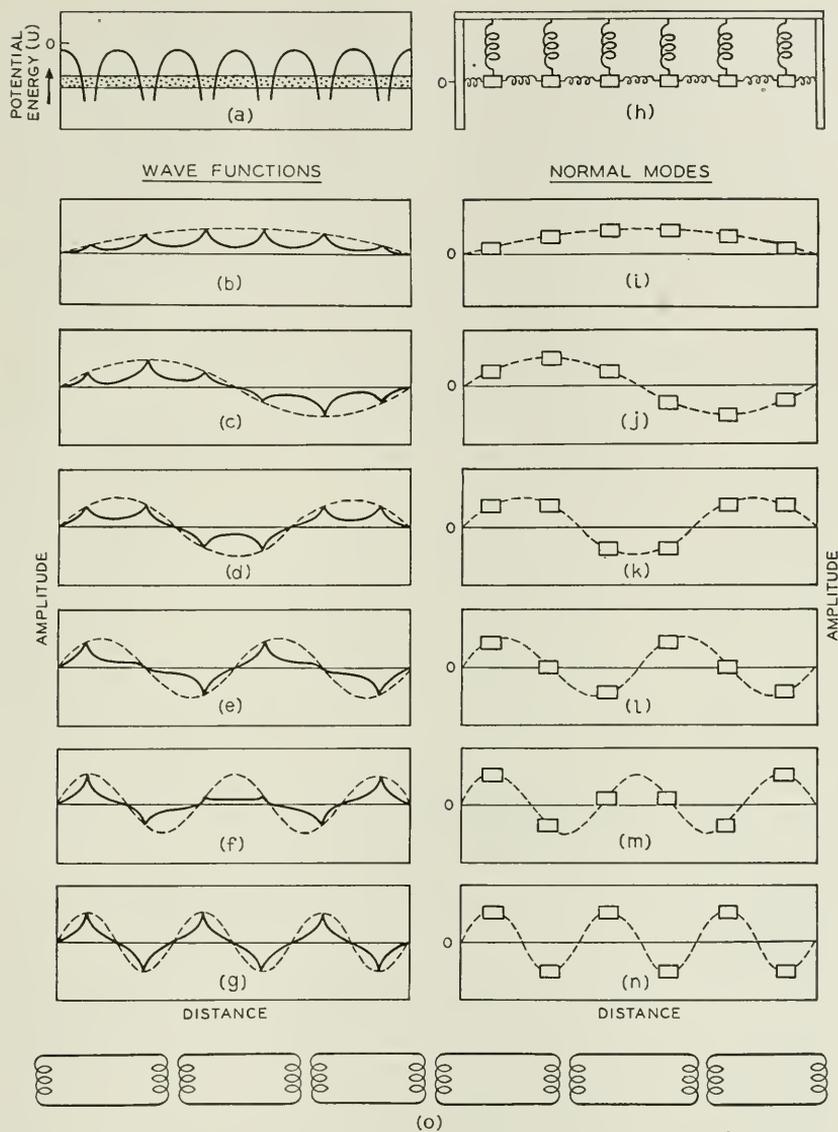


Fig. 11—A one-dimensional crystal and some mechanical and electrical analogues.

- (a) The potential energy of an electron for points on a line through the nuclei.
- (b) to (g) Wave functions for points on the same line.
- (h) Coupled oscillators.
- (i) to (n) Their normal modes of vibration.
- (o) Coupled circuits.

the situation is similar to that which arises through neglecting "edge effects" in calculating the capacity of a parallel plate condenser.

In Fig. 11 we show also a series of coupled oscillators with boundary conditions corresponding to those prescribed for the atoms. For this case there are six coupled oscillators, which when uncoupled had six independent normal modes of vibration all with the same frequency, like that shown for the single oscillator of Fig. 7*d*. After coupling there are six normal modes all having different frequencies; the standing wave patterns corresponding to these are shown in Figs. 11*i* to 11*n*. A similar splitting of frequencies occurs when the members of a set of electrical circuits are placed in close proximity as indicated in Fig. 11*o*. For them the situation is more complicated than for the mechanical oscillators; each mechanical oscillator has but a single frequency, whereas each circuit has a fundamental and a sequence of overtones. Each possible frequency for the electrical circuits is split by coupling into a set of six.

In Figs. 11*b* to 11*g* are shown the proper electron wave functions which arise from the 1*s* atomic states. These wave functions have different energies. When the atoms were separated there were six 1*s* wave functions for the six atoms and each of these gave two states—one for each spin. After coupling we find six crystal wave functions and twelve crystal quantum states, the same number of states for each spin as before. This illustrates a fundamental theorem concerning wave functions in crystals which holds for two and three dimensions as well as for one and is true no matter how large the number of atoms in the crystal. This theorem, which we shall refer to as the "conservation of states," may be stated as follows: consider a set of N similar isolated quantum mechanical systems; they may be single atoms or molecules. Any particular quantum state is then repeated N times over, once for each system. Now bring the systems together so that the energy levels have split up. Then for each N -times-repeated quantum state of the isolated systems, we find a set of N crystal quantum states. In other words, putting the systems together may change the energies and wave functions of the quantum states but no states are gained or lost in the process.

In Fig. 12 we indicate how the energy levels of the states depend upon the lattice constant. Each energy level in the figure corresponds to two states, one for each spin. For simplicity only two atomic levels are shown here. Higher energy levels split appreciably at larger lattice constants because of the greater spatial extension of their wave functions. For any particular lattice constant the energy levels arising from a given atomic state lie in a certain band of energy. The

number of states in the band is, of course, proportional to the number of atoms; however, if the number of atoms is large, the width of the band is independent of the number of atoms. This concept of allowed bands of energies for the crystal states plays the same role in crystals as the concept of energy levels in the atom. We shall refer to 3s bands

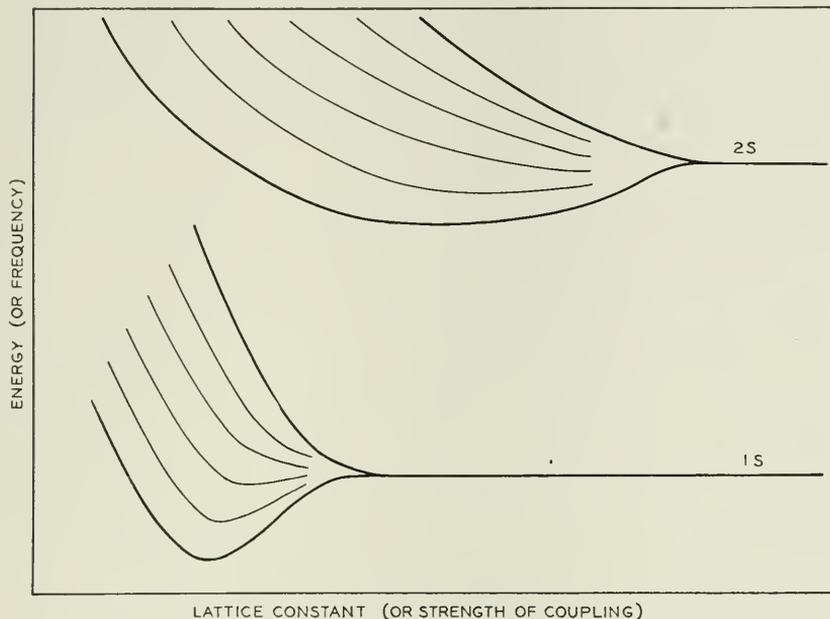


Fig. 12—Dependence of energy levels upon lattice constant or frequency of vibration upon strength of coupling.

and 3d bands of energy levels in crystals in much the same way as we refer to the 3s and 3d atomic energy levels from which these bands arise.

We must emphasize that like the molecular states, the crystal states do not belong to the atoms individually but instead belong to entire system of atoms.

Before proceeding with the application of these ideas to crystals with large numbers of atoms, we shall digress by anticipating several subjects to be taken up in the next paper. For the energy levels of isolated atoms the quantum numbers n , l , m , and m_s were satisfactory. For a crystal, however, there will be many crystal quantum states in an energy band all arising from atomic levels having the same values of n , l , m , and m_s . A new quantum number is therefore needed to distinguish the various crystal states in an energy band one from another. In Fig. 11 we see that the wave function of each crystal state is asso-

ciated with a wave form, shown dashed. This wave form is in every case of such a wave-length that it has an integral number of half wave-lengths along the edge of the crystal.^{7a} The number of half wave-lengths is a suitable quantum number for the wave functions in the crystal and a more general consideration of it in the next paper will lead us into the theory of the "Brillouin zone" and the zone structure of energy bands. The second subject concerns the transmission properties of the crystal. The set of coupled circuits of Fig. 11 constitutes a length of transmission line. A line of this type is a simple filter network and as such it has bands of frequency in which power will be transmitted and bands in which it will not. The allowed frequencies lie in the transmitting bands. The system of coupled mechanical vibrators likewise constitutes a mechanical filter. Just as the mechanical and electrical systems can transmit power in their allowed bands, a crystal can transmit an electron whose energy is in an allowed band. The electrons in an allowed band, however, can produce a net current only if the band is partially filled. Electrons in wholly filled energy bands, although individually representing tiny currents to and fro in the crystal, can produce—we shall find—no net current as their individual currents cancel out in pairs. On this basis the theorist explains the difference between metals and insulators as follows: in a metal some of the energy bands are partly filled, but in an insulator each energy band is either completely filled or completely empty.

Distributions of Quantum States in Energy Bands

When there are a very large number of atoms in the crystal, it is impractical to represent the energy levels by distinct lines as was done for the case of six atoms in Fig. 12 and another scheme must be used. For a crystal of macroscopic dimensions the number of levels in the band is of the order of 10^{24} , that is a million million million million. When so many levels are placed so close together, a continuous band of allowed energies is suggested. Actually, of course, only a discrete set of allowed energies is possible, the total number in the band being that required by the conservation of states. We shall now consider the distribution in energy of these quantum states; that is, how many lie in a given range of energy between E and $E + dE$. Let us call

^{7a} For a three-dimensional crystal having the external shape of a cube, the three-dimensional wave function has an integral number of half wave-lengths along lines parallel to each edge of the crystal. This condition is illustrated in a simplified form by the wave patterns for the two-dimensional drum head shown in Fig. 7; for each normal mode, there is an integral number of half wave-lengths parallel to each boundary of the membrane, and, in fact, the values of these numbers are given by p and q .

this number dN ; it will depend upon E and be proportional to dE and we may write

$$dN = N(E)dE, \tag{3}$$

where the function $N(E)$ represents the "number of quantum states per unit energy" at E . This equation, like many in statistical mechanics requires special interpretation, because if dE is small enough—less than the spacing between levels in the band—it may include no levels. If, however, we always use small but not infinitesimal values for dE , so many levels will be included in it that equation (3) is quite satisfactory. In Fig. 13a we represent qualitatively the distribution in

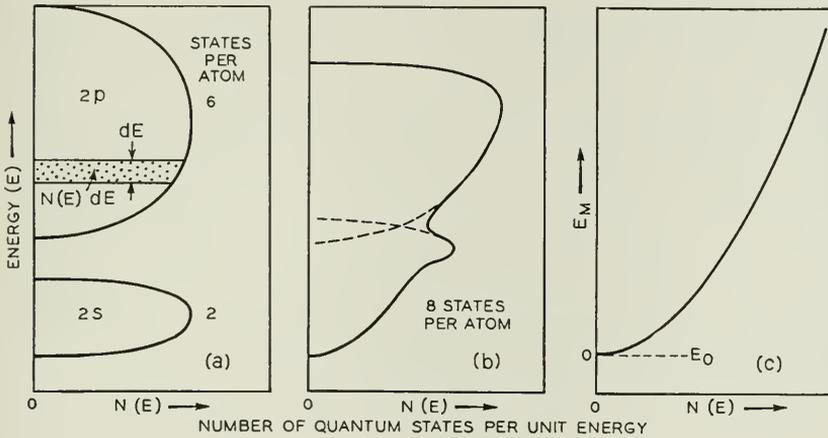


Fig. 13—Distribution of energy states in energy.

- (a) For two separate energy bands.
- (b) For overlapping energy bands.
- (c) For free electrons.

energy for two energy bands. We plot $N(E)$ horizontally so as to retain the vertical scale for E . The area under the curve for the $2p$ levels is three times that for the $2s$. This is because the number of states in the $2p$ and $2s$ bands are respectively six times and two times the number of atoms in the crystal. The $1s$ band lies too low to be shown on this figure; its levels will be concentrated over a very narrow range in energy in keeping with the small splitting suggested in Fig. 12.

It is possible for the energy band arising from one atomic energy level to overlap the energy bands arising from other atomic levels. We shall be concerned below with several cases where this occurs for various crystals. When it does occur the states in the bands become mixed up and it is no longer possible to decide which atomic level was

the parent of each state in the band. This confusion is of no consequence, however, for it does not interfere with using the distribution in energy curves when they are obtained. Furthermore, the conservation of states holds when the bands overlap so that the total number of states per atom in the combined bands is the sum of the number of states per atom in the separated bands. In Fig. 13*b* we represent a distribution qualitatively similar to that occurring in various metals where *s* and *p* bands overlap. The number of states per atom in the combined bands is eight, four for each spin.

A very important distribution-in-energy curve is that of the case of "free electrons." This is the distribution one obtains by imagining that the electrons in a crystal are perfectly free—that is, subjected to no electrostatic forces whatever—but that they are required to remain within a certain prescribed volume. The distribution of quantum states in energy for this case is represented in Fig. 13*c*. At low temperatures the electrons tend to occupy the lowest states consistent with Pauli's principle and the system is referred to as a "degenerate electron gas." With the aid of the distribution curve, the energy and pressure of this gas can be calculated. We shall require its energy for a discussion of the binding energy of sodium, but we shall give here only the equation of the curve, leaving the calculation of the energy until later.⁸ According to the theory, then, for the case of free electrons the number of states per unit energy is given by

$$N(E) = \frac{4\pi V}{h^3} (2m)^{3/2} E^{1/2}, \quad (4)$$

where *V* is the volume of container, *h* is Planck's constant, *m* the mass and *E* the energy of the electron; for free electrons *E* is all kinetic energy, there being no potential energy. For the case of the alkali metals, calculations show that the wave functions for the valence electrons are very similar to the wave functions for free electrons. For these metals we can use Eq. (4) to calculate energies.

Before utilizing the concepts of energy bands in a discussion of the binding energies of crystals, we must define two symbols to be used in describing the energy of a state in the band. For this purpose we arbitrarily separate the energy *E* of a crystal state into two parts: one of these is denoted by *E*₀ and stands for the energy of the lowest state in the band and the other is *E*_{*M*} which stands for the energy which the state possesses in excess of *E*₀—that is, its energy above the bottom

⁸ The reader will find a derivation of this curve given in K. K. Darrow's article "Statistical Theories of Matter, Radiation and Electricity," *Bell System Technical Journal*, Vol. VIII, 672, 1929 or *Physical Review Supplement*, Vol. I, 90 (1929), and in various texts on quantum statistics and the theory of metals.

of the band. We shall find in the next paper that the quantum states in a band represent electrons traversing the crystal with various average speeds. The state E_0 has an average speed of zero. The subscript "M" has been assigned with these ideas in mind and stands for "motion," implying that an electron with energy greater than E_0 has an energy of motion E_M . In general both E_0 and E_M are actually composite energies containing both kinetic and potential energy; only in the case of free electrons is E_M purely an energy of motion. We shall not use in this paper the property of motion connected with the values of E_M ; however, we shall use the division of the energy into two parts, E_0 and E_M , and we shall for convenience refer to the latter as an "energy of motion."

We shall next apply the concept of the energy band to a determination of the binding energies of several types of crystals. It is one of the principal merits of the theory of energy bands in crystals that we can treat many different crystal types on the basis of the same set of ideas. As we shall point out later, however, the band theory is most appropriate for metals: for ionic and valence crystals other theories are better suited.

Energy Bands and Binding Energies of Metals

For several metals the wave functions and distribution of states in energy have been found by solving Schrodinger's equation for the electrons in the metal. We shall discuss sodium since it constitutes one of the simplest cases and is the first metal for which good calculations were carried out.

A sodium atom, Na, contains ten electrons in filled K and L shells and one valence electron in the M shell; its electron configuration is $1s^2 2s^2 2p^6 3s$. When the atoms are assembled together as in the metal, the $3s$ atomic state gives a wide band which overlaps the $3p$ band while the lower levels widen only very slightly.

The formation of the energy bands⁹ is shown in Fig. 14. Since the K and L bands are very narrow, it is possible to neglect the changes in the wave functions of the electrons occupying them and to concentrate upon the valence electrons. The valence electrons then move in a potential field produced by the Na^+ ions and the other electrons.

It can be shown by a lengthy argument that for the case of a monovalent metal, the energy of the metal as a whole is very nearly equal to the sum of the energies of the valence electrons.¹⁰ It is rather

⁹ J. C. Slater, *Phys. Rev.*, 45, 794 (1934).

¹⁰ An exact statement of the situation is too involved for this paper. The reader can find a more complete discussion in Mott and Jones "The Properties of Metals and Alloys," Chapter IV.

natural that the valence electrons should contribute so largely to the binding since the complete shells of electrons making up the Na^+ ion, that is the K and L electrons, are only slightly affected by bringing the atoms together to form a metal. The result that the energy of the metal is the sum of the energies of the valence electrons is

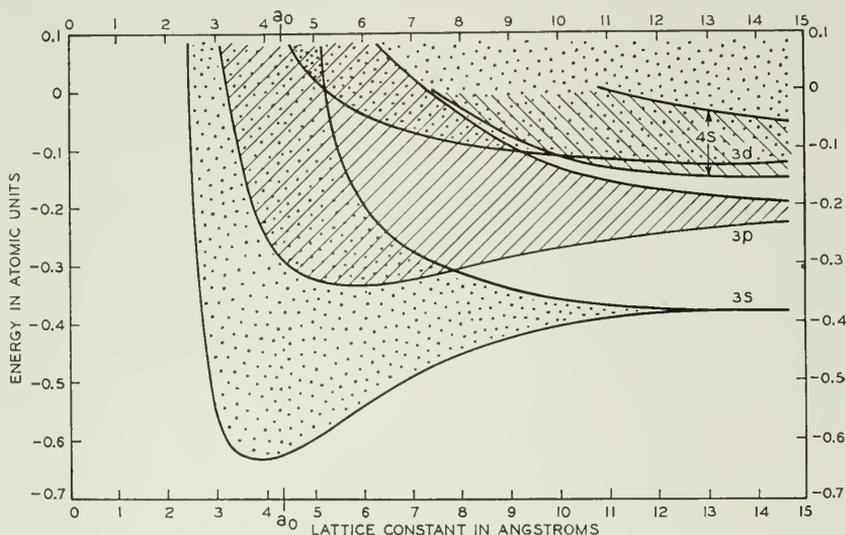


Fig. 14—Energy bands for sodium versus lattice constant.

of great importance in applying the theory. We shall discuss below how the energies of the various states in the band depend upon the arrangement of the atoms; some of these states are occupied and the energy of the crystal can be found by adding the energies of the occupied states. In this way we can find how the energy of the crystal depends upon the arrangement of the atoms and can find what arrangement makes the energy least. According to theory the arrangement of least energy is the stable one and the one which should be found in nature. The remainder of this section will be devoted to discussing the energies of the quantum states in metals and the energies of the electrons which occupy them.

The first satisfactory solutions of Schrodinger's equation for electrons moving in the field of a metal were obtained for sodium by Wigner and Seitz.¹¹ They assumed, in keeping with the findings of experiment, that the sodium atoms were arranged on a body-centered

¹¹ E. Wigner and F. Seitz, *Phys. Rev.*, 43, 804 (1933) and 46, 509 (1934) and E. Wigner, *Phys. Rev.*, 46, 1002 (1934).

cubic lattice. They did not, however, assume that the lattice constant was that given by experiment but instead carried out calculations for each of several assumed values for the lattice constant lying on both sides of the experimental value. The results of their calculations are shown in Fig. 15. The curve marked E_0 is the energy of the lowest

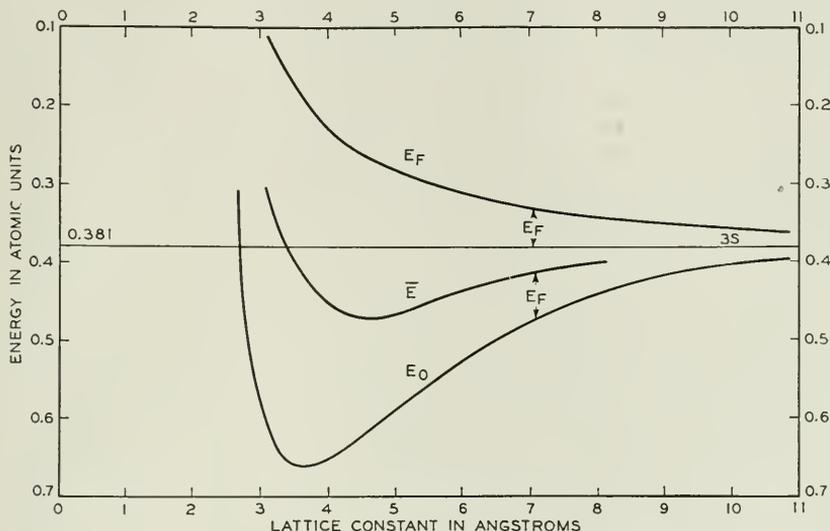


Fig. 15—Energy for sodium versus lattice constant.

level in the valence band. Only two electrons, one with each spin, can occupy this energy level and all others must occupy states of higher energy—that is, only two electrons can have zero value for the “energy of motion” E_M and all others must have larger values. By a method of calculation described below, it can be shown that the average energy of motion of a valence electron is given by the curve marked E_F in the figure. Hence the total energy per valence electron in the metal, which for a monovalent metal is equal to the energy per atom, is represented by the curve marked \bar{E} in the figure; $\bar{E} = E_0 + E_F$. Figure 15 exhibits the dependence of this energy upon the lattice constant. The abscissa of the minimum in the \bar{E} curve gives the theoretically predicted value for the equilibrium lattice constant. The binding energy or heat of sublimation is defined as the energy required to separate the metal into isolated atoms; it is the difference in energy between the minimum of the curve and the value of \bar{E} for infinite lattice constant—that is, for free atoms. Finally, the curvature of the curve at its minimum is a measure of the energy required to compress or expand the crystal and from it a value for the compressibility can

be obtained. In Table I, we compare theoretical and experimental values of lattice constant, binding energy, and compressibility calculated by the method described above. The theoretical values were computed by Bardeen¹² who has added some refinements and corrections to the original calculations.

TABLE I

	Li		Na	
	Calc.	obs.	Calc.	obs.
Lattice Constant (angstroms) . . .	3.49	3.46	4.53	4.25
Heat of Sublimation (Kg. cal./gm. atom)	34	39	23	26
Compressibility (cm ² /dyne)	8.4×10^{-12}	7.4×10^{-12}	12.0×10^{-12}	12.3×10^{-12}

Although the theory can give quite satisfactory values for the various physical quantities shown in Table I, it cannot as yet predict precisely what crystalline form a metal like sodium will take. In carrying out the calculations discussed above, it was assumed that the atoms were arranged in a body-centered cubic lattice. Now the correct theoretical procedure would be to calculate the energy for all conceivable arrangements of the atoms and then to select that arrangement giving the least energy of all as the theoretically predicted equilibrium arrangement. This program is, of course, too laborious to be practical—furthermore experience shows that metals, with but few exceptions, crystallize in one of three forms: body-centered cubic, face-centered cubic, and hexagonal close-packed. For this reason it might be regarded as sufficient to calculate the energies for the face-centered cubic and hexagonal close-packed and to compare these with that for the body-centered cubic. When such calculations are carried out, however, it is found that the minimum energies calculated for the three forms differ among themselves by amounts which are negligible in view of approximations necessary in making the calculations. Hence the theory cannot predict with any certainty which form really has the lowest energy; it does predict, however, that all three forms do have nearly the same energy and gives a value for this energy. Actually the binding energy of sodium must be greatest for the body-centered cubic lattice because this form is the one that occurs in nature and so must be the form of lowest energy. However, it is probable that the difference in energy between the various possible allotropic

¹² J. Bardeen, *Jour. Chem. Phys.*, 6, 367, 372 (1938).

forms for sodium is really very small—so small that we should not expect the present theory to evaluate it. Some indication that the energy of caesium is very nearly the same in the body-centered form and face-centered form (or possibly the hexagonal close-packed form) is furnished by a transformation at high pressures observed by Bridgman. In the next paper we shall meet a case where the theory does seem able to differentiate between the energy of face-centered and body-centered structures. In general, however, the procedure is to use the crystal structure found by x-rays and to calculate the energy for a series of values of the lattice constant as was done for sodium.

We must now return to a discussion of the curves E_0 and E_F of Fig. 15. About the curve E_0 we shall only say that it is obtained by solving Schrodinger's equation for and finding the energy and wave function of an electron in the lowest state in the energy band. The wave function for this state, however, possesses the interesting feature of being very nearly the same as the wave function for a free electron having zero energy of motion. From this fact it is possible to draw the conclusion that the distribution in energy of motion of the valence electrons in sodium is the same as the distribution in energy of free electrons in an electron gas. Accepting this conclusion, we can then use the formulas given for the distribution of states for free electrons in order to calculate the mean energy of motion of the valence electrons in sodium. The results of this calculation, which we give in a footnote,¹³ lead to the energy curve E_F . This energy curve is, from its

¹³ We shall first derive a general expression for E_F without specifying the particular form of $N(E)$. Since in this footnote all energies are measured from E_0 , we shall omit the subscript M from E_M and use simply the symbol E in the equations. The total number, denoted by n , of atoms in the crystal is equal to the total number of valence electrons. Let the volume of the crystal be V . Because of the duplicity due to the spin there are $2n$ states in the band, and half of them will accommodate the n electrons so that the band will be filled only up to a certain energy E_{\max} . We must therefore have

$$n = \int_0^{E_{\max}} N(E) dE. \quad (i)$$

Once the distribution function $N(E)$ is known, this equation serves to determine E_{\max} . The average energy of motion of an electron in these occupied states is, from the definition of an average, the total energy of motion divided by the total number of electrons:

$$E_F = \frac{1}{n} \int_0^{E_{\max}} EN(E) dE. \quad (ii)$$

Substituting the value of $N(E)$ for free electrons into the first equation gives

$$n = \frac{8}{3} V(2mE_{\max}/\hbar^2)^{3/2} \quad (iii)$$

The quantity V/n is the volume per electron which in the case of a monovalent metal

definition, the average energy per electron of a degenerate electron gas. In a degenerate electron gas the electrons have the least possible energy consistent with Pauli's principle and with the distribution of quantum states in energy. For reasons associated with the origin of the statistical mechanics of electrons—that is, with the Fermi-Dirac statistics—the energy E_F is called the “Fermi energy” and given the subscript F . The energy E_F is far greater than the average energy per particle of an ordinary classical gas. We shall see below how this fact accounts for the very small specific heat of the electron gas. From the dependence of the energy upon volume, the pressure of the electron gas can be calculated. It is usually very large, for sodium it is about 50,000 atmospheres. The force that prevents this pressure from blowing the metal apart is represented by the E_0 curve, which gives decreasing energy with decreasing lattice constant and corresponds to a force pulling the atoms together. A more detailed discussion of these forces will be taken up in the third paper of this series.

Other Metals

Calculations similar to those for sodium can be carried out for other metals. The band structure as calculated for copper by Krutter¹⁴ is shown in Fig. 16. Ten electrons per atom can be accommodated in the $3d$ band and two per atom in the $4s$. For copper the $3d$ band is filled—in keeping with the fact that the Cu^+ ion consists of filled K, L, and M shells. From the discussion of molecules given above

is the same as the volume per atom. Denoting by Ω the value of V/n , we find

$$\bar{E}_{\text{max.}} = \left(\frac{3}{\pi}\right)^{2/3} \frac{\hbar^2}{8m} \Omega^{-2/3} = 36.1\Omega_0^{-2/3}\text{eV} \quad (\text{iv})$$

where Ω_0 is the volume per atom in cubic Angstroms. For a body-centered cubic lattice with lattice constant a Angstroms, $\Omega_0 = a^3/2$. Substituting the expression for $N(E)$ into the equation for E_F gives

$$E_F = \frac{3}{5} E_{\text{max.}} = 21.6\Omega_0^{-2/3}\text{eV}. \quad (\text{v})$$

Expressing E_F in atomic units and Ω_0 in terms of the lattice constant, we find

$$E_F = 2.54a^{-2}. \quad (\text{vi})$$

This is the equation of the curve for Figure 15. The values of $E_{\text{max.}}$ calculated from the above equations for a series of metals are

Metal	Li	Na	K	Rb	Cs	Cu	Ag	Au
$E_{\text{max.}}(\text{eV})$	4.74	3.16	2.06	1.79	1.53	7.10	5.52	5.56

¹⁴ H. M. Krutter, *Phys. Rev.*, 48, 664 (1935).

we should expect encroachment repulsions between these ions when their wave functions begin to overlap. In the band picture this repulsion results from the spreading of the $3d$ band; since the band spreads more to higher energies than to lower energies and since it is full, the average energy of an electron in it increases as the lattice constant decreases. Thus the same result, repulsion between closed shells, is

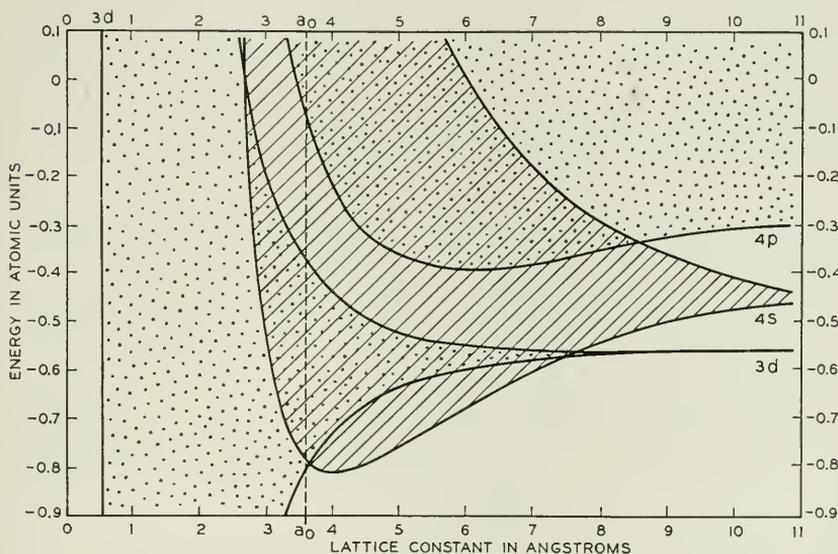


Fig. 16—Energy bands for copper versus lattice constant.

found for the ions in a metal as for the rare gas atoms. For elements whose atoms have partially filled $3d$ levels the situation is quite different. For them only part of the levels of the $3d$ band will be filled and there will be a decrease in the energy of the $3d$ electrons in the metal as compared to the atom. This has been proposed by Seitz and Johnson as an explanation of the fact that the highest binding energies for the metals of a transition series occur for those that have approximately half-filled $3d$ bands and for which consequently nearly all of the $3d$ electrons have lower energies than in the atomic state.¹⁵ The very high melting point metals—columbium, molybdenum, tantalum, and tungsten—come approximately at the middle of their transition series. In Table II we give the binding energies for a number of the transition elements.

¹⁵ F. Seitz and R. P. Johnson, *Jour. App. Phys.*, 8, 84, 186, 246 (1937).

TABLE II

HEATS OF SUBLIMATION FOR SEVERAL METALS INCLUDING THE TRANSITION ELEMENTS
IN KILOCALORIES PER GRAM ATOM *

K 19.8	Ca 48	Se 70	Ti 100	V 85	Cr 88	Mn 74	Fe 94	Co 85	Ni 85	Cu 81	Zn 27.4
Rb 18.9	Sr 47	Y 90	Z 110	Cb —	Mo 160	Ma —	Ru 120	Rh 115	Pd 110	Ag 68	Cd 26.8
Cs 18.8	Ba 49	La 90	Hf —	T 185	W 210	Re —	Os 125	Ir 120	Pt 127	Au 92	Hg 14.6

* Taken from F. R. Bichowsky and F. D. Rossini "The Thermochemistry of the Chemical Substances," Reinhold (1936), except for T which was taken from D. B. Langmuir and L. Malter, *Phys. Rev.* 55, 1138 (1939).

Energy Bands of Diamond

In Fig. 17 we show the band structure for diamond as calculated by Kimball.¹⁶ The configuration of the carbon atom is $1s^2 2s^2 2p^2$ and

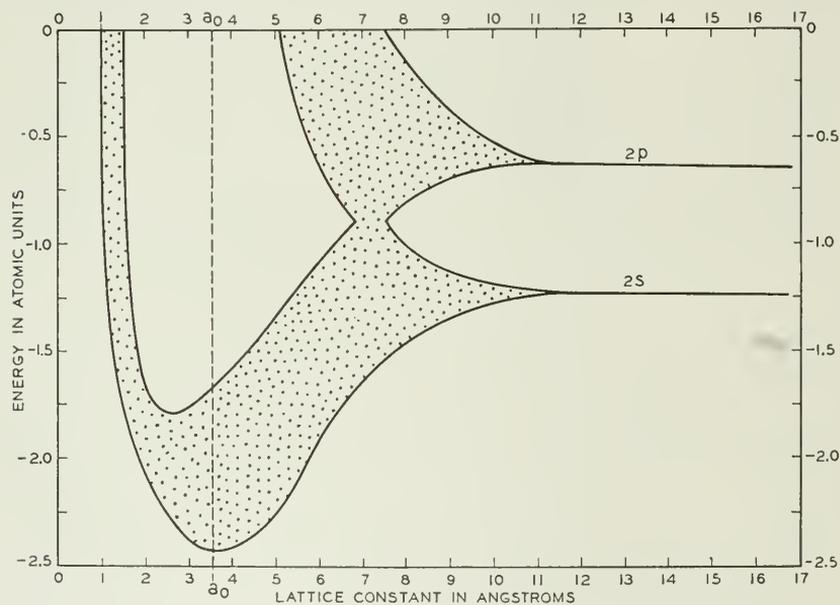


Fig. 17—Energy bands for diamond versus lattice constant.

all four of the L shell electrons are involved in the binding. At large lattice constants the lower band contains two states per atom and the upper six. The lower band is completely filled, the upper only one-

¹⁶ G. E. Kimball, *Jour. Chem. Phys.*, 3, 560 (1935). Some unimportant features resulting from approximations in Kimball's work have been modified in this figure.

third filled. To the left of the crossing of the bands, Kimball finds that both bands contain four states per atom so that the lower is filled and the upper is empty. The actual spacing in diamond occurs to the left of the crossover and, as we shall see in the next paper, the resultant filled band and empty band arrangement explains the absence of electrical conductivity for diamond. The diagram suggests an explanation for the conductivity in graphite; one of the lattice constants of graphite is known to be larger than the abscissa of the crossover of Fig. 17; hence in graphite there are partially filled bands and conduction.

The general downward trend of the bands in Fig. 17 indicates a strong binding energy for diamond; but quantitative calculations of the total energy have not been made.

The type of binding involved in diamond is quite like the binding of metals save that, owing to the absence of partially filled bands, there is no electrical conductivity. In both cases the energy arises from the lowering of energy levels as the atoms come together. In chemical terminology the binding of diamond is referred to as "homopolar" signifying that the atoms are all similarly charged, or rather uncharged. In crystals containing ions rather than neutral atoms, the cohesion is due largely to electrostatic forces and one refers to binding as "heteropolar" or "ionic."

Energy Bands and Binding Energies of Ionic Crystals

The energy band theory can be applied to the calculation of the binding energy of ionic crystals. Before discussing this application, however, it will be instructive to examine a somewhat simpler approach to the problem.

A sodium chloride molecule consists of a sodium ion and a chlorine ion. These ions have charges of $+e$ and $-e$ respectively and have a mutual electrostatic energy of

$$-\frac{e^2}{r}, \quad (5)$$

where r is their distance of separation. This electrostatic energy, which we shall refer to as the "coulomb energy," decreases with decreasing interatomic distance. If the ions are close together, as they are in a molecule, the energy of encroachment due to the overlapping of their closed shells must be considered; this energy increases with decreasing interatomic distance. The equilibrium distance is the one that makes the total energy, coulomb plus encroachment, a minimum.

Closely similar calculations can be carried out for a crystal. One finds the total coulomb energy of all the ions and the total encroachment energy; and then one finds the lattice constant that makes the total energy a minimum. The total encroachment energy is easily found; only atoms which are nearest neighbors in the lattice have appreciable overlapping with each other and it is therefore a straightforward and simple calculation to find the total number of encroachments in the crystal. The coulomb energy is not quite so simply found, however, because the electrostatic interaction of a given ion with its nearest neighbors is no more important than its interaction with its vastly larger number of more distant neighbors. The electrostatic problem is solved as follows: one considers a NaCl lattice which is perfect except for the absence at one lattice point of a Na^+ ion; one finds by known techniques of electrostatics the value at the vacant lattice point of the electrostatic potential due to the remaining ions; this potential is negative and has a value

$$-\phi = -\frac{Me}{4a} = -\frac{3.49e}{a}, \quad (6)$$

where a is the lattice constant and M is a numerical constant known as Madelung's constant, which has a particular value for any special lattice; for the NaCl lattice, $M = 13.94$. If now a Na^+ ion is placed in the vacant lattice point, its electrostatic energy will be $-e\phi$. Similarly the electrostatic potential at a vacant Cl^- lattice point is $+\phi$ and the electrostatic energy of a Cl^- placed there is $-e\phi$. The total electrostatic energy per NaCl molecule in the lattice, however, is not $-2e\phi$ but only $-e\phi$; the factor 2 does not occur since otherwise the electrostatic interaction between each pair of ions would be included twice.¹⁸ The total energy per molecule for the crystal can be found by combining the coulomb and the encroachment energies, and the equilibrium lattice constant and binding energy per molecule thence can be derived.

Using wave functions for Na^+ and Cl^- ions obtained by D. R. Hartree, who has found solutions of Schroedinger's equation numerically, the encroachment energies in NaCl have been evaluated by R. Landshoff.¹⁹ For the lattice constant and binding energy for NaCl he obtains 5.88\AA and $165\text{ Kg.-cal./gm. atom}$ while experiment gives 5.63\AA and $183\text{ Kg.-cal./gm. atom}$.

Some very important theoretical work of a semi-empirical nature has

¹⁸ To see that this is true in a simple case, use the procedure described above to calculate the electrostatic energy of an isolated NaCl molecule.

¹⁹ *Zeits. f. Phys.*, 102, 201 (1936).

been carried out for the alkali halides. In it an analytical expression suggested by theory and containing adjustable constants has been used for the closed shell repulsions. The adjustable constants have been determined from certain data and then used for predictions which can be compared with other data. Using a relatively small number of adjustable constants, Born and Mayer,²⁰ Mayer and Helmholz,²¹ and Huggins and Mayer²² have calculated a much larger number of values for lattice constant and binding energy for many alkali halides with an agreement with experiment of the order of one per cent.

Let us now consider NaCl using the band picture. We shall reach the rather surprising conclusion that there is no fundamental difference between the results obtained from it and those just deduced from the ionic picture described above.

In Fig. 18 we show qualitatively the behavior of the bands for NaCl.²³ In the ionic state, an electron is transferred from the Na $3s$ to the Cl $3p$. The general shifting of the bands is explained as follows. The wave functions corresponding to the Cl⁻ $3p$ band, like all energy band wave functions, are distributed over the whole crystal. They are not, however, equally intense at Na⁺ and at Cl⁻ ions; instead they are definitely concentrated about the Cl⁻ ions. The electrostatic potential at a Cl⁻ ion, due to the remainder of the crystal, has the same value (6) as was found in discussing the ionic method. Since the charge on the electron is $-e$, the energy of each of the states in the Cl⁻ $3p$ band varies with a in the same manner as does $-e\phi$. A similar argument shows that the Na⁺ energy bands vary as $+e\phi$. At a certain lattice constant, the Cl⁻ $3p$ and $3s$ bands and the Na⁺ $2p$ and $2s$ bands begin to widen. Since these bands are full, this widening gives the customary encroachment energy just as it was obtained in the ionic picture. The shifting of the bands similarly gives the coulomb energy. To see this we note that per NaCl molecule there are 18 electrons in the Cl⁻ bands where energies vary as $-18e\phi$ and that there is also one chlorine nucleus with charge $+17e$ whose energy varies as $+17e\phi$. This leaves a net effect of $-e\phi$ for the Cl⁻ ions. Similarly a net effect of $-e\phi$ comes from the electrons and nuclei of the Na⁺ ions. As in the case of the ionic method the sum, $-2e\phi$, of these energies really contains each ionic energy twice and the total electrostatic energy per NaCl molecule is $-e\phi$. So far as calculating energies is concerned, the two methods give equivalent results; the advantage, if any, lies

²⁰ *Zeits. f. Phys.*, 75, 1 (1932).

²¹ *Zeits. f. Phys.*, 75, 19 (1932).

²² *Jour. Chem. Phys.*, 1, 643 (1933).

²³ J. C. Slater and W. Shockley, *Phys. Rev.*, 50, 705 (1936).

with the ionic method rather than the band method because of the more immediate physical interpretation of the former.

We may remark that in the discussion of metallic sodium, it was not necessary to consider the potential energy of the nuclei and the closed

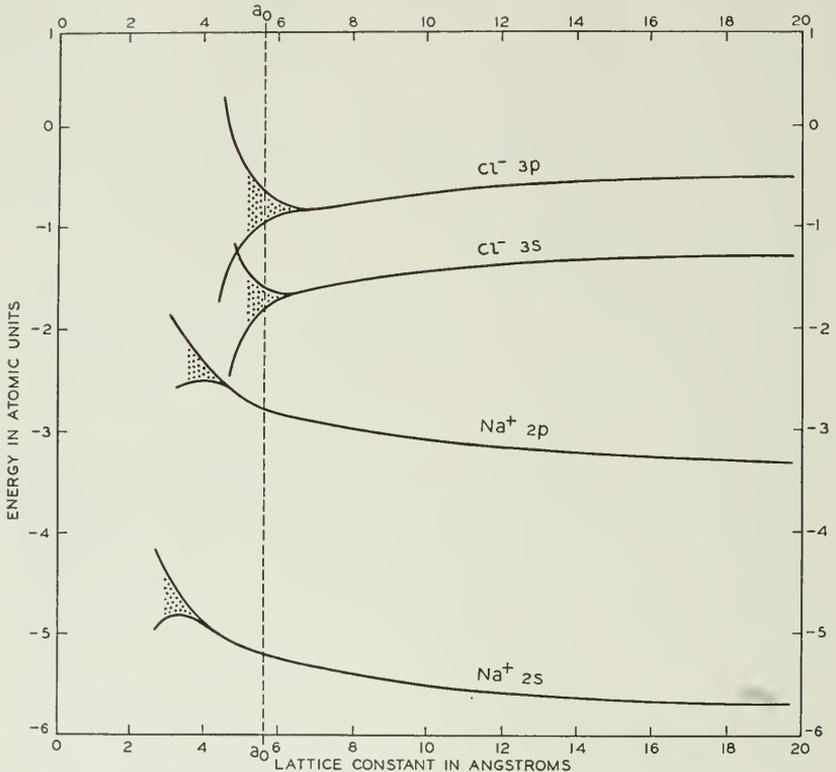


Fig. 18—Energy bands for sodium chloride versus lattice constant.

shell electrons as was done in NaCl. This is because sodium metal is not ionic—although we think of it as consisting in part of Na^+ ions, the electrostatic forces between them are suppressed by the shielding effect of the electron gas. In an ionic crystal, like NaCl, there is no electron gas and the coulomb energy must be considered in the manner described above.

Energy Bands for Other Crystals

There are chemical compounds which lie between the homopolar and ionic types. For example in the sequence of compounds NaF, MgO, AlN, SiC the compounds are progressively less and less definitely ionic—the least ionic, SiC or carborundum, being homopolar. Simi-

larly there are compounds, in particular intermetallic compounds, which are more like metals than like either ionic crystals or valence crystals. Thus there is an intermediate field which connects all three of the simple types of binding. Good computations are lacking for these intermediate cases; we shall return to a discussion of some aspects of them in connection with semiconductors in the next paper.

CONCERNING A CLASSIFICATION OF CRYSTALS

In the last section we saw how the concept of the energy band can explain the binding energies of a number of different types of crystals. Although the band theory has the merit of being very general it has the disadvantage of being at the same time rather abstract. Other theories have been developed to explain the cohesion of particular types of crystals; and, while lacking the generality of the band theory, they have the advantage of a more immediate physical interpretation in their own particular fields. In this section we shall digress from the exposition of the band theory in order to describe briefly some of the simpler viewpoints of the other theories.

We have discussed in the last section three types of binding. Sodium exemplified the metallic type; diamond, the homopolar or valence type; and sodium chloride, the ionic type. The distinction between the valence bond and the metallic bond is not very clearly indicated in the band theory; the only difference there had to do with the degree of filling of the bands. There is another difference, however, which has been long familiar to chemists. The homopolar compounds are usually characterized by "directed valence." Thus the "tetrahedral carbon atom" is a familiar concept of organic chemistry. In crystals in which homopolar binding is dominant the atoms are arranged so that each atom has the proper valence bonds with its neighbors. In diamond each carbon atom is tetrahedrally surrounded by four other carbon atoms. In silicon carbide, carborundum, a similar situation prevails: each carbon is tetrahedrally surrounded by four silicons and vice versa. These crystals are said to have a "coordination number" of four, or $z=4$, meaning that each atom has four nearest neighbors. In crystals of the divalent elements—sulphur, selenium and tellurium—each atom has two near neighbors and the valence condition is satisfied; these crystals have a coordination number of two. The monovalent halogens form crystals in which each atom has one near neighbor, coordination number one. In the metals, however, the neighbors of a given atom are as many as eight or twelve—do these large coordination numbers imply that the metals have eight or twelve electron pair bonds with their neighbors?

According to the quantum mechanical theory of valence, which in itself forms a theory with as many ramifications as the band theory, the electron configuration $1s^2 2s^2 2p^2$ of carbon is especially suited for forming "electron pair bonds" with other atoms. In forming these bonds the wave functions from one atom and another become distorted so as to overlap and form a high electron concentration along the line between the atoms; the energy levels being incompletely filled for the atoms, this overlapping does not produce a repulsion but instead a binding together like that produced by the overlapping wave function in Fig. 9*b* in the hydrogen molecule. The carbon atom is capable of forming four such bonds and forming them most effectively along four lines, making the tetrahedral angles with each other.

Recently Brill²⁴ and his collaborators using x-ray analysis have determined the electron concentration in diamond, in which the carbon atoms are arranged in a tetrahedral manner. The results of their investigations are shown in Fig. 19A.²⁵ It is easily seen that the electrons are concentrated in the bonding directions forming homopolar bonds between the atoms.

The energy band theory, we have said, does not give the clearest picture of the valence crystals; it is, however, especially suited to treatments of the metallic bond. According to the band theory the valence electrons constitute an electron gas—that is, instead of forming electron pair bonds with localized overlapping of the wave functions, they form instead a more or less uniform region of negative charge. In this negative charge the positive ions float. Since the ions repel each other they tend to arrange themselves so as to use their space to best advantage and this requires that they take up one of the "close-packed" arrangements. Let us see why this is true. The close-packed arrangements are those obtained by trying to pack rigid spheres as compactly together as possible. For these arrangements then, the volume per sphere is less than for other arrangements; that is, the close-packed arrangements are the ones which give a minimum volume per sphere for a prescribed value for the distance between sphere centers. Conversely, the close-packed arrangements must be the ones which give a maximum value for the distance between neighboring sphere centers for a given value of the volume per sphere. Since the energy of motion, E_F , of the electron gas and, although we have not shown why, the energy E_0 , depend for a metal mainly upon the volume, in metals we are interested in cases where the volume per

²⁴ R. Brill, H. G. Grimm, C. Hermann and Cl. Peters, *Ann. d. Physik*, 34, 393 (1939).

²⁵ The writer is indebted to Professor Grimm for his permission to reproduce Fig. 19 from his article: *Naturwissenschaften* 27, 1 (1939).

atom is closely prescribed. For a given volume per atom, as we have seen above, the close-packed arrangements are the ones which give the largest separation between neighboring positive ions; and since the positive ions repel each other, the close-packed arrangements will give the lowest energies. This accounts for the fact that the custom-

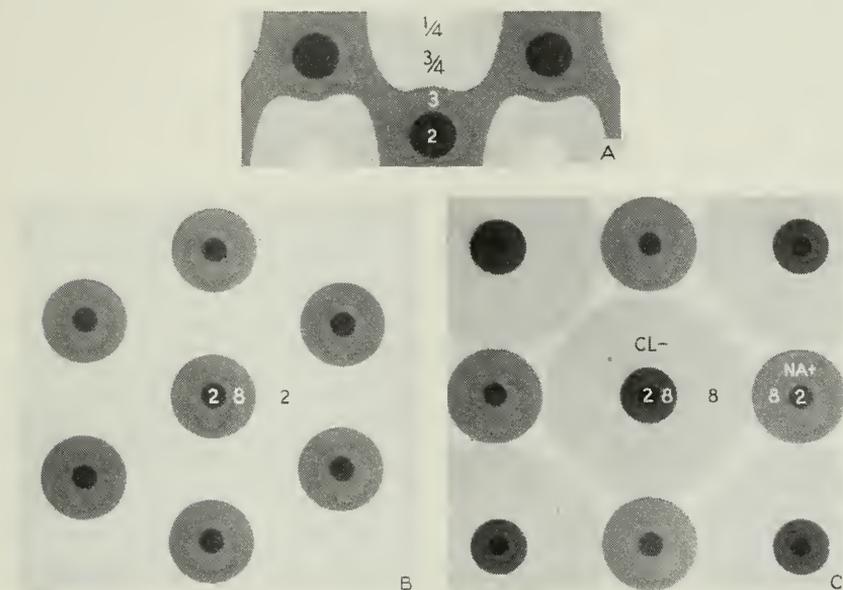


Fig. 19—Electron charge densities in crystals. The numerical values give the number of electrons per atom in the space corresponding to each intensity of shading.

- (A) In diamond.
- (B) In magnesium.
- (C) In sodium chloride.

ary metallic lattices are the body-centered cubic lattice, the face-centered cubic, and the close-packed hexagonal. A further discussion of this physical picture of the nature of the metallic state will be given in the third paper of this series. In Fig. 19B we show qualitatively the electron density of metallic Mg according to Grimm.²⁵ It is seen that the valence electrons give a uniform negative charge in which the positive ions are embedded.

We have seen in the preceding section that the band theory of the alkali halides is essentially equivalent to the ionic theory. A large fund of evidence attests to the validity of the ionic theory, one item being the electron concentrations determined for sodium chloride by Brill²⁴ and his collaborators. These are represented in Fig. 19C; we

²⁵ Loc. cit.

²⁴ Loc. cit.

see that the electrons are closely held about the ions with very little overlapping of the closed shells.

In addition to the metallic, homopolar, and ionic bonds, there is still another interatomic force known as the Van der Waals force—a very weak force compared to the other three. We shall not discuss its origin here except to say that it arises from the spontaneous and mutual polarization of two atoms or molecules when in the neighborhood of each other. It is responsible for the “*a*” term in the Van der Waals equation for gases. When a crystal is formed from organic molecules, such as a crystal of benzene, the forces holding them together are the weak Van der Waals forces. This is the reason why “molecular crystals” have low melting points and binding energies. Although the Van der Waals forces are much smaller than the other three, they are not entirely negligible in comparison and in some of the calculations referred to in the last section, their effects are included.

It is interesting to note that in a single crystal of a given chemical compound, several of the various forces may be operative at once in a rather separable way. A classification of this sort for crystals has been discussed by Grimm.²⁵ For example the crystal mica, which cleaves so naturally into sheets, consists of planes of atoms bound together chiefly by valence forces, the binding between the planes being due to ions lying between and in the planes. Thus mica is held together in two directions by strong valence forces and in the other by weaker ionic forces. In asbestos the atoms are arranged in parallel rows, being held together in the rows by valence forces; the rows, on the other hand, are held to each other by ionic forces. The ionic bonds are more easily broken and asbestos crystals exhibit a typical fibrous structure. Mica and asbestos are intermediate members of a sequence of which diamond with all valence binding and sodium chloride with all ionic binding constitute the extremes. We shall give one more example: cellulose consists of long chains of carbon, oxygen and hydrogen, the chains held to each other by Van der Waals forces; it is an example of valence binding in one direction and Van der Waals binding in the other two.

This section has been a digression, as the main purpose of these papers is to illustrate the band theory of solids. It would hardly be fair to concentrate on this, however, without pointing out, as has been done in this section, that, although the band theory has great generality, it is best adapted for a certain class of solids and that other viewpoints are more natural for solids outside of this class.

²⁵ *Naturwissenschaften*, 27, 1¹(1939).

THERMAL PROPERTIES OF CRYSTALS

In this section, as in the last, we shall digress from a straightforward exposition of the theory of energy bands and discuss the theories of specific heat and thermal expansion. These theories are well worth discussing on their own merits and furthermore their results and methods can be applied later to other topics. Thus the thermal vibrations that account for the specific heat will be shown in the second paper of this series to account for the resistance of metals. The discussion of thermal expansion given here will in the next section on magnetism be extended to an explanation of the unusual expansion properties of magnetic materials, in particular to an explanation of the very small expansion of invar. We shall, however, make use of the band theory once in this section by showing why the free electrons in a metal do not normally make an appreciable contribution to the specific heat.

In the introduction to this paper we pointed out that the specific heat per gram atom of a solid should be by classical theory $3R$ —coming half from the kinetic energy and half from the potential energy of the atoms. This prediction is in reasonable agreement with experiment for many crystals at high temperatures. As the temperature is lowered, however, the observed specific heat decreases in such a way as to approach zero when the absolute zero of temperature is approached. This decrease in the specific heat at low temperatures, as well as the value $3R$ at high temperatures, is readily explained by quantum mechanics. In order to understand the explanation we must inquire into the atomic vibrations of a crystal.

In considering atomic vibrations we are really concerned with the motions of the nuclei. The electrons act as a cement to hold the nuclei in their equilibrium positions and exert restoring forces on them when they are displaced. (We shall see below why the electrons do not partake of the thermal energy.) The nuclei are effectively mass points in this theory and for quantum mechanical reasons, which we shall not discuss, they are incapable of acquiring thermal energy of rotation; hence so far as the crystal vibrations are concerned, we need consider only their translational or rectilinear motions. A crystal containing N atoms has $3N$ degrees of freedom since each nucleus can move in three dimensions. In order to find the specific heat of a crystal we must find the normal modes of vibration. The system of coupled oscillators in Fig. 11 represents reasonably well the normal modes of vibration for a one dimensional crystal whose atoms have only one degree of freedom. There is a similar set of normal modes for

a three dimensional array of atoms and, once the forces between the atoms are known, the frequency of vibration of each of the modes can be found. This means that so far as thermal vibrations are concerned, we can consider the crystal as equivalent to a set of $3N$ oscillators whose frequencies are those of the normal modes. We must next discuss the specific heat of a single oscillator.

According to classical statistical mechanics, a harmonic oscillator in a temperature bath at absolute temperature T will have an average thermal energy equal to kT , where k is Boltzmann's constant. The value kT is only an average value, we emphasize, and the oscillator will have other energies some of the time, the probability of each energy being given by known equations. The probability is very small, however, that the oscillator acquires more than two or three times kT of thermal energy. In a very large system of oscillators, the fluctuations of energy of the oscillators tend to cancel out and the probability of any appreciable fractional deviation of the total energy from its mean value is very small. If N is the number of molecules in a gram molecule ($N = 6.06 \times 10^{23}$), then $Nk = R$, the gas constant, = 1.99 cal. per gm. molecule per degree C. Hence the energy of $3N$ oscillators is $E = 3NkT = 3RT$ and the specific heat is $C = dE/dT = 3R$; this classical result that the specific heat of one gram atom of solid is $3R$ is known as the DuLong-Petit Law.

According to quantum mechanics, an oscillator of frequency ν has a set of quantum states whose energies are $\frac{1}{2}h\nu$, $(1 + \frac{1}{2})h\nu$, $(2 + \frac{1}{2})h\nu$, etc. The oscillator can take on only these energies. If it is in a heat bath of temperature T , however, it will sometimes have one allowed energy and sometimes another and as for the classical case we shall be concerned with its average energy. At absolute zero, the average energy is, of course, $\frac{1}{2}h\nu$. Now the probability of the oscillator gaining much more than kT of thermal energy is very slight. Hence the average energy of the oscillator remains at $\frac{1}{2}h\nu$ until thermal energy becomes large enough to excite it to the next state which is $h\nu$ higher, and consequently so long as kT is much less than $h\nu$ the quantum oscillator acquires much less thermal energy than would a classical oscillator. For kT much greater than $h\nu$, the oscillator will spend an appreciable fraction of its time in many of the quantum states and, as may be shown mathematically, the quantum restriction is no longer of importance so far as the average energy is concerned and the value kT is obtained just as in the classical case. In Fig. 20 the dependence upon temperature of the average energy and the specific heat for a quantum oscillator are shown.

The specific heat of the crystal is just the sum of the specific heats of its oscillators. Since the oscillators have different frequencies they have different specific heats and in order to add up the specific heats of all of them it is necessary to know how the various frequencies of the

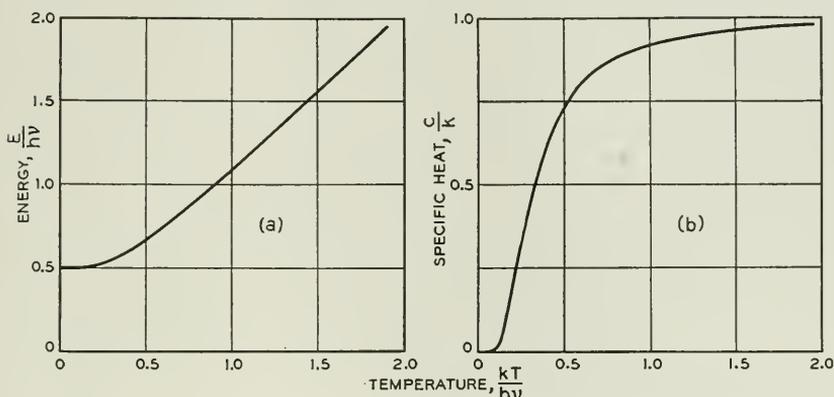


Fig. 20—Thermal behavior of an oscillator according to quantum mechanics.

- (a) Energy versus temperature.
 (b) Specific heat versus temperature.

oscillators are distributed. Once this distribution in frequencies is known it is merely a matter of summation to find total specific heat. The problem of finding the distribution in frequency of the oscillators was first solved by Debye. The low-frequency vibrations are very simply found for they are merely the acoustic vibrations of the crystal; they are very similar to the normal modes shown for the square membrane of Fig. 7g. For these low-frequency vibrations it can be shown by a straightforward argument, which is too long to give here, that the number dN of oscillators whose frequencies lie between ν and $\nu + d\nu$ is

$$dN = V4\pi \left(\frac{2}{C_T^3} + \frac{1}{C_L^3} \right) \nu^2 d\nu, \quad (7)$$

where C_T and C_L are the velocities of transverse and longitudinal waves in the solid and V is its volume.²⁶ Debye assumed that this distribution held for all the normal modes. There is of course a highest frequency of vibration, ν_{\max} , and the total number of normal modes must be $3N$; hence Debye concluded that

$$\begin{aligned} 3N &= \int_0^{\nu_{\max}} V4\pi \left\{ \frac{2}{C_T^3} + \frac{1}{C_L^3} \right\} \nu^2 d\nu \\ &= V4\pi \left\{ \frac{2}{C_T^3} + \frac{1}{C_L^3} \right\} \frac{\nu_{\max}^3}{3}. \end{aligned} \quad (8)$$

²⁶ For a derivation see P. Debye, *Ann. d. Physik*, 39, 789 (1912).

From this equation ν_{\max} can be found if N/V and the velocities C_T and C_L are known. Knowing ν_{\max} and the distribution in frequency, Debye summed the specific heats of all the oscillators and obtained the specific heat of the solid. According to this theory the specific heat vanishes at $T = 0$ and is proportional to T^3 near $T = 0$. At high temperatures it approaches the classical value of $3R$. A measure of the temperature at which the classical value is closely approached is given by the maximum frequency of atomic vibration ν_{\max} ; when kT is greater than $h\nu_{\max}$, all the modes of vibration including the highest make substantial contributions to the specific heat. The temperature at which this occurs is known as the Debye temperature and denoted by the symbol θ_D ; obviously $\theta_D = h\nu_{\max}/k$. The specific heat given by Debye's equation is a function of T/θ_D only and can thus be represented by the expression $C(T/\theta_D)$; so that by this theory all crystals should have the same curve for specific heat versus temperature except for changes in the temperature scale corresponding to the different values of their Debye temperatures.

TABLE III

DEBYE TEMPERATURES IN DEGREES KELVIN USED IN FIGURE 21

Pb 88	Tl 96	Hg 97	J 106	Cd 168	Na 172	KBr 177
Ag 215	Ca 226	KCl 230	Zn 235	NaCl 281	Cu 315	Al 398
Fe 453	CaF ₂ 474	FeS ₂ 645	C 1860			

In Fig. 21 is shown a compilation of specific heat data.²⁷ For each substance a value of θ_D (given in Table III) has been chosen so as to obtain the best agreement with experiment and the values of the specific heat have then been plotted as a function of T/θ_D . The Debye theory relates to specific heat at constant volume and in it no allowance is made for the energy due to thermal expansion. The experimental points are derived from measurements of specific heat at constant pressure which have been transformed by using a thermodynamical relationship so as to give specific heat at constant volume.

For these curves θ_D was chosen so as to obtain the best fit. It is, however, possible to calculate θ_D from theory by using the elastic constants of the material to evaluate C_T and C_L and then substituting in Eq. (8). For sodium the elastic constants have been calculated entirely from theory by the methods described in the section on

²⁷ Taken from E. Schroedinger, *Handbuch der Physik*, Vol. X, p. 307 (1926).

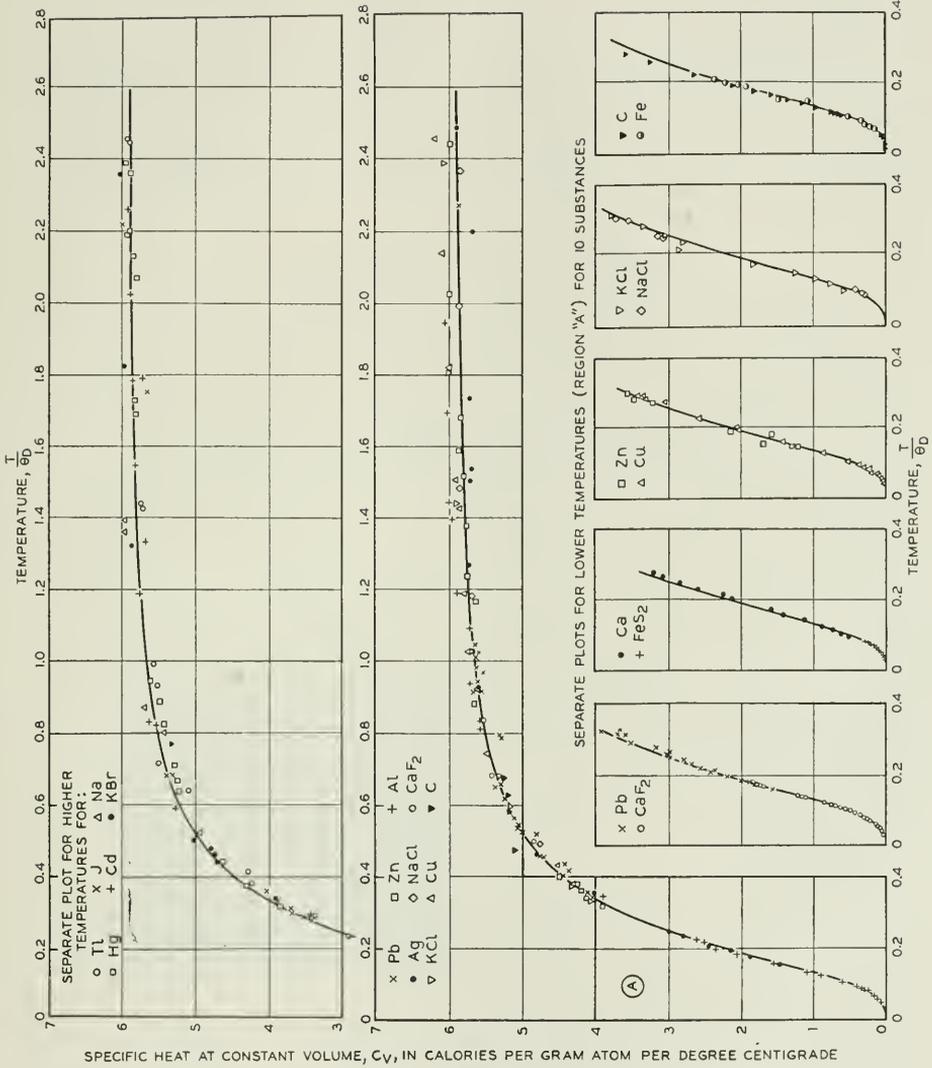


Fig. 21—Comparison of the Debye specific heat curve with experiment.

"Electrons in Crystals" and extensions of them* to be discussed in the third paper of the series. Using the theoretical values one obtains a value of 143° K for θ_D , whereas the value that fits experiment best is 172° K.

Recently calculations have been made from a model of the crystal as an assemblage of atoms rather than as a continuum as postulated in deriving Eq. (7)—that is, a model like the coupled oscillators, rather than like the stretched membrane, is used. These calculations, principally by Blackman, have explained some discrepancies between the Debye theory and experiment.

The Specific Heat of the Electrons

We must now see why the electrons contribute only slightly to the specific heat. Let us consider a case like that of sodium where we have a partially filled band. At the absolute zero of temperature, the electrons will fill all the levels below a certain energy E_1 and all the higher levels in the band will be empty (Fig. 22a). Now at temperature T some of the electrons will be excited to higher states; since, however, an electron cannot gain more than about kT of energy thermally, only those electrons whose energies lie in a range kT below E_1 can be excited. Electrons occupying states farther down in the band cannot acquire kT of thermal energy for, if they did so, they would have to move to states already occupied and such an act is forbidden by Pauli's principle. In order to demonstrate what a small fraction of the electrons can gain energy thermally, we point out that the width of the energy band is usually 4 or 5 ev while the value of kT in electron volts is $T/11,600$ and room temperature corresponds to a kT of about .03 ev. The electrons which do gain thermal energy have a normal value for the specific heat but constitute only about one per cent of all the valence electrons.

It might be maintained that the above argument is specious and that the electrons could all gain energy kT ; this would not violate Pauli's principle because the electrons would move upward in the band as a unit, each moving into a state vacated by another electron. This contention is found to be wrong; one finds by using the statistical mechanics appropriate to electrons that the distribution of the electrons among the energy levels is given by the Fermi-Dirac distribution function.²⁸ According to this, the distribution of the electrons among the levels would be as indicated in Fig. 22b. The proba-

* K. Fuchs, *Proc. Roy. Soc.* 157, 444 (1936).

²⁸ For a discussion of the Fermi-Dirac statistics see K. K. Darrow, *The Bell System Technical Journal*, Vol. VIII, p. 672, 1929, or *The Physical Review Supplement*, Vol. I, p. 90, 1929.

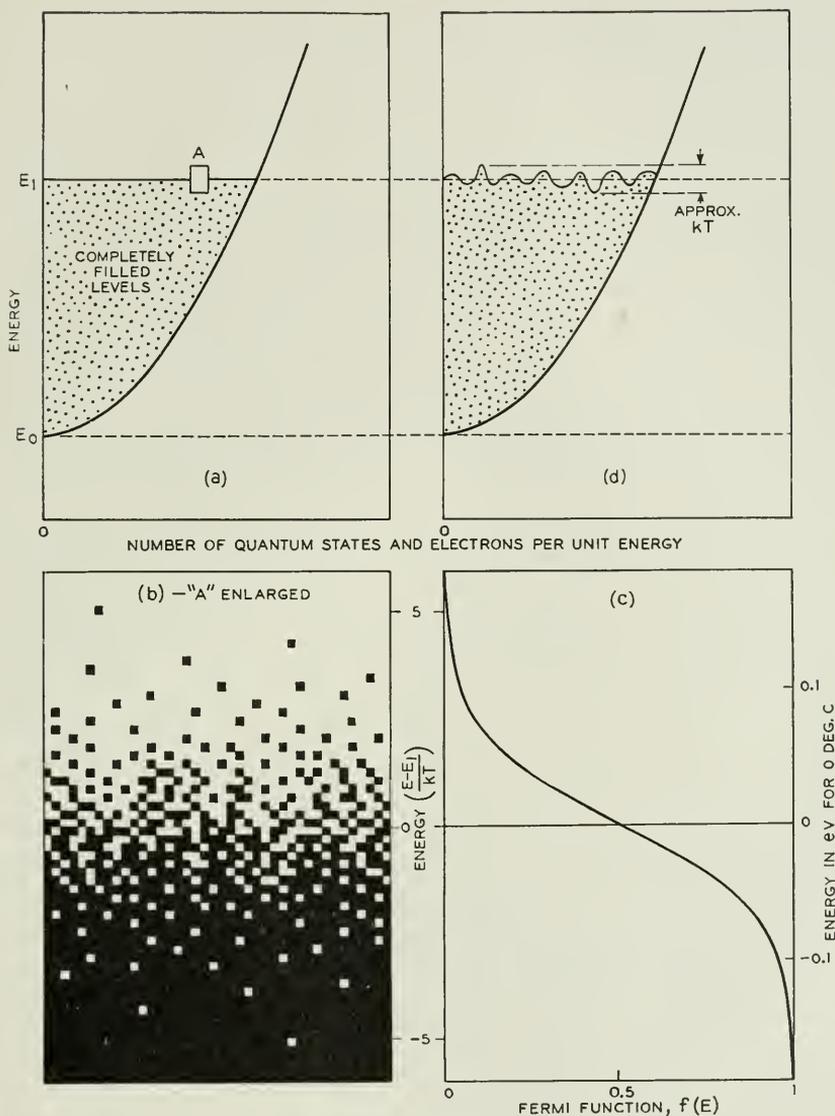


Fig. 22—Specific heat of the electrons.

- (a) Distribution of electrons in energy for the absolute zero of temperature.
- (b) Enlargement of part of (a) but for room temperature; each unit of area represents a quantum state.
- (c) The Fermi distribution function.
- (d) A water tank analogue.

bility that any particular energy state be occupied is given by the Fermi-Dirac factor f

$$f = \frac{1}{e^{\frac{E-E_1}{kT}} - 1} \quad (9)$$

This factor is shown in Fig. 22*c* and the corresponding filling of energy levels is shown schematically in 22*b*. A physical picture which is helpful in understanding this result may be obtained by considering the distribution of energy levels, Fig. 22*a*, to be the cross-section of a trough or tank. If we pour water into this tank it will fill to a certain level, E_1 . If we let each molecule of water in the tank represent an electron in the crystal, then the distribution in energy of the electrons is correctly represented by the distribution in height of the molecules. Thermal agitation is represented by shaking the tank; this will produce surface ripples as in Fig. 22*d* which represent crudely the Fermi-Dirac distribution.

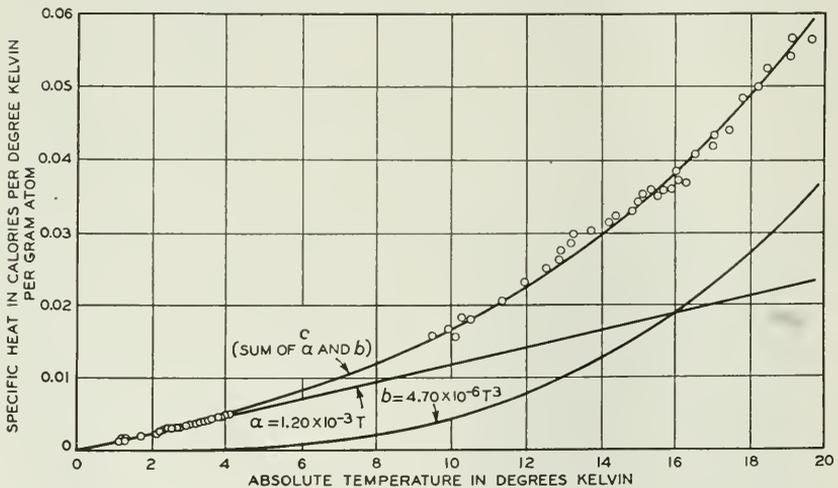


Fig. 23—Specific heat of iron at low temperature.

Under certain conditions, however, the electronic specific heat is not negligible. We have seen that the number of electrons participating in specific heat is proportional to kT and that these have a more or less normal specific heat. Hence the electronic specific heat is proportional to T . On the other hand, at low temperatures the Debye specific heat is proportional to T^3 . Hence for sufficiently low temperatures the electronic specific heat is the larger. In Fig. 23 we give the specific

heat of iron near absolute zero.²⁹ The theoretical curve *c*, which is seen to represent the experimental data quite well, is the sum of two terms represented by curves *a* and *b*. *a* is linear in the temperature and represents the electronic specific heat while *b* is cubic and represents that due to lattice vibrations. Numerical calculations from theory of the slope of curve *a* which could be compared with the observed slope are not available. Curve *b*, we have said, is just the Debye curve and is drawn as if the Debye temperature were 462°, a value which is in good agreement with 453°, the value deduced from the specific heat at higher temperatures in connection with Fig. 21. At very high temperatures the electronic specific heat will again be of importance. But at high temperatures it is necessary to apply corrections to the Debye theory and the writer is not acquainted with any unambiguous evidence for electronic specific heat in that case.

Thus we see that only a very small fraction of the electrons of a partially filled band contribute to the specific heat. It is the Pauli principle which restrains the remainder. We shall see in the next paper why the Pauli principle does not interfere with the conduction of electricity. For the case of an insulator—that is, a crystal each of whose bands is either wholly filled or wholly empty—it is still harder for electrons to arrive at empty states and the electronic specific heat is quite negligible. Hence all of the specific heat for an insulator is of the atomic vibration type discussed in the Debye theory.

The Theory of Thermal Expansion

In order to understand the theory of thermal expansion we must study the curve representing energy versus lattice constant for the solid. This is shown qualitatively in Fig. 24. We note that the energy curve is unsymmetrical about its minimum. We may describe its behavior by saying that it is harder to compress than to expand the solid. This statement is illustrated by a comparison of the expansion and the compression which can be produced by a given energy *E*; it is seen that the asymmetry of the curve causes the expansion produced by this energy to be greater than the compression. Now the origin of thermal expansion is as follows: owing to thermal agitation—that is, atomic vibrations—regions of the crystal are alternately expanding and contracting; since the expansions occur more readily than the contractions, there is on the average a net expansion. The greater the temperature the greater this net expansion; hence we find that the size of the solid increases with increasing temperature. This explanation of thermal expansion can be made clearer by considering, not a solid,

²⁹ W. H. Keesom and B. Kurrelmyer, *Physica* 6, 633 (1939).

but a diatomic molecule. Suppose Fig. 24 gives the dependence of the energy of a molecule upon the internuclear distance. Suppose the molecule is given vibrational energy corresponding to E on the figure.

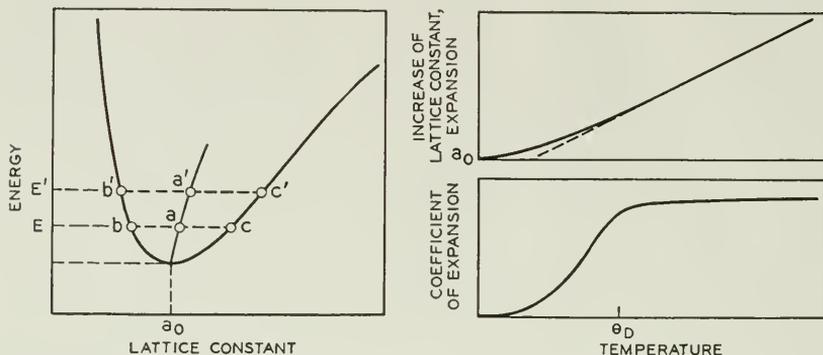


Fig. 24—The theory of thermal expansion. The asymmetry of the curve for the energy of a crystal versus the lattice constant is responsible for the thermal expansion.

Then the nuclei will vibrate between positions b and c on the figure. Since c lies more to the right of the equilibrium position than b does to the left, the mean distance of separation, a , lies to the right of a_0 . Increasing the vibrational energy to E' increases the mean separation to a' . This shows that the asymmetry of the potential curve results in a continuous increase in mean internuclear separation with increasing energy of vibration. A crystal is, in a sense, an assemblage of diatomic molecules, each pair of nearest neighbors having a potential energy curve like that of Fig. 24, and its expansion is explained in the same way.

The theory outlined above can be made quantitative. From it we obtain the interesting result that the thermal expansion coefficient is proportional to the specific heat. This is a rather natural result: we have seen that the total expansion is proportional to the thermal energy; hence the rate of expansion with increasing temperature, i.e. the thermal expansion coefficient, should be proportional to the rate of increase in thermal energy with increasing temperature, i.e. to the specific heat. The relationship embodying this statement is known as Grüneisen's law and is expressed by the equation

$$\alpha = \gamma \frac{K}{V} C_v, \quad (10)$$

where α is the volume coefficient of thermal expansion (three times the linear coefficient), K is the compressibility, V the volume of one gram

atom, and C_V the specific heat per gram atom at constant volume. γ is a parameter which measures the asymmetry of the curve and is defined as follows: if we think of the solid as being compressed by an external pressure, the forces between the atoms will change and the Debye temperature will increase. If the curve were a parabola—that is, perfectly symmetrical—the Debye temperature would not change. γ is defined by the relationship

$$\gamma = - \frac{\partial \ln \theta_D}{\partial \ln V}. \quad (11)$$

The γ , K , and V are nearly constant for a given substance. Hence the thermal expansion curve is practically the same as the specific heat curve except for a constant factor. In Fig. 25 we give the thermal

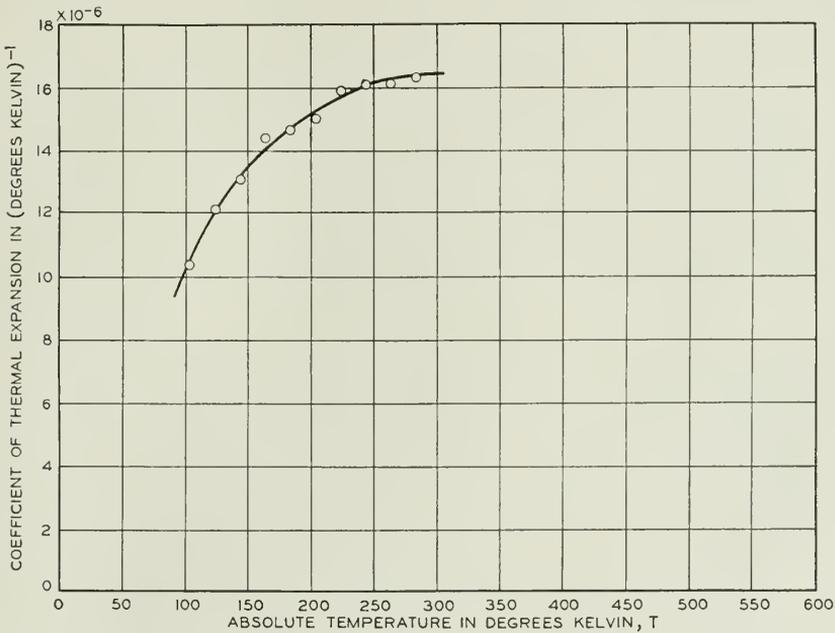


Fig. 25—Coefficient of thermal expansion versus temperature for copper.

expansion of copper.³⁰ The Debye temperature was chosen to give the best fit. We see that the theory of thermal expansion gives as good agreement with experiment as does the theory of specific heat. If Grüneisen's law were perfectly satisfied, the same Debye temperature would be found for both the thermal expansion and the specific heat curves. The relatively small difference between the two values, 325 for expansion and 315 for specific heat, is a measure of the validity of Grüneisen's law.

³⁰ E. Grüneisen, *Handbuch der Physik*, X, p. 43 (1926).

Grüneisen's law applies only to simple crystals; we shall see in the next section that it is not applicable to the anomalous expansion associated with ferromagnetic transformations nor is it applicable to the abnormal expansions of the order-disorder transformations in alloys.

MAGNETIC EFFECTS

In this section we return to a discussion of the energy band theory and this time introduce the magnetic moment associated with the spin of the electron. It is the spin magnetic moment which when added to the concept of energy bands leads to explanations of para and ferromagnetism.

When a body is placed in a magnetic field it becomes magnetized; in other words it acquires a magnetic moment. Ferromagnetic materials become very easily magnetized in the field with their magnetic moments parallel to the field and they may remain magnetized after the field is removed. Paramagnetic materials are also magnetized in the direction of the field but only very weakly compared to ferromagnetic materials and only while they remain in the field. Diamagnetic materials are magnetized in a direction opposite to the field and, like paramagnetic substances, only weakly and while in the field. These magnetic effects are produced by the electrons in two distinct ways. In the first place, the motion of the electron as a whole produces a current and this current, like the ordinary macroscopic currents in a wire, produces a magnetic field. Conversely, an externally applied magnetic field affects the motions of the electrons in a body and can thereby magnetize it; this process accounts for the diamagnetism of diamagnetic bodies but it may contribute to the paramagnetism as well. It is not with this first way in which electrons can behave magnetically but rather with the second way, described below, that we shall be concerned. The first way, which is mentioned for completeness, involves a theory too complicated for treatment in this article. In the second place, an electron can behave magnetically by virtue of its spin: the rotation of the electron about its own axis produces a magnetic moment which is anti-parallel—because the charge of the electron is negative—to the angular momentum due to the spin. A magnetic field tends to align the spin magnetic moments of the electrons and to make them contribute to the paramagnetism. We shall see below that this process accounts for the paramagnetism of non-ferromagnetic metals. We shall see also that the magnetism of ferromagnetic bodies is due to the magnetic moment of the electron spin but that the energy involved in the theory of ferromagnetism is not an interaction between the magnetic dipoles of the electrons but is

instead an electrostatic exchange energy like that discussed for atoms in connection with Figs. 4 and 6.

Paramagnetism and Diamagnetism

Let us consider first the so-called "weak spin paramagnetism." This occurs in metals, since they have partially filled bands. In the presence of a magnetic field the spin of the electron is quantized so that the component of its angular momentum in the field direction is either $+\frac{1}{2}\hbar$ or $-\frac{1}{2}\hbar$ where $\hbar = h/2\pi$ ($h =$ Planck's constant) is the quantum mechanical unit of angular momentum. The corresponding components of magnetic moment along the field are $-\mu_\beta$ and $+\mu_\beta$ where μ_β is the quantum mechanical unit of magnetic moment known as the Bohr magneton. Letting $-e$ be the charge and m the mass of the electron and c be the speed of light, we have from the quantum theory

$$\mu_\beta = e\hbar/2mc. \quad (12)$$

The ratio of mechanical moment (i.e. angular momentum) to magnetic moment, taken without regard to sign, is called the "gyro-magnetic ratio." For the spin of the electron its value is mc/e , but for the motion of the electron as a whole, its value is $2mc/e$. Because of the difference between these two values, experimental measurements of the gyro-magnetic effect play a decisive role in the experimental verification of the electron spin theory of ferromagnetism in a way which we shall describe below.

Half the quantum states in an energy band of a crystal have angular momentum components along the magnetic field of $\frac{1}{2}\hbar$ and the other half of $-\frac{1}{2}\hbar$. In Fig. 26*a*, we have divided the states in the band into two groups, corresponding to the two spins. We shall refer to one of these as "the band with plus spin" and to the other as "the band with minus spin." When a magnetic field is applied, the energies of the electrons are changed. Thus if an electron in the lowest state of the band with minus spin has an energy E_0 before the field is applied, it has an energy of $E_0 - \mu_\beta H$ afterwards; the second term represents, of course, the energy of the magnetic dipole μ_β when parallel, as distinguished from anti-parallel, to the field—the situation for minus spin. All the states in the band with minus spin will be thus altered in energy. Similarly all the states in the band with plus spin are displaced upwards in energy by $\mu_\beta H$. This is the situation represented in Fig. 26*b*. After the displacement we find that some of the electrons in the band with plus spin have higher energies than empty states in the band with minus spin; such an arrangement is not stable and the electrons will change their quantum states so as to produce the lowest energy possible consistent with the distribution of energy levels shown

in Fig. 26*b* and with Pauli's principle. The arrangement of lowest energy is shown in Fig. 26*c*; electrons have shifted from the band of plus spin to states of lower total energy in the band of minus spin until the two bands are filled to the same energy level, indicated by the solid horizontal line. As the figure shows, the number of electrons shifted will be the number lying in the energy range $\delta E = \mu_\beta H$.³¹

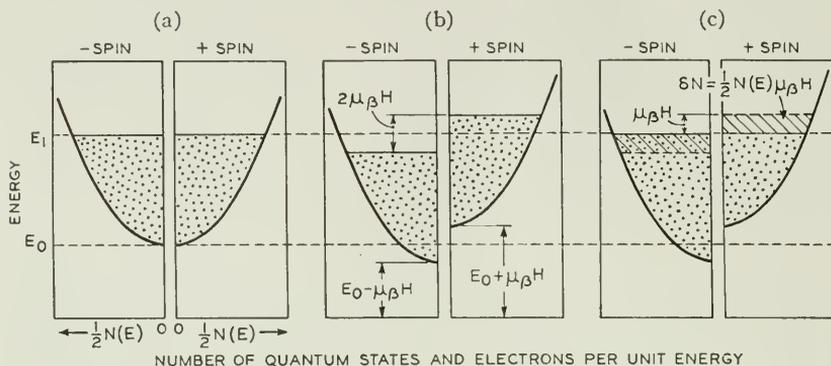


Fig. 26—The paramagnetism of free electrons.

- (a) Distribution of electrons in energy.
 (b) Displacement of levels by a magnetic field.
 (c) Distribution of electrons in energy in a magnetic field.

The number of states, δN , lying in this energy range in the band of plus spin, which contains of course half the states in the band, is according to equation (3)

$$\delta N = \frac{1}{2}N(E_1)\delta E = \frac{1}{2}N(E_1)\mu_\beta H. \quad (13)$$

The magnetic moment of these states is

$$\delta M_+ = -\mu_\beta \delta N = -\frac{1}{2}N(E_1)\mu_\beta^2 H. \quad (14)$$

The minus sign occurs because the angular momentum and the magnetic moment of an electron are in opposite directions; the states of plus spin have minus moments in Fig. 26.

The electrons that occupied these states before the field was applied now occupy states with minus spin and produce a magnetic moment of

$$\delta M_- = \frac{1}{2}N(E_1)\mu_\beta^2 H. \quad (15)$$

Hence the minus band gains a plus moment and the plus band loses a

³¹ We have here assumed that the fractional change in $N(E)$ in the interval $\mu_\beta H$ is negligible; this assumption is reasonable. For a field of 10,000 gauss, $\mu_\beta H$ is only 5.77×10^{-5} ev while $E_1 - E_0$ is of the order of several ev.

minus moment and, since the net moment of Fig. 26a is obviously zero, the net moment produced by the magnetic field is

$$\delta M = \delta M_- - \delta M_+ = N(E_1)\mu_\beta^2 H. \quad (16)$$

The susceptibility of a material, denoted by χ , is defined as the magnetic moment produced per unit volume per unit field:

$$\chi_s = \frac{\delta M}{VH} = \frac{N(E_1)}{V}\mu_\beta^2. \quad (17)$$

The subscript “s” is a reminder that this susceptibility was produced by the spin magnetic moment of the electron.

Since the moment produced is in the direction of the field, χ_s is positive; the susceptibility is of the paramagnetic type. As for its magnitude: in the monovalent metals, as we have said before, the distribution of levels in the bands is well approximated by the free electron formula (4). Using this, we find

$$\chi_s = \frac{4\pi}{h^3} (2m)^{3/2} (E_{\max})^{1/2} \mu_\beta^2. \quad (18)$$

where E_{\max} ($= E_1 - E_0$) is the maximum kinetic energy in the band.

Before comparing susceptibilities calculated from this expression with experimental values, we must discuss diamagnetism. The electrons in the partially filled band of Fig. 26 give formula (18) because of their spin magnetic moments. They give a susceptibility also because of their motion through the crystal. For the case of free electrons, this susceptibility is negative—that is, it is a diamagnetic susceptibility, and, according to a theory we cannot discuss here, in magnitude it is one third of χ_s . Denoting it by χ_m (“m” for motion of the electron as a whole), we have

$$\chi_m = - (1/3)\chi_s. \quad (19)$$

The electrons in the filled bands, corresponding to electrons in closed shells in the ionic cores of the metal, also give rise to diamagnetism. They can give no spin paramagnetism because there is no possibility of transferring electrons from a *filled* band of one spin to a *filled* band of the other spin—this would require putting more electrons in the band of one spin than it has quantum states, a violation of Pauli’s principle. Denoting by χ_i the susceptibility of the ionic cores of the metal, we have for the net susceptibility χ the equation

$$\chi = \chi_s + \chi_m + \chi_i. \quad (20)$$

Specializing this for the case of free electrons in the valence electron

band gives

$$\chi = (2/3)\chi_s + \chi_i = \frac{8\pi}{3} \left(\frac{2m}{\hbar^2} \right)^{3/2} \mu_B^2 (E_{\text{max}})^{1/2} + \chi_i \quad (21)$$

In Table IV we give theoretical and experimental values for the susceptibilities of the simple metals. The values of χ_i are obtained from theory for lithium and by experiment for the other metals.

TABLE IV
MAGNETIC SUSCEPTIBILITIES *

	Li	Na	K	Rb	Cs
χ_s	1.5	0.68	0.60	0.32	0.24
χ_i	-0.1	-0.26	-0.34	-0.33	-0.29
$\chi = \frac{2}{3}\chi_s + \chi_i$	0.9	0.2	0.06	-0.12	-0.15
χ observed †	0.5	0.51	0.40	0.07	-0.10

* This Table is taken from N. F. Mott and H. Jones, "The Theory of the Properties of Metals and Alloys," Oxford 1936, p. 188.

† K. Honda, Ann. d. Physik 32, 1027 (1910) and M. Owen, Ann. d. Physik 37, 657 (1912).

Although equation (17) for the spin susceptibility χ_s in terms of $N(E_1)$ is generally true, the relationship that $\chi_m = -\chi_s/3$ is true only for the case when $N(E)$ is the free electron distribution.³² For some metals $N(E)$ differs greatly from that for free electrons and then larger values of χ_m may occur. The high diamagnetism of bismuth is explained in this way. In the next paper, we shall discuss the meaning of the freeness of electrons; however, a discussion of electron diamagnetism lies beyond the scope of this paper.³³

Ferromagnetism

The shift of electrons from one band to another for the paramagnetic behavior shown in Fig. 26 persists only so long as the magnetic field is applied. When the magnetic field is removed, the stable arrangement is as shown in Fig. 26a, equal numbers of electrons having each spin. The situation is quite different in ferromagnetic materials and, for reasons discussed below, in the stable arrangement there are many more electrons of one spin than of the other.

Two things are important for the occurrence of ferromagnetism: the exchange effect as illustrated in Figs. 4 and 6 and the structure of the bands arising from the 3*d* levels. The 3*d* levels, as is shown in

³² An even more stringent condition is actually required.

³³ For the diamagnetism of electrons in closed shells the reader is referred to K. K. Darrow's article, "The Theory of Magnetism," *Bell System Technical Journal*, Vol. XV, 1936, and in particular to Page 247.

Fig. 6, are only partially filled for free atoms of the ferromagnetic elements iron, cobalt, and nickel. We shall show first how the partially filled $3d$ bands together with exchange forces can produce ferromagnetism and later discuss the theory of why only the last three of the eight transition elements are ferromagnetic.

The splitting of the atomic energy levels into bands is shown in

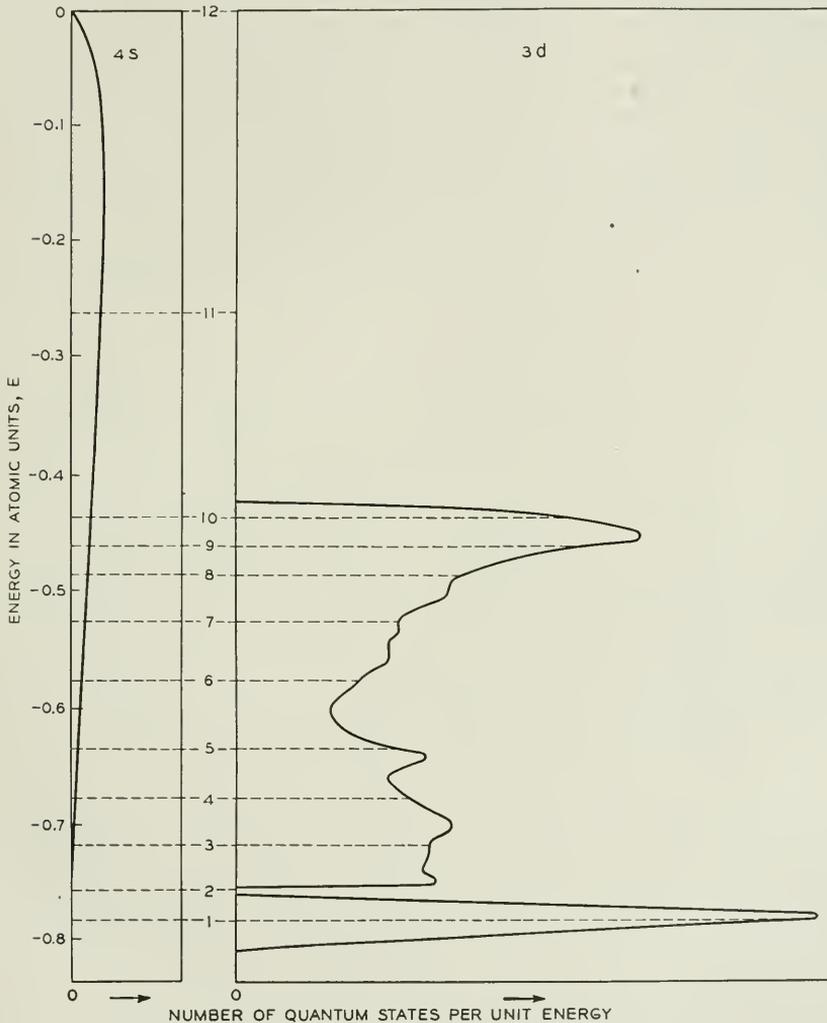


Fig. 27—Distribution of states in energy for copper. The distribution is probably quite similar for iron, cobalt, and nickel and in the absence of calculations for these other metals, this figure will be used for them. The total number of quantum states per atom in the $4s$ and $3d$ bands having energies less than the ordinates of the dashed lines are given by the corresponding integers.

Fig. 16. The $3d$ levels give a band capable of containing ten electrons per atom, five with each spin; and the $4s$ band can hold two electrons per atom, one with each spin. Curves representing $N(E)$ for these bands, calculated for the case of copper by Slater and Krutter, are shown in Fig. 27. We see that the $4s$ band is much wider in energy than the $3d$ and that it contains only one-fifth as many electronic states. The band structure will be similar for all the transition elements; the energy scales, however, will be different. As is shown in Fig. 6 the $3d$ electrons are more tightly bound for copper than for nickel or chromium. Corresponding to this tighter binding, the $3d$ wave functions of copper extend less in space than those of nickel and chromium and consequently they overlap less between atoms and the $3d$ band is narrower for copper. Progressing towards decreasing atomic number in the sequence of elements from copper to scandium, the $3d$ band will continually widen; and this widening, as we shall see later, can help account for the absence of ferromagnetism for the elements before iron in the periodic table.

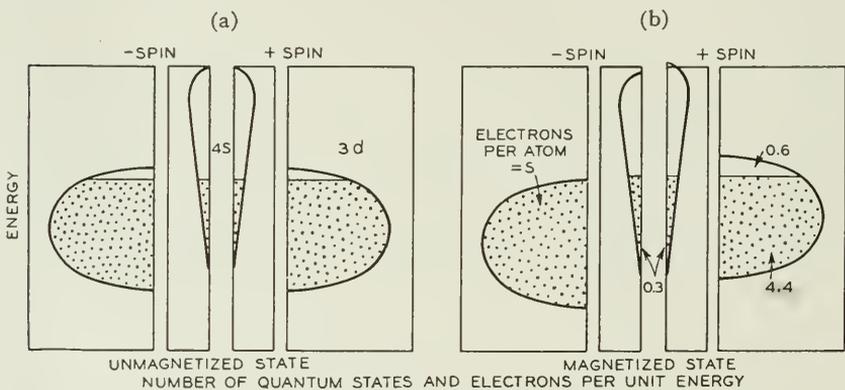


Fig. 28—The ferromagnetism of nickel.

In Fig. 28, we give a simplified representation of the $4s$ and $3d$ bands split into two sets according to the spin. (We may, if we wish, suppose that a magnetic field is applied along which the spin is quantized, but that the field is so weak that the displacement of the energy bands produced by it is negligible; this supposition is not necessary, however, for regarding the spin we shall need only the fact that all the electrons in the + spin band have parallel spins which are anti-parallel to those in the - spin band.) For the element nickel there are 28 electrons, 10 of which are in the $3d$ and $4s$ bands. They can fill the bands as indicated in Fig. 28a. Let us compare this distribution with the electron

configuration of the atom, Fig. 6; we see there that there are unequal numbers of electrons of the two spins. This inequality is produced by the exchange effect which lowers the more occupied set of $3d$ levels in respect to the less occupied set and produces a stable arrangement with the $3d$ levels of one set completely filled. This exchange effect operates in the same way in metallic nickel. In Fig. 28*b* we show the distribution which results when electrons are shifted from the $3d$ band of plus spin to that of minus spin until the latter is filled. The exchange effect produces the displacements of the bands as shown. The arrangement in Fig. 28*b* is stable; in order for electrons to be transferred from the filled minus $3d$ band to the plus band, they would have to increase their energy, a fact which is expressed by drawing the diagram so that the lowest vacant quantum states are appreciably above the highest energy state of the filled $3d$ band. Thus for nickel an unbalanced distribution of spins prevails both for the free atom and the metal.

The Energy of Magnetization

The argument presented above for the stability of the magnetized state shown in Fig. 28*b* is not really rigorous. We saw that if one electron was transferred from the filled $3d$ band to one of the vacant states, its energy and, therefore, the energy of the crystal would be raised. In other words, the magnetized state has less energy than a state which is slightly less magnetized. This fact in itself does not prove that the magnetized state is stable; it proves only that it is metastable—i.e., that its energy is less than the energy of other states which differ from it slightly; in order to establish the stability of the magnetized state, it is necessary to prove that its energy is less than the energy of any other state including that of the unmagnetized state shown in Fig. 28*a*. We may illustrate this necessity by considering the following hypothetical behavior: as the magnetization is reduced from that of Fig. 28*b* to zero (the value for Fig. 28*a*), the energy might at first increase and then decrease—decreasing so much finally that the energy would be lower for the unmagnetized than for the fully magnetized state. We shall, therefore, discuss the difference in energy between the fully magnetized and unmagnetized states; theory shows that this quantity is the fundamental one whose value determines whether or not ferromagnetism occurs.

Let us consider the change in energy in going from the unmagnetized state to the magnetized state in Fig. 28. This change in energy can be separated into two contributions, one positive and one negative. The positive contribution comes from an increase in “Fermi energy” or

"energy of motion," which was discussed in connection with the binding energy of metals. This energy is positive because after the shift to the magnetized state, electrons have moved from states in the band of plus spin to states which lie higher—in respect to the bottom of the bands in both cases—in the band of the minus spin; that is, the electrons which have moved from one band to the other have all gained "energy of motion." The negative contribution to the energy comes from the exchange effect. This causes the lowering of the filled band and the raising of the unfilled band; since there are more electrons in the lowered band than in the raised band, there is a net decrease in energy due to this exchange effect. Thus we have a positive change in Fermi energy and a negative change in exchange energy in going from the unmagnetized to the magnetized state. If the exchange energy has a greater change than the Fermi energy, the energy of the magnetized state is lower and the metal is ferromagnetic.

No satisfactory calculations have as yet been made for these energy differences. In order to calculate them, accurate values for the distribution of states in the $3d$ band are needed, and the mathematical methods available for computing this distribution are not as yet very satisfactory. Next the exchange effect energy must be found; this is also difficult to calculate accurately. Finally, the description given here is over-simplified; in particular another energy term, known as the correlation energy, must be included; this energy acts somewhat like an exchange energy but between the bands of different spins and it tends to cancel out the exchange energy. Although these difficulties greatly mar the usefulness of the theory of ferromagnetism represented in Fig. 28, this theory is able to correlate a large amount of experimental material in a very natural way; and since it is the theory based on the concepts of energy bands, it is the one that we shall discuss in this paper. In passing, however, we must state that there are other theories of ferromagnetism which in some ways are more successful and in other ways less successful than the band theory. Some of these are atomic rather than band theories. An example of this type of difference in method of attack was given in the discussion of the binding energy of sodium chloride; two treatments were given: for one the basis being the ions and for the other the energy bands. In the case of sodium chloride, however, the theoretical equivalence of the two methods is easily demonstrated. In the case of ferromagnetism, the two theories are not equivalent and are both simplifications of a more complex and as yet unsatisfactorily explored intermediate case.

Although no satisfactory calculations of the energy difference between the magnetized and unmagnetized states of metals exist, the

theory must be regarded as representing great progress over non-wave-mechanical theories. The reason is this: in older theories of ferromagnetism the energy was supposed to come from the magnetic interaction between the magnetic dipoles, and it turned out that the energies calculated in this way were at least a thousandfold too small. The energies calculated in the new theory are adequate in magnitude but have nothing to do with the magnetic moment of the electron; they arise from the exchange energy, which is, as we have said before, an electrostatic energy resulting from the wave-mechanical treatment of Pauli's principle. It is the laws governing the spin quantum number of the electron, not the magnetic moment, which are responsible for the energy of magnetization; the externally observed magnetic field of a ferromagnetic material is merely a superficial indication of more fundamental electrostatic forces.

Intrinsic Magnetization

According to our theory, the low energy state and therefore the stable state of metallic nickel is a magnetized one. If one picks up a piece of nickel at random, however, it may not appear to be magnetized. This apparent absence of magnetism is due to the presence of "domains." According to the domain theory—which is a very well established branch of magnetic theory—a block of nickel will consist of a number of microscopic domains, each highly magnetized, but having their magnetic moments pointing at random in a number of directions so that on the average there is no magnetism. The application of a magnetic field aligns the magnetic moments of these domains and, since they are then all parallel, one can measure the total magnetization of the sample. A field strong enough to line up all the domains is said to produce "saturation" because a further increase in field will give no further increase in magnetization. It is customary and convenient to divide the total or saturation magnetic moment of the material by the total number of atoms, thus finding the average magnetic moment per atom, and to express this value in Bohr magnetons. The resultant value is called the intrinsic magnetization per atom and is denoted by β .³⁴ For example, if a crystal had one electron per atom and all the electrons had their spins parallel, then all their magnetic moments would be parallel, too, and the intrinsic magnetization would be unity, $\beta = 1$.

For nickel the intrinsic magnetization is 0.6 Bohr magnetons per atom. The following argument shows how easily such a fractional number can be accounted for by our theory. Nickel has 10 elec-

³⁴ The "intrinsic magnetization" is customarily defined as the magnetic moment per unit volume when the moments of the domains are parallel.

trons per atom in the $3d$ and $4s$ bands. The $3d$ band with minus spin is supposed full, containing five electrons per atom. The $4s$ band (both spins) can contain two electrons per atom, and from Fig. 27 we see that it is about one-fourth full; suppose it contains 0.6 electrons per atom; the remaining electrons go to the $3d$ band with plus spin which is not quite full but has a "hole" in it of 0.6 electrons per atom. There are equal numbers of electrons of each spin in the $4s$ band and their magnetic moments cancel.³⁵ The net magnetic moment arises from the unbalance of 0.6 electrons per atom between the two parts of the $3d$ band. This unbalance will correspond to a magnetization of 0.6 Bohr magnetons. The theory is not capable of predicting the number 0.6 exactly; however, this number is entirely consistent with what can be said about the distribution of levels in the band. In the "atomic" theories of magnetism, it is supposed that each atom has a certain magnetic moment. From the results of the gyromagnetic experiments,³⁶ one concludes that the magnetization is due to electron spin. Since an atom whose magnetism is due to electron spin must have a magnetic moment equal to an integral multiple of the Bohr magneton,³⁷ the "atomic" theory is forced to assume that 40 per cent of the nickel atoms are unmagnetized and that 60 per cent have one Bohr magneton, or else that 70 per cent are unmagnetized and 30 per cent have two Bohr magnetons or at any rate that there are at least two kinds of atoms. These rather awkward assumptions are not required in the band theory, the reason being, as is suggested in Fig. 28, that the electrons are not thought of as belonging to the atoms individually but to the crystal as a whole.

The intrinsic magnetization of ferromagnetic material decreases with increasing temperature. In the band theory this is explained as follows: at a temperature T some of the electrons are excited from the filled band to the partially filled band; as the temperature is increased more are shifted. Furthermore, if we compare the effects of two equal increments of temperature, one occurring at a higher temperature than the other, the one at the higher temperature will have the greater effect. This is because at the higher temperature more electrons have been shifted; hence the exchange effect displacement of the band of one spin in respect to the band of the other spin is less and electrons need

³⁵ Actually there will be a slight exchange effect in the $4s$ band; however, it will be so slight that the magnetic moment produced can be neglected.

³⁶ If a piece of iron is suspended so that it can rotate and then is magnetized, it will acquire an angular momentum. The ratio of angular momentum to magnetic moment should be mc/e if the magnetization arises from electron spin and $2mc/e$ if it arises from motion of the electron as a whole. Experiment gives the following fractions of the former value: for iron 1.03, for cobalt 1.23, for nickel 1.05.

³⁷ For a discussion of this theorem see K. K. Darrow's article, "Spinning Atoms and Spinning Electrons," *Bell Sys. Tech. Jour.*, XVI, 319 (1937).

not gain so much energy to shift from the more full to the less full band. Hence at the higher temperature there is more decrease in magnetization per degree rise in temperature than at the lower temperature. A logical consequence of this reasoning is that the magnetization decreases more and more rapidly as the temperature increases

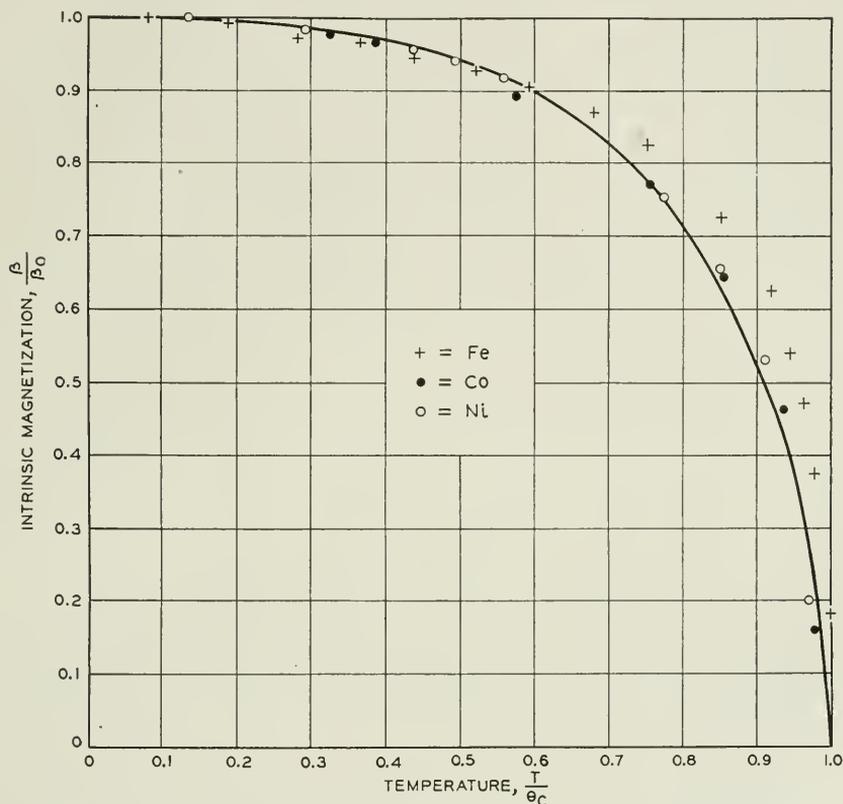


Fig. 29—Intrinsic magnetization versus temperature. The horizontal scale represents the temperature divided by the Curie temperature and the vertical scale, the intrinsic magnetization divided by the intrinsic magnetization at absolute zero. The theoretical curve is derived from quantum mechanics.

and becomes zero at a certain critical temperature, which is known as the Curie temperature and denoted by θ_c . A more complete discussion of the theory of the temperature dependence of magnetism would belong in a paper devoted solely to the theory of magnetism.³⁸ In

³⁸ See, for example, K. K. Darrow, *Bell Sys. Tech. Jour.* XV, 224 (1936), R. M. Bozorth, "The Present Status of Ferromagnetic Theory," *Bell Sys. Tech. Jour.*, XV, 63 (1936) and texts such as J. H. Van Vleck, "The Theory of Electric and Magnetic Susceptibilities," Oxford, 1932, E. C. Stoner "Magnetism and Matter," Methuen and Company, Ltd., London, 1934, and F. Bitter "Introduction to Ferromagnetism," McGraw-Hill Book Co., New York, 1937.

this paper we shall use the fact that the magnetism changes with the temperature to explain the anomalous expansion of ferromagnetic materials. In Fig. 29 we show the variation in intrinsic magnetization with temperature as observed for iron, cobalt and nickel.

Variation of Intrinsic Magnetization with Composition

Let us consider how the intrinsic magnetization should vary from element to element in the transition series, supposing always that the temperature is so low that thermal effects can be neglected. The element next to nickel is cobalt; cobalt has one less electron than nickel so that the $3d$ band and partially filled $4s$ band for it will have one less electron in them. Because of the relatively small number of quantum of states in the $4s$ as compared to the $3d$ band, this deficit will be made up mainly by the $3d$ band which will therefore contain not 4.4 as for nickel but instead 3.4 electrons leading to an unbalance of 1.6 Bohr magnetons per atom. The observed β for cobalt is 1.7 in good agreement with this.

One can obtain electron atom ratios intermediate between cobalt and nickel by forming alloys. We shall speak of the electron concentration, C , of these and other alloys, meaning by this term the total number of electrons available for the $3d$ and $4s$ bands divided by the total number of atoms. So long as the minus spin half of the $3d$ band remains full and so long as the number of electrons in the $4s$ band does not vary much, the value of β will be a linear function of the electron concentration varying from ~ 1.6 to ~ 0.6 as the concentration varies from 9 for cobalt to 10 for nickel. In Fig. 30 are given the intrinsic magnetizations plotted against electron concentration for a series of alloys. It is seen that from cobalt to about halfway between nickel and copper, an increase in C produces, very nearly, a numerically equal decrease in β . This means that the increase in C goes toward filling up the holes in the $3d$ band and reducing the unbalance and hence β . Some alloys are included in Fig. 30 for which the two elements are not adjacent in the periodic table; their values of β also conform to the values predicted from their electron concentrations.

The very natural way in which the band theory accounts for the results shown on Fig. 30 is its principal success in the theory of ferromagnetism.

The bend in the curve between iron and cobalt is not very satisfactorily explained at present. One theory is that for iron neither $3d$ band is entirely full; but this explanation is said to be inconsistent with the observed dependence of magnetization upon temperature at low

temperatures. Another theory is that there are only 0.2 electrons in the 4s band, thus leaving the remaining 7.8 electrons of iron distributed 5 to one 3d band and 2.8 to the other, leaving an unbalance of 2.2; this theory is unsatisfactory because it would require an inexplicable displacement upwards of the 4s band compared to the 3d in going from

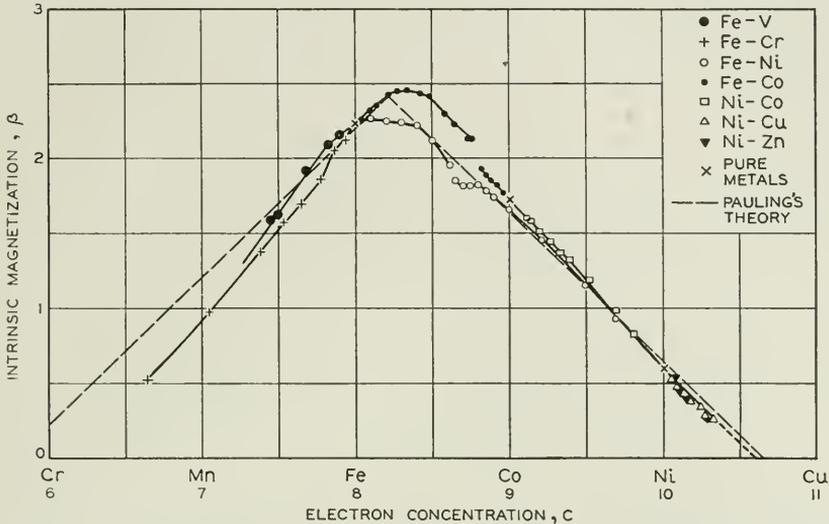


Fig. 30—Intrinsic magnetization versus electron concentration.

The data for this figure were obtained from the following sources:
 Fe-V and Fe-Cr M. Fallot *Ann. de Physique* 6, 305-387 (1936).
 Fe-Co R. Forrer *J. de Physique et le Radium*, 1, 325-339 (1930).
 Fe-Ni M. Peschard *Comptes Rendus* 180, 1836 (1925).
 Ni-Co P. Weiss, R. Forrer, and F. Birch *Comptes Rendus* 189, 789-791 (1929).
 Ni-Cu and Ni-Zn V. Marian *Ann. de Physique* 7, 459-527 (1937).

cobalt to iron. Another theory has been proposed by Pauling³⁹; he has stated it in the "atomic" language but it can be translated into the band language as follows: the 3d band is broken into two parts, an upper part containing 4.88 levels per atom, 2.44 for each spin, and a lower part separated from the upper by an energy gap and containing 5.12 levels per atom, 2.56 for each spin. A number of electrons per atom varying from 0.6 for nickel to 0.7 for cobalt are in the 4s band; for simplicity we shall suppose that this number has a constant value of 0.65 electrons per atom. According to this simplification, one of the upper parts of the 3d band has 0.65 holes for nickel. This band will become empty if the electron concentration is decreased by 1.79 ($= 2.44 - 0.65$)—that is, for a concentration of $10 - 1.79 = 8.21$.

³⁹ L. Pauling, *Phys. Rev.*, 54, 899 (1938).

If the concentration is decreased below 8.21, electrons will be removed from the upper part with the other spin; this will result in a decrease in the unbalance and hence in β , which has for $C = 8.21$, a value of 2.44—corresponding to one filled and one empty upper part; and this decrease will be numerically equal to the decrease in C . Accordingly, the value of β for iron, $C = 8$, is $2.44 - 0.21 = 2.23$. The numbers 2.44 and 2.56 were, of course, chosen so as to obtain this agreement for iron. This theory of Pauling expresses reasonably well the variations in β for all the alloys of Fig. 30.

Criterion for Ferromagnetism

We must now see how the theory explains the absence of ferromagnetism for the remaining transition elements. We have seen that the exchange energy lowers and the Fermi energy raises the energy of the magnetized state compared to the unmagnetized state. These two effects very nearly cancel even for the magnetic elements iron, cobalt, and nickel. For the other elements in the transition series, which are not ferromagnetic, the Fermi term apparently exceeds the exchange term. We shall give a theoretical reason for expecting this result.

In the first place we must indicate how nearly the effects cancel. Let us take cobalt, which has nine electrons in the $3d$ and $4s$ bands, as an example. From Fig. 27 we see that for cobalt in the unmagnetized state both $3d$ bands are filled to about -0.46 atomic units. In the magnetized state one band is filled by electrons which have come from levels with less energy of motion in the other band. Since the top of the $3d$ band comes at about -0.42 units on Fig. 27, the average gain in energy for each transferred electron is about 0.04 units. Since the number of electrons transferred is 1.7 per atom, the increase in Fermi energy is 0.068 atomic units or 0.9 eV per atom. From an analysis of thermal measurements the value for the actual energy of magnetization is found to be about 0.2 eV per atom; a value which is only about one fourth of the predicted increase in the Fermi energy. Hence the exchange energy exceeds the Fermi energy by only 25 per cent and the two energies nearly cancel.

The variation in the structure of the $3d$ band from element to element was discussed in connection with Fig. 27; we concluded then that the bands become wider as we recede in the periodic table from nickel towards scandium. Greater band width means greater Fermi energy in the magnetized state and this effect opposes the occurrence of ferromagnetism. The exchange energy can also change. Calculations by Slater,⁴⁰ which unfortunately are too over-simplified to bear much

⁴⁰ J. C. Slater, *Phys. Rev.*, 49, 537, 931 (1936).

weight, show that for manganese the Fermi energy outweighs the exchange energy so that manganese is not ferromagnetic at all. In Fig. 31 we represent the state of affairs predicted for chromium; the exchange

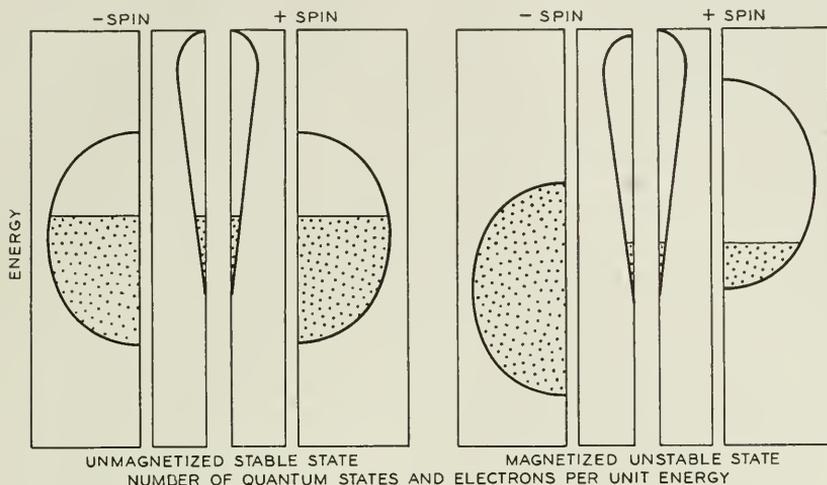


Fig. 31—The absence of ferromagnetism for chromium.

energy is over-balanced by the Fermi energy and for this metal the unmagnetized arrangement has the least energy and is the stable state.

A very instructive curve can be drawn to illustrate the criterion for the occurrence of ferromagnetism. It is shown in Fig. 32. The vertical scale is the energy of the unmagnetized state, E_U , minus the energy of the magnetized state, E_M . When $E_U - E_M$ is positive, the magnetized state has the lower energy and will be the stable state,

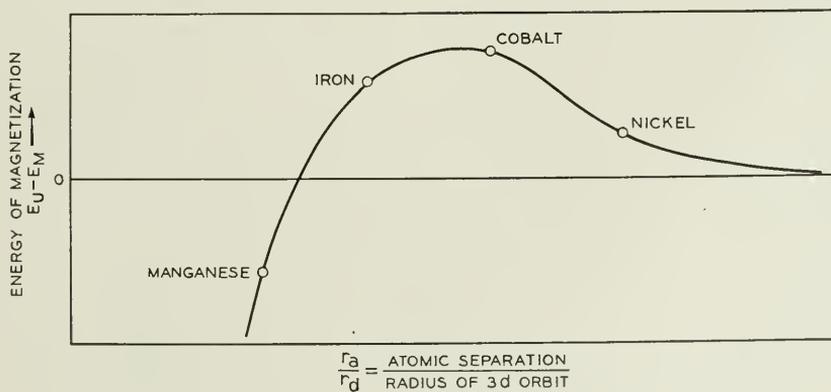


Fig. 32—Criterion for the occurrence of ferromagnetism.

and when $E_U - E_M$ is negative, the reverse is true. Hence a positive value for $E_U - E_M$ is a necessary and sufficient condition for ferromagnetism. The variable on the horizontal scale is r_a (the distance between nearest neighboring atoms in the crystal) divided by r_d (the average radius for the $3d$ wave function). Small values of r_a/r_d mean crowding together of the atoms, large values of the Fermi energy, and no ferromagnetism. Certain values of r_a/r_d , such as are found for iron, cobalt, and nickel, favor ferromagnetism. Very large values of r_a/r_d mean widely separated atoms and low Fermi energy and, consequently, ferromagnetism; however, for very widely separated atoms, the energy of interaction between them is small and so is the energy of magnetization. The curve shown in Fig. 32 is only qualitative. The theory that the curve should have this form was first worked out by Bethe using the "atomic" rather than the band theory of magnetism; for the reasons discussed above, however, no quantitative theoretical curve is available. Ratios of r_a/r_d have been calculated by Slater* and occur as indicated for several elements. This curve can be considered from either of two viewpoints. We may imagine that r_a remains constant, as it does approximately for the transition elements, and that r_d varies from element to element; we then get the result shown in Fig. 32. On the other hand we may consider a definite chemical element thus fixing r_d ; then Fig. 32 tells us how the energy of magnetization depends on the lattice constant or volume of the sample. We shall use this in the following paragraphs to explain the effects of magnetism upon thermal expansion.

Magnetism and Thermal Expansion

In Fig. 33*a* we show a solid curve which represents for iron in the magnetized state the dependence of the energy E_M upon the lattice constant a . In Fig. 33*b* is shown, on a relatively enlarged energy scale, the value of $E_U - E_M$ as taken from Fig. 32 with r_d thought of as fixed, and a the lattice constant in place of r_a . The position of curve (*b*) has been adjusted so that the point marked \circ , corresponding to iron in Fig. 32, comes at the equilibrium distance or minimum of the E_M curve. Adding the solid curves of (*a*) and (*b*) (adjusting the energy scales, of course) gives the dashed curve representing the energy E_U of the unmagnetized state shown in Fig. 33*a*. We are now in a position to make predictions about the thermal expansion of iron.

Let us imagine that the iron is somehow made to stay in the magnetized state. Then its expansion curve, lattice constant versus temperature, will be shown as in Fig. 33*c* by the solid heavy line. Next

* J. C. Slater, *Phys. Rev.*, 36, 57 (1930).

imagine it maintained in the unmagnetized state; in this state the equilibrium lattice constant is smaller than for the magnetized case and the expansion curve is shown dashed. The curves for fixed

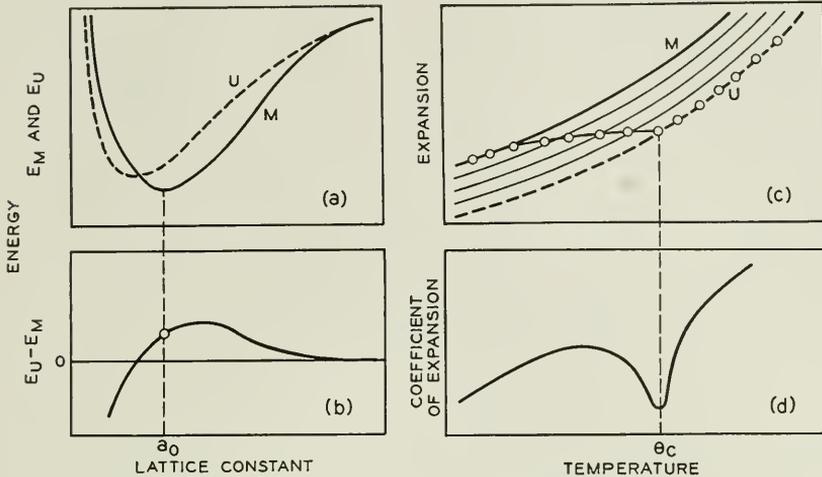


Fig. 33—Theory of the thermal expansion of iron.

- Energy in magnetized (M) and unmagnetized (U) states versus lattice constant.
- Difference in energies versus lattice constant.
- Lattice constant versus temperature.
- Thermal expansion coefficient versus temperature.

intermediate degrees of magnetization are shown as light lines. Now as the iron is heated the magnetization does not stay constant but decreases with temperature and becomes zero at the Curie temperature θ_c . In Fig. 33c this corresponds to a continuous shifting from the line of higher magnetization to the lines of lesser magnetization with increasing temperature as indicated by the curve with circles. We see that the rate of expansion—that is, the thermal expansion coefficient, which is defined as the derivative of the curve divided by a —should have an irregular form as shown in Fig. 33d.

In Fig. 34 we show observed thermal expansion curves for a series of iron nickel alloys,⁴¹ showing that the expansion for iron rich alloys agrees with that predicted from Fig. 33. The reader may verify that had the curve of Fig. 33b been adjusted to correspond to nickel, the anomalous expansion would have been in the opposite direction, as is found experimentally for the nickel rich alloys.

The more rapid the transition from the magnetized to the unmag-

⁴¹ Figures 34 and 35 are taken in a modified form from J. S. Marsh, "Alloys of Iron and Nickel," Vol. I, Special-Purpose Alloys, 1938, McGraw-Hill Book Co.

netized curve, the greater will be the anomaly in expansion. For the alloy invar the Curie point occurs at about 200° C. and the transition is so rapid that the magnetic effect nearly cancels the normal expansion.

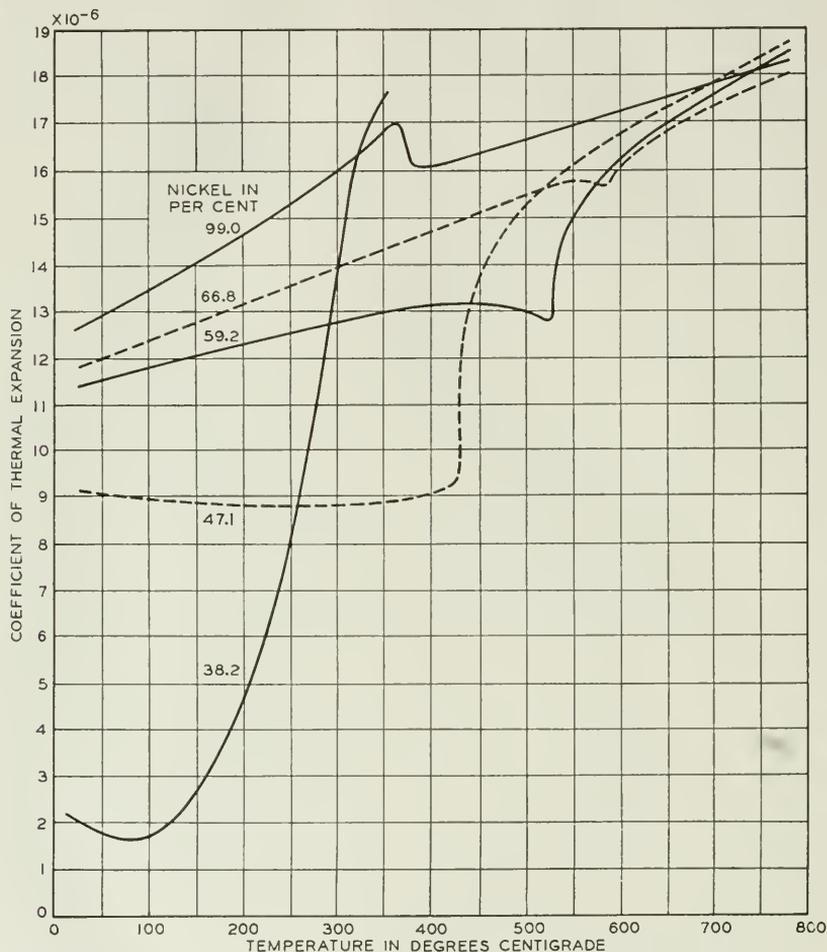


Fig. 34—Coefficients of expansion for iron-nickel alloys versus temperature.

Figure 35 shows a curve for the thermal expansion of an iron-nickel alloy containing 36.5 per cent Ni, corresponding to Fig. 33c. The flat region implies an expansion coefficient of nearly zero.

Grüneisen's law is definitely violated by metals having expansion effects of the sort associated with ferromagnetic changes. Grüneisen's law, it will be recalled, states that the thermal expansion coefficient is proportional to the specific heat. For all ferromagnetic transforma-

tions, the specific heat has a peak at the Curie temperature. For invar, however, the thermal expansion suffers a dip at the Curie temperature. Hence the proportionality between specific heat and thermal expansion coefficient does not hold. Even for cases where the expansion coeffi-

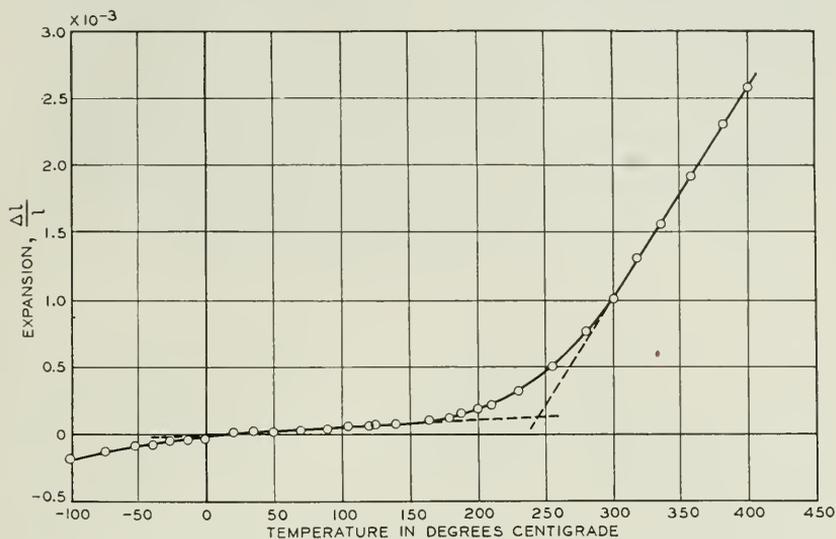


Fig. 35—Expansion of invar versus temperature.

cient has a peak, as in nickel for example, the proportionality does not hold. The reason for the failure of Grüneisen's law is easily found and reflects in no way upon validity of the law for the cases to which it is intended to apply. Grüneisen's law is derived by assuming that the crystal has a single definite energy versus volume curve. For ferromagnetic materials this is not true as is evinced by the two curves of Fig. 33*a*.

In this paper we have been concerned with the important but inactive attributes of electrons associated with their energies. We have seen how the variations of the electronic energy levels can be used to explain a number of the important properties of solids. In the next paper, we shall discuss the more dynamic subjects of electron velocities and accelerations.

ACKNOWLEDGMENTS

The writer would like to express his gratitude to Dr. R. M. Bozorth for discussions of the section on magnetism, to Dr. K. K. Darrow for criticisms and suggestions, to Mr. A. N. Holden for many valuable comments on the manuscript and for the preparation of the crystal of Fig. 1, and to Mr. B. A. Clarke for much valuable advice and assistance with the figures.

Dial Clutch of the Spring Type*

By C. F. WIEBUSCH

The mathematical theory is developed for the spring clutch which consists of two coaxial cylinders placed end to end and coupled torsionally by a coil spring fitted over them. Relations are derived whereby it is possible to design spring clutches in terms of the requirements and the constants of the spring material. Experimental verification of the relations is given. The theory of residual and active stresses as applied to the springs is discussed.

THE operation of all present day machine switching telephone systems depends on the use of the telephone dial. The dial originates the current pulses required to operate the step-by-step, panel, or crossbar switching equipment and for the reliable functioning of this equipment the pulses must occur within a closely limited frequency range. The stepping pulses are produced during the unwinding of the dial from the position to which it has been wound by the subscriber and it is this unwinding which must occur at a constant speed. To accomplish the speed control a governor depending on centrifugal force is used. It is not desirable that the governor come into action on the windup of the dial as this would put an extra load on the user's finger and slow up the operation of dialing. A clutch which holds in one direction of rotation and is free in the other direction is therefore interposed between the governor and the finger wheel with its associated circuit interrupting mechanism. In the past, the most commonly used clutch consisted of a pawl and ratchet, but this has now been replaced by the spring clutch because of its quietness and lower cost. A partially assembled dial using a spring clutch is shown in Fig. 1.

The ideal clutch for a dial governor would be one offering zero coupling torque during dial windup and an infinite positive coupling in the other direction. In practice the free torque in the windup direction need only be small compared to the torque of the main spring, while the holding torque in the other direction need only be great enough to withstand the main spring torque plus any helping torque that a

* Essentially the same material was presented at National Meeting of Applied Mechanics Division of The American Society of Mechanical Engineers, New York, N. Y., June 14-15, 1939, and published in *Journal of Applied Mechanics*, September 1939, under the title of "The Spring Clutch."

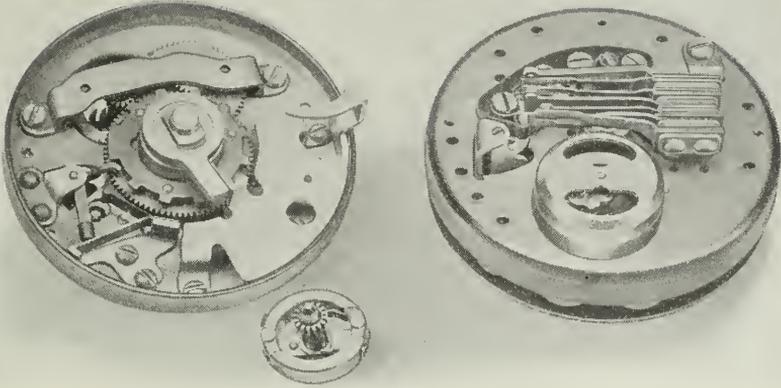


Fig. 1—Dial governor and front and back view of partially assembled dial.

subscriber may impatiently exert. The first limit is easy to set up on the basis of the mechanical constants of the dial; the fixing of the latter limit required measurements on the strength of a considerable number of persons. It was found that the maximum force that an ordinary man can exert with any finger at the finger hole of a dial is about six pounds. When the proper factors of safety have been added to these limits it is possible to specify exactly the requirements on the dial clutch. The remainder of this paper is devoted to the development of the relations and a discussion of the problems involved in designing a spring clutch to meet a given set of requirements.



Fig. 2—Telephone-dial clutch.

The type of spring clutch to be discussed here consists of two cylinders placed end to end, rotating on a common axis, and torsionally coupled by the friction between the cylinders and a coil spring fitted over the cylinders. A photograph of such a clutch, assembled and apart, from a telephone dial is shown in Fig. 2. In spite of widespread use there seems to be little theoretical discussion of this device in the literature.

It is obvious that if the driving drum be rotated in the direction to wind up the spring and decrease the diameter, the spring will grip the cylinders and will be capable of exerting more torque than it would in the direction of rotation which tends to unwind the spring. Equations are to be developed which will permit the calculation of these two torque values in terms of the physical dimensions and the material constants of the clutch.

TORQUE OF SPRING CLUTCH IN THE FREE DIRECTION

In Fig. 3 assume that the spring is fastened to the left-hand arbor in order that any slipping which may take place must occur on the driving drum on the right. Assume also that the spring is so formed that the

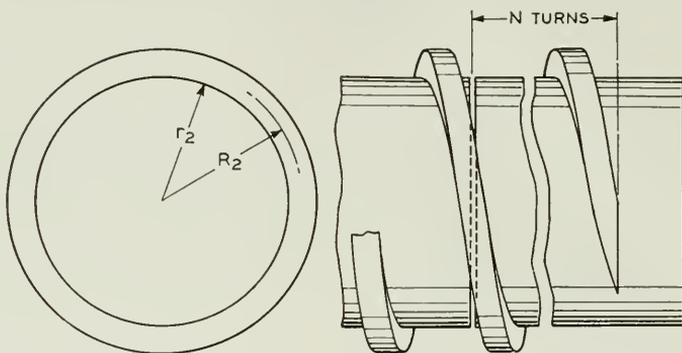


Fig. 3—Diagram of spring clutch.

inward radial force on the drum per unit length of the material is constant when no torque is applied.

NOTATION

- l = length along the line of contact of the spring on the arbor, measured from the free end to any point, in.
- μ = coefficient of friction between the spring and the arbor
- r_2 = radius of the arbor, in.
- R_2 = radius to the neutral bending axis of the spring when on the arbor, in.
- N = number of turns on the right-hand arbor
- P = compression in the spring wire at any point due to the applied torque; this is not the stress in the material but the resultant force acting across the entire cross section of the wire, lb.
- f_0 = radial force of spring on arbor when no torque is applied, lb. per in. of contact line.

As compression exists in the spring wire at any point when the arbor is turned to make the spring unwind, there will be a radial force subtracting from f_0 at every point. This subtracting force is P/r_2 . The increase of compression in the wire along the length of the line of contact due to friction is

$$dP = \mu(f_0 - P/r_2)dl, \quad (1)$$

which upon integration gives

$$l = - (r_2/\mu) \ln [f_0 - (P/r_2)]C,$$

where C is a constant of integration equal to $1/f_0$ since $P = 0$ at $l = 0$. Hence

$$P = r_2 f_0 (1 - e^{-\mu l/r_2}). \quad (2)$$

Since $l = 2\pi r_2 N$,

$$P = r_2 f_0 (1 - e^{-2\pi N\mu}). \quad (3)$$

Since the torque is equal to Pr_2

$$T = r_2^2 f_0 (1 - e^{-2\pi N\mu}) \text{ (in.-lb.)}. \quad (4)$$

It will be observed that for any but fractional values of $N\mu$ the exponential term becomes very small and

$$T = r_2^2 f_0 \quad (N\mu > 1). \quad (5)$$

If $N\mu = 1.0$ this expression is in error by only 0.2 per cent. It can thus be seen that provided $N\mu$ does not become too small, variations in N or μ do not affect the torque exerted. The torque will depend only on the radius of the arbor and on the force f_0 which is controlled entirely by the dimensions and the elastic properties of the spring.

TORQUE OF SPRING CLUTCH IN THE GRIPPING DIRECTION

If the torque is applied to the clutch in the direction to wind up the spring, instead of unwind it as in the previous case, the force P'/r_2 due to the tension P' in the spring wire adds to the inward force f_0 and the relation corresponding to equation (1) is

$$dP' = \mu(f_0 + P'/r_2)dl. \quad (6)$$

From which by the same method as before

$$P' = r_2 f_0 (e^{2\pi N\mu} - 1). \quad (7)$$

The corresponding torque is

$$T' = r_2^2 f_0 (e^{2\pi N\mu} - 1) \text{ (in.-lb.)}. \quad (8)$$

In this case the torque increases rapidly with an increase in either N or μ especially for large values of $N\mu$. The coefficient of friction is in general a rather variable factor. It is to be expected that the slipping torque will also be variable, but since a lower limit can in general be set for this coefficient it will always be possible to make the number of turns of the spring such as to give any desired lower limit of torque.

THE RADIAL FORCE ON THE ARBOR

One method of evaluating the force f_0 occurring in the torque relations depends on equating the potential energy of strain per unit length of the wire when on the arbor to the work done in expanding the spring from its free diameter to the diameter of the arbor.

Let

E = Young's modulus for the spring material, psi

I = the area moment of the wire section, in.⁴

h = the radial thickness of the wire, in.

R_1 = free radius to the neutral axis, in.

r_1 = free inner radius of the spring, in.

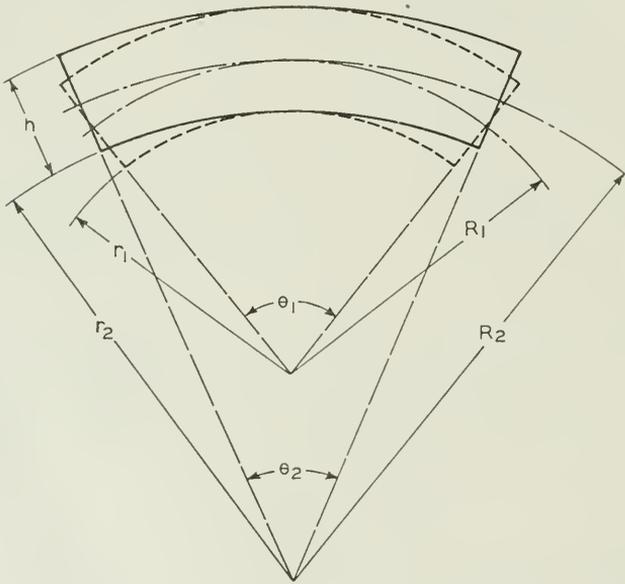


Fig. 4—Element of spring in initial and expanded condition.

Consider the portion of spring wire shown in Fig. 4 straightened out from the initial radius of curvature R_1 to the radius R_2 . The

fibers above the neutral axis will be compressed while those below the neutral axis will be stretched. Let y be the distance of any given fiber from the neutral axis. The length of the undistorted fiber will be $L = (R_1 + y)\theta_1$ while after bending this same fiber will have the length $L' = (R_2 + y)\theta_2$. The strain or the change in length per unit length is

$$\frac{L - L'}{L} = 1 - \frac{(R_2 + y)\theta_2}{(R_1 + y)\theta_1}. \quad (9)$$

Since along the neutral axis there is no change in length, $\theta_2 = L_0/R_2$ and $\theta_1 = L_0/R_1$. Substituting these values in equation (9) gives

$$\text{Strain} = (y/R_2)(R_2 - R_1)/(R_1 + y). \quad (10)$$

The potential energy per unit volume in a material strained in tension or compression is

$$W/V = (E/2)(\text{Strain})^2. \quad (11)$$

Substituting the value of strain from equation (10) in equation (11) the energy density at any point of the deflected wire will be

$$\frac{W}{V} = \frac{E}{2} \left[\frac{y(R_2 - R_1)}{R_2(R_1 + y)} \right]^2. \quad (12)$$

Let b represent the width of the wire at the point y . Then for a wire symmetrical about the neutral axis the strain energy per unit length of the wire becomes

$$\frac{W}{l} = \int_{-h/2}^{h/2} \frac{E}{2} \left[\frac{y(R_2 - R_1)}{R_2(R_1 + y)} \right]^2 \frac{R_1 + y}{R_1} b dy, \quad (13)$$

where the factor $(R_1 + y)/R_1$ represents the ratio of the length of the fiber at the point y to the length along the neutral axis. If all values of y , and hence also $h/2$, are small compared to R_1 there will be little error made in neglecting the y , which carries both positive and negative values, in the expression $R_1 + y$. Hence

$$\frac{W}{l} = \int_{-h/2}^{h/2} \frac{E}{2} \left(\frac{R_2 - R_1}{R_2 R_1} \right)^2 b y^2 dy. \quad (14)$$

The integral of $b y^2 dy$ is the area moment I of the section and therefore

$$\frac{W}{l} = \frac{1}{2} EI \left(\frac{R_2 - R_1}{R_2 R_1} \right)^2. \quad (15)$$

This must be equal to the work done per unit length of the neutral

axis, by the force per unit length $F(\Delta R)$ working through the distance ΔR where $\Delta R = R_2 - R_1$. That is

$$\int_0^{\Delta R} F(\Delta R) d\Delta R = \frac{1}{2} EI \left[\frac{\Delta R}{R_1(R_1 + \Delta R)} \right]^2. \quad (16)$$

Differentiating both sides with respect to ΔR gives $F(\Delta R)$ for the left-hand side, and after simplification

$$F(\Delta R) = \frac{EI}{R_1} \frac{\Delta R}{(R_1 + \Delta R)^3}. \quad (17)$$

Substituting for ΔR its value $R_2 - R_1$ gives

$$F(\Delta R) = EI(R_2 - R_1)/R_1 R_2^3. \quad (18)$$

The equivalent force per unit length measured along the surface of the arbor must be larger than this value in the ratio of R_2/r_2 since the same total force is here distributed over a shorter length. This latter force is f_0 ; hence

$$f_0 = \frac{R_2}{r_2} F(\Delta R) = EI \frac{R_2 - R_1}{R_1 r_2 R_2^2} \text{ lb. per in.} \quad (19)$$

The value of f_0 calculated from this equation in terms of the constants of the spring material and the dimensions of the spring may be used in equations (4) and (8) to calculate the free torque and the slipping torque of the clutch.

EXPERIMENTAL CHECK OF THE FREE-TORQUE RELATION

In order to check the validity of the relation for the free torque, equation (4), and that for the radial force on the arbor, equation (19), the free torque of a given spring on arbors of various diameters as well as the torque for different numbers of turns on the same arbor was measured. The spring of 0.0085×0.022 -in. phosphor-bronze ribbon was attached to a short vertical shaft suspended by a torsion fiber of measured torsional stiffness. The free end of the spring was placed over a vertical arbor capable of rotation. The arbor was rotated and the angle of twist of the torsion fiber was measured thus giving a measure of the slipping torque. The precision of the measurements of torque was about 0.5 per cent although the sensitivity to small changes was about 0.2 per cent. No measurable increase of torque occurred by increasing the number of turns on the arbor beyond six. This is to be expected if the coefficient of friction exceeds about 0.12.

For all succeeding measurements seven to eight turns were used. As a further check on the independence of the torque and the coefficient

of friction for this number of turns a measurement was made before and after oiling the arbor and spring with a light lubricating oil. There was a decrease in torque of approximately 0.25 per cent.

The inside diameter of the spring as measured by a taper gage was 0.180 in. \pm 0.001 in. The torque for arbors ranging in size from 0.182 to 0.193 in. was determined and is shown in Fig. 5. This curve ex-

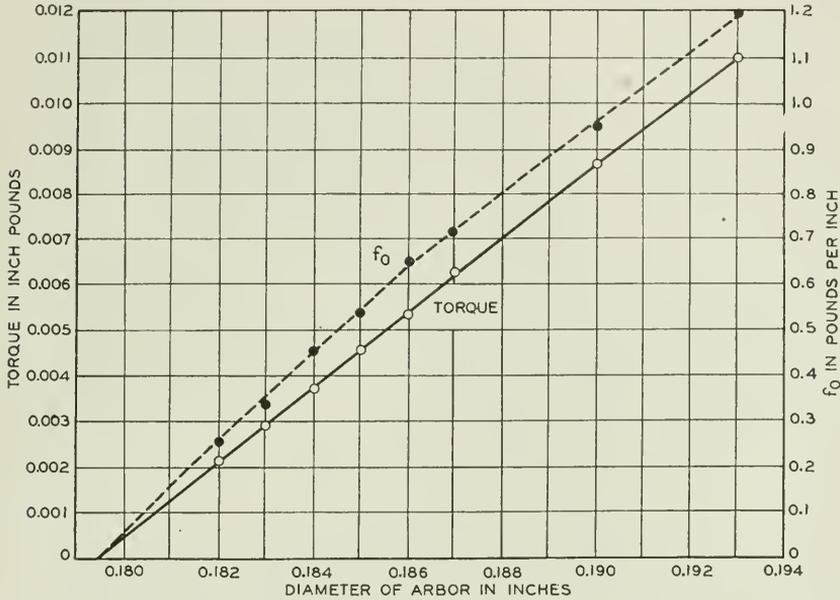


Fig. 5—The free torque and the radial force on the arbor for a phosphor-bronze spring on arbors of different diameters.

trapolated to zero torque gives, for an accurate measure of the inside diameter of the spring, 0.1794 in. Using this value and the quantity EI , determined by obtaining the resonant frequency of a straight short length of the ribbon, of which the spring was made, vibrating as a fixed free reed, the radial force on the arbor as calculated by equation (19) is shown by the dotted curve as a function of the arbor diameter $2r_2$. The points indicate values of f_0 obtained from the measured values of torque by the use of equation (5). The two sets of values agree within about 2 per cent.

As a further check on the validity of the calculations under practical conditions, the free torque of a phosphor-bronze spring on a dial-governor arbor was measured for various numbers of turns of the spring engaging on the slipping arbor. The torque due to bearing friction alone with no spring in place was also measured and found to

be approximately 0.001 in.-lb. This constant value was subtracted from the other measured values of torque. The resulting values of free spring torque are plotted as a function of the number of turns in Fig. 6. The torque values calculated by equations (4) and (19) and

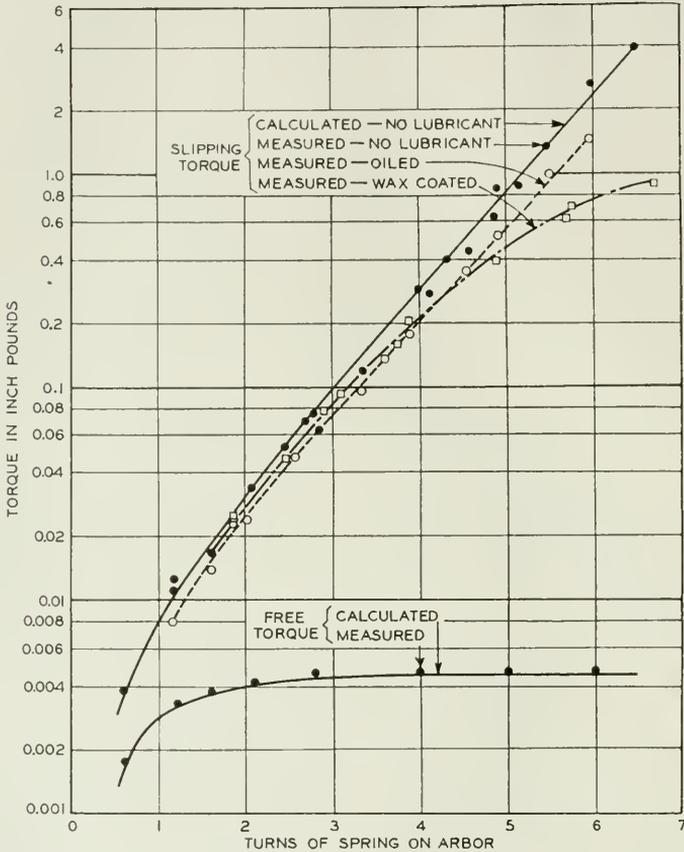


Fig. 6—Dependence of torque on the number of turns of the spring engaging the arbor.

using the value of $\mu = 0.165$ obtained as described in the next section, are shown on the curve to indicate the agreement.

EXPERIMENTAL CHECK OF THE HOLDING-TORQUE RELATION

It is to be expected that any relation for which the coefficient of friction is a controlling factor will be difficult to check accurately. Equation (8) for the slipping torque is of this type. It is possible however by taking a large number of measurements to establish the

validity of the equation and then by determining limiting values for the coefficient of friction, to use this equation as a design relation, especially if only a minimum torque limit is set. Measurements of the slipping torque, as a function of the number of turns, were made on the same dial clutch that was used for checking the free torque.

This slipping torque was not steady as was the case in the free direction but varied as much as ± 20 per cent. An average value was taken in each case. An uncertainty also existed regarding the number of turns engaging the rotating arbor. Since the crossover from one arbor to the other requires practically one whole turn, slight differences in the arbor diameters may result in a gain or loss of almost half a turn. In addition to these factors there was an end effect due to the fact that the free end of the spring wire was cut off square rather than beveled but this factor although calculable was neglected in view of the other uncertainties. The results of these measurements are shown in Fig. 6.

From equation (8) it can be seen that for large values of N the plot of T' versus N will be a straight line provided μ is independent of the force between the spring and the arbor. The slope of this straight line when multiplied by the proper constant, which can be shown to be 0.368, gives the coefficient of friction. For the experimental points shown in Fig. 6 this value of μ is 0.165. Using this value of μ in equation (6) the calculated curve was plotted. Considering the uncertainties involved the calculated and measured curves are in good agreement.

The dotted curve of Fig. 6 shows the effect of lubricating the clutch with a light machine oil. This resulted in only a small decrease of the coefficient of friction. The curve shown by the dashes illustrates the effect of lubricating a clutch with spermaceti. The coefficient was no longer a constant but decreased with an increase in load.

SPRING STRESSES

In determining the load that a spring clutch will withstand, first without stretching which will result in backlash, and second without breaking, initial as well as load stresses must be considered. The initial stresses are made up of the residual stresses due to forming the spring, plus the stresses due to expanding the spring to fit the arbor, that is from an inner radius r_1 to an inner radius r_2 .

Of these limiting load values the easiest to calculate is the torque required to break the spring. This is given by the product of the radius of the spring and the breaking strength of the spring wire. Loads much smaller than this value would stretch some of the fibers of the spring, especially those in which a high initial stress already

existed. As will be shown, such stretching will cause the radius to increase at those portions of the spring where the applied stresses are highest, that is, for those turns near the dividing line of the arbor.

Substituting for y , in equation (10), the distances from the neutral axis to the extreme inner and outer fibers will give the strain in these fibers. Provided R_1 is considerably larger than $h/2$, half the thickness of the material, the distance to these extreme fibers becomes $h/2$ and the y in the denominator can be neglected in comparison to R_1 . The maximum fiber stress due to placing the spring on the arbor is this value of strain multiplied by Young's modulus. Then

$$S_0 = \frac{h}{2R_2} \frac{R_2 - R_1}{R_1} E. \quad (20)$$

This stress is in the form of compression for the outer fibers and tension for the inner. Since a load on the clutch results in a tension in the spring, the stress given by equation (20) must be added to the load tension stress to get the total stress on the inner fibers of the spring. This initial stress therefore reduces the load-carrying capacity of the clutch.

RESIDUAL SPRING STRESSES

When a straight wire is wound upon an arbor to form a coil spring the strain on the inner and outer fibers must exceed that corresponding to the yield-point and plastic-flow results. To simplify the discussion assume the idealized stress-strain curve shown by the heavy lines of Fig. 7(a). The stress distribution across any section of the wire while wound on the winding arbor will be as shown by the heavy lines of Fig. 7(b) where S_{YP} is the yield-point stress, the maximum stress that the material will sustain. The moment, across the section, required to produce this bending is

$$M = \int_{-h/2}^{h/2} Sbydy \quad (21)$$

where b is the width of the wire at any point of y distance from the neutral axis and S is the stress at the same point. If the spring is released, it expands to a radius R_1 in which condition the external and the internal moments are both zero. It is now possible by applying the same moment as was given by equation (21) to reduce the radius of curvature again to R_0 but without causing additional plastic flow. The added stress distribution produced by this second bending must therefore follow a straight line as shown by S_2 - S_2 , which together with the residual stresses in the relaxed condition (radius R_1) must

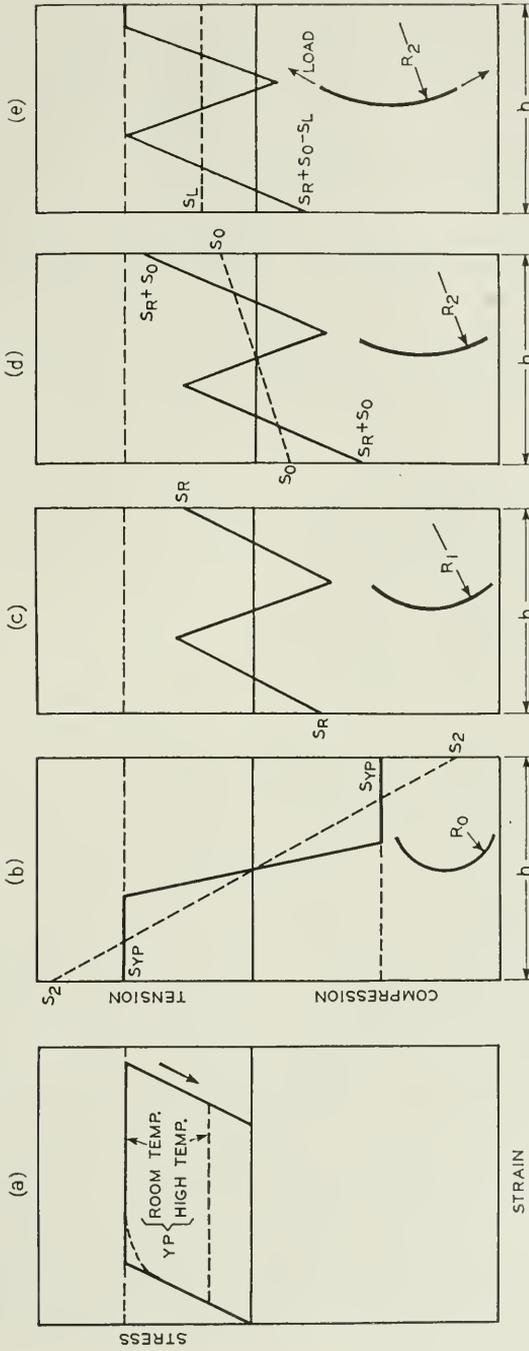


Fig. 7—Idealized stress conditions in clutch springs.

equal the distribution $S_{YP}-S_{YP}$ resulting from the original forming operation. Therefore the stress distribution in the relaxed condition (radius R_1) must be the difference between $S_{YP}-S_{YP}$ and S_2-S_2 or as indicated in Fig. 7(c). The value of S_2 is of course so determined that the moment as specified by equation (21) is the same for the dotted-line as for the solid-line stress distribution.

The value of $S_R = S_2 - S_{YP}$ is relatively easy to determine for rectangular and round wire on the basis of the straight-line stress-strain characteristic if the bending has been sufficiently severe to have caused plastic flow almost to the neutral axis. The moment given by the actual stress distribution will then differ but little from that obtained by equation (21) with S replaced by S_{YP} , a constant. The values of S corresponding to the S_2-S_2 distribution are given by

$$S = 2S_2y/h. \quad (22)$$

For a rectangular wire b is a constant and equating the moments corresponding to the two stress distributions gives

$$\int_{-h/2}^{h/2} S_{YP}bydy = \int_{-h/2}^{h/2} 2\frac{S_2}{h}by^2dy,$$

from which $S_2 = (3/2)S_{YP}$ or

$$S_R = S_2 - S_{YP} = (1/2)S_{YP} \text{ (rectangular wire)}. \quad (23)$$

Similar integrations in the case of a round wire of radius $h/2$ for which $b = 2\sqrt{[(h/2)^2 - y^2]}$ gives

$$S_R = 0.7S_{YP} \text{ (round wire)}. \quad (24)$$

These equations give the residual fiber stresses in the extreme inner and outer fibers under the assumed conditions. In any case where the stress-strain characteristic is known a correct value for the residual stress can be obtained by graphical integration. The values given by equations (23) and (24) will, however, be fair approximations even in cases where the stress-strain characteristic is curved provided it does not exhibit strain hardening to a decided extent. Since a limited amount of plastic flow takes place for any stress above the proportional limit, which is generally far below the yield point, the residual stress may be sufficient to cause a small amount of creep. Any additional stresses will then cause permanent deformation of the spring.

The analysis will be continued on the basis of the idealized straight-line characteristic. Figure 7(d) shows the result of placing the spring

on the clutch arbor. The line S_0 - S_0 shows the stresses added by this expansion where S_0 is given by equation (20). The sum of these stresses and those shown in Fig. 7(c) gives the total stresses as shown by the solid line of Fig. 7(d). If now a load be put on the clutch a uniform stress S_L will be added but this stress for even relatively light loads may be sufficient to cause the total stress on the inner fibers to exceed the yield point as is indicated in Fig. 7(e). The inner fibers are consequently stretched and when the load is released and the spring taken from the arbor it will be found that the center turns of the spring have expanded. Even with the spring on the arbor if the clutch is turned in the free direction it will be noticed that these center turns raise off the arbor. It was shown in the paragraphs on the clutch torque in the free direction that the torque did not increase appreciably after the first few turns. This can be explained by the fact that as soon as the outward radial force due to the compression along the wire is equal to the initial inward radial force of the spring on the arbor the friction on these turns vanishes. The value of the compression will be fixed by a relatively few end turns. Hence if the inward force of some of the center turns decreases due to their stretching this compression will be sufficient to expand the turns to clear the arbor. If S_L is still further increased the stretch will be sufficient to cause the diameter of the center turns to exceed the arbor diameter even when no torque is applied in the free direction.

Since the yield point of metals decreases at higher temperature it is possible to produce a spring having lower residual stresses by the proper heat-treatment. If the wire is wound on an arbor and then heated, additional plastic flow takes place since the maximum stress that can be sustained at the high temperature is that shown in Fig. 7(a) as the high-temperature yield point. If the spring is then cooled and released the expansion will not be as great as for the untreated spring. The residual stresses will again be given by equation (23) or (24) where S_{YP} is taken as the lowest yield point reached in the temperature cycle. In Fig. 8, (a), (b), and (c) show the stresses in the heat-treated specimen corresponding to those shown in Fig. 7, (c), (d), and (e), for the untreated spring. Figure 8(c) shows that for the same load stress as for Fig. 7(e) no permanent deformation has taken place. It is of course important that the strain-relieving temperature should not go high enough to lower permanently the strength of the material. This limit¹ for phosphor bronze is about 320° C.

To determine the stress-temperature characteristics of 18-8 stainless

¹ "Better Instrument Springs," Robert W. Carson, *Trans. A.I.E.E.*, vol. 52, September, 1933, p. 869.

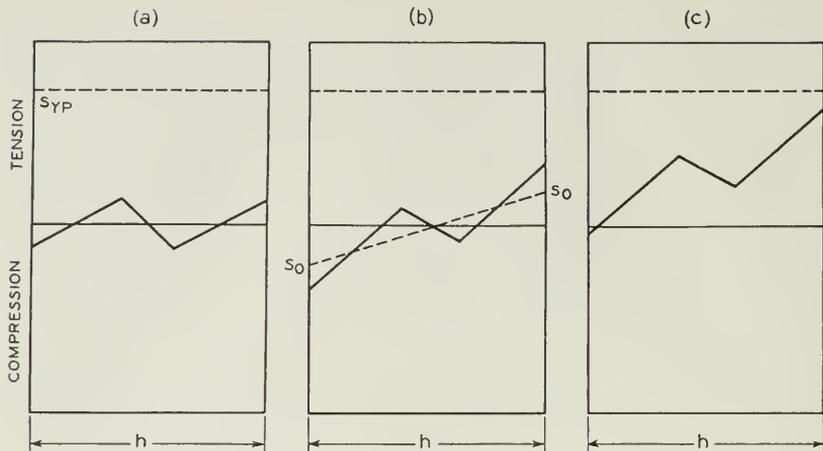


Fig. 8—Stress conditions in a strain-annealed clutch spring.

steel² a number of springs of 0.0068×0.021 -in. ribbon were wound on 0.1486-in. arbors and given various heat-treatments. They were then cooled, released from the arbors, and measured for inside diameter. Table 1 gives the results of these measurements.

TABLE 1
EFFECT OF HEAT-TREATMENT ON SPRINGS
Ribbons of 18-8 stainless steel, 0.0068×0.021 in., heat-treated on
0.1486-in. winding mandrels

Heat-treatment temp., °C.	Time, hr.	Diam. after release, in.	Residual stress, psi
25	..	0.228	111000
100	4	0.206	88000
200	4	0.188	66000
300	4	0.182	58000
400	4	0.175	47000
470	4	0.169	37000
500	4	0.166	33000
400	$\frac{1}{4}$	0.177	51000
400	$\frac{1}{2}$	0.176	49000
400	1	0.176	49000
400	2	0.175	47000
400	3	0.175	47000
400	4	0.175	47000

The residual stresses were calculated by noting from equation (23) that $S_R = S_2/3$ and then obtaining S_2 from equation (20) rewritten as

$$S_2 = \frac{h}{2R_1} \frac{R_1 - R_0}{R_0} E. \quad (25)$$

² 8 per cent nickel, 18 per cent chromium.

Straight pieces of the stainless-steel ribbon were given the same series of heat-treatments as the springs. Young's modulus was determined for each of these samples and a bending test was also applied to determine whether the wire had been permanently annealed. No appreciable effect was noted. The proportional limit and the ultimate tensile strength of this ribbon at room temperature as measured on a tensile-testing machine were 41,600 and 252,000 psi, respectively. It is thus seen that except with the high-temperature anneals the residual stress alone exceeds the proportional limit and any additional stress will cause a permanent deformation.

It is also possible to obtain a favorable residual-stress distribution, that is, an initial compression on the inner and a tension on the outer fibers. If the released spring having the stress distribution shown in Fig. 7(c) is expanded until considerable plastic flow takes place on the inner and outer fibers the stress distribution of Fig. 9(a) is obtained, which on release results in the residual-stress distribution as shown in Fig. 9(b).

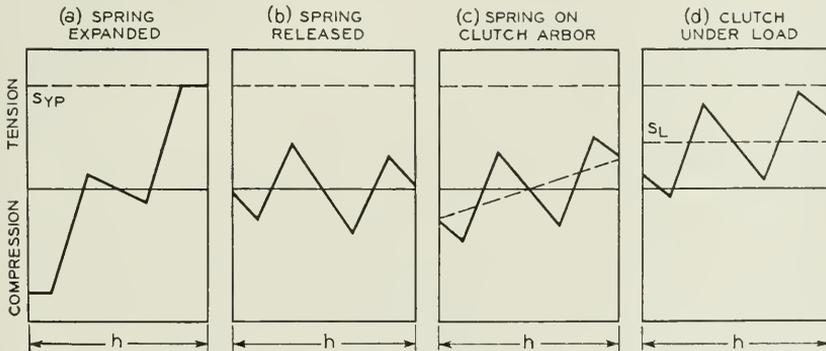


Fig. 9—Stress conditions in an expanded clutch spring.

To verify the validity of these arguments three springs of stainless-steel ribbon with different preliminary treatments and one of heat-treated phosphor bronze were tested for backlash as a function of previous loading. The backlash angle was measured from a point of slipping in the free direction to the point in the holding direction at which it would sustain a load of 0.05 in.-lb. An initial load of 0.5 in.-lb. was then applied and removed and the backlash measured as before. This was repeated for various loads up to the breaking point of the spring. The results are shown in Fig. 10. In the case of the untreated stainless steel the backlash began to increase immediately. Its higher initial value was probably due to the unavoidable stressing occasioned

by assembling the spring on the clutch arbor and to the 0.05-in.-lb. testing torque. The backlash of the other three samples remained constant to slightly above 0.5 in.-lb. and then began to rise. At low loads the phosphor bronze was better while at higher loads the heat-treated stainless steel had the advantage; the breaking load for the latter was also about 30 per cent higher.

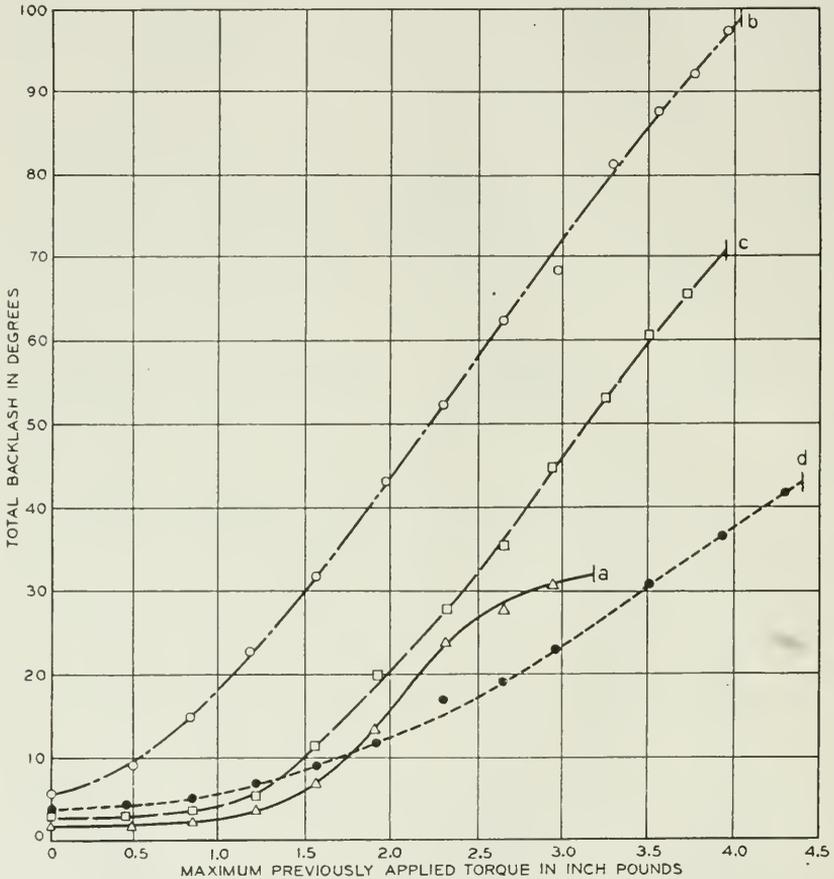


Fig. 10—Effect of overload on clutch backlash. Arbor diameter, 0.190 in.

(a) Phosphor-bronze-ribbon spring, 0.0085×0.022 in., heat-treated 4 hr. at 230°C ., 0.1835 in. ID.

(b) 18-8 stainless-steel ribbon, 0.0068×0.021 in., wound on 0.127-in. mandrel and released, 0.183 in. ID.

(c) 18-8 stainless-steel ribbon, 0.0068×0.021 in., wound on 0.120-in. mandrel and released, 0.170 in. ID, mechanically expanded to 0.183 in. ID.

(d) 18-8 stainless-steel ribbon, 0.0068×0.021 in., wound on 0.157-in. mandrel, heated 4 hr. at 470°C ., cooled, and released, 0.182 in. ID.

CONCLUSION

The relations developed in the preceding sections are sufficient to determine uniquely the correct spring dimensions for a spring clutch of specified free and gripping torque provided the material constants and the cross-sectional shape (round or rectangular) of the wire, as well as the length and diameter of the clutch arbor, are specified. Choice of values for the last two factors is based largely on permissible heating resulting from the slipping in the free direction. Furthermore, as would generally be the case, if only maximum values of the free torque and minimum values of the gripping torque are given a number of solutions can be obtained from which to choose the most convenient. By combining the relations derived, in a manner to permit step-by-step calculation, the design of spring clutches is reduced to a simple routine.

Abstracts of Technical Articles from Bell System Sources

*Recent Observations on the Relation between Penetration, Infection and Decay in Creosoted Southern Pine Poles in Line.*¹ C. H. AMADON. The relation between poor penetration and decay, and the necessity for rational and adequate penetration requirements in treating specifications, are now fairly well understood by producers and users of creosoted southern pine poles. The purpose of this brief paper is to supplement the information presented in the Proceedings for 1936 and 1937 on the behavior of these poles in line under actual service conditions.

*Tarnish Studies. The Electrolytic Reduction Method for the Analysis of Films on Metal Surfaces.*² W. E. CAMPBELL and U. B. THOMAS. A method is described for analysis of tarnish films on metals by electrolytic reduction at the cathode. Its suitability is demonstrated for the rapid and accurate measurement of oxide films on copper varying in average thickness from monomolecular layers to 1000 Å. It is shown to be useful for reduction of mixed oxide-sulfide films on copper and silver. The method is used to measure the oxide films on freshly reduced copper after one-half hour's exposure to oxygen or air. Such films are shown to be 10–20 Å thick. A thicker film, measuring 30–70 Å is found to be produced by abrasion of copper in air, water, benzene or toluene. Adaptations and modifications are discussed which give wide analytical application to the method.

*An Electrochemical Study of the Corrosion of Painted Iron.*³ H. E. HARING and R. B. GIBNEY. The corrosion protective value of approximately 50 different paints was determined by means of an electrochemical method which has been previously described. This determination involved the measurement of the change in the potential of the painted iron with time when wet with water for 24 hr. or less. It was found that the interpretation of the time-potential curves which were automatically plotted by a recording vacuum tube electrometer, was facilitated if the test was conducted in a nitrogen atmosphere. The results obtained with the electrochemical or potentiometric method compared favorably with those obtained in a

¹ *Proc. American Wood-Preservers' Association*, 1939.

² *Electrochemical Society Preprint* 76–25.

³ *Electrochemical Society Preprint* 76–24.

one-year outdoor exposure test. Such differences as were found were shown to be due either to deterioration or improvement in the paint film as the result of weathering.

*Characteristics of Modern Microphones for Sound Recording.*⁴ F. L. HOPPER. Factors influencing the choice of a microphone for sound recording are considered. The characteristics of a new miniature condenser transmitter and amplifier, as well as a number of other types of microphones now in use, are included.

*Cold-Cathode Gas-Filled Tubes as Circuit Elements.*⁵ S. B. INGRAM. The application of electronic devices to the local systems plant is still in its infancy. One of the first of these devices to receive extensive use is the cold-cathode gas-filled tube. As a sensitive relay it is beginning to make its appearance in a number of telephone control and signaling circuits, being best known for its use in the standard four-party subscriber set where its rectifying property enables it to discriminate between positive and negative polarity for selective ringing. Compared with other types of vacuum tubes the cold-cathode tube has the advantages that it operates without cathode heating power, has the ability to start immediately when a signal is applied, and does not deteriorate when not passing current. These advantages make it particularly suitable for use in telephone circuits where intermittent service is common and long life and economical operation are required.

The paper describes the structure and electrical characteristics of cold-cathode tubes. Their properties as circuit elements are then illustrated in a number of typical basic circuits.

*Inductive Coordination with Series Sodium Highway Lighting Circuits.*⁶ H. E. KENT and P. W. BLYE. This paper describes the wave-shape characteristics of the sodium-vapor lamp and discusses the relative inductive influence of various series circuit arrangements in which such lamps are employed. A method is outlined by means of which the noise to be expected in an exposed telephone line may be estimated. Measures are described which may be applied in the telephone plant or in the lighting circuit to assist in the inductive coordination of the two systems. These measures need be considered only when a considerable number of lamps is involved, since noise induction is negligible

⁴ *Jour. S.M.P.E.*, September 1939.

⁵ *Elec. Engg.*, July 1939.

⁶ *Elec. Engg.*, July 1939.

when there are only a few lamps as, for instance, at highway intersections.

*A Cardioid Directional Microphone.*⁷ R. N. MARSHALL and W. R. HARRY. A microphone is described which has uniform directivity over a wide frequency range. This is made possible by placing in a single instrument a dynamic type pressure microphone element and a ribbon type "velocity" element, and electrically equalizing the outputs before combination. The resultant directional pattern is a heart-shaped curve or cardioid, giving a fairly wide pick-up zone in front and a substantial dead zone at the back of the instrument. Because of the unusually rugged ribbon employed, the new microphone is much less susceptible to wind noise than ordinary ribbon types. Housed in an aluminum case, the microphone weighs only $3\frac{1}{4}$ lbs. High output level, low impedance, and high quality, together with the excellent directivity, promise to make the cardioid microphone an important tool for the motion picture sound engineer.

*Fractional-Frequency Generators Utilizing Regenerative Modulation.*⁸ R. L. MILLER. By the application of the principle of regeneration to certain modulation systems, a generator of submultiple or other fractional-frequency ratio may be obtained.

A simple example is obtained by considering a second-order modulator whose output is connected back to a conjugate input by means of a feedback loop including an amplifier and a selective network. If an input frequency f_0 is applied, it is found that a frequency component $f_0/2$ appearing in the feedback path will modulate with the applied frequency to produce sidebands of $f_0/2$ and $3f_0/2$. The network and amplifier, being especially efficient for the frequency $f_0/2$ and having a gain higher than the modulator loss, will reinforce this component causing it to build up to some steady-state value. Similar processes are possible by which greater submultiple ratios may be obtained.

Since the output wave is obtained by a modulation process involving the input wave, it will appear only when an input is applied and then bears a fixed frequency ratio with respect to it. Experiments show that the ability of the generator to produce a fractional frequency is independent of phase shift in the feedback path. Circuits are possible in which the amplitude of the fractional-frequency wave will bear a linear relation to the input wave over a reasonable range and at the

⁷ *Jour. S.M.P.E.*, September 1939.

⁸ *Proc. I.R.E.*, July 1939.

same time maintain a constant phase angle between the two waves. Typical circuits are discussed which make use of copper oxide as the modulator elements.

*Seasonal Cosmic-Ray Effects at Sea Level.*⁹ R. A. MILLIKAN, H. V. NEHER and D. O. SMITH. By sending a Neher self-recording electroscope in a 10-cm lead shield repeatedly on a slow Norwegian steamer over the route Vancouver-Los Angeles, around South America and return to Los Angeles and Vancouver, we find (1) as heretofore an equatorial dip measured from Los Angeles of seven per cent on the western side of South America, eight per cent on the eastern side; (2) no measurable seasonal effect, or winter-summer differences, at all in the voyage from Los Angeles to the Straits of Magellan; (3) as heretofore constancy in cosmic-ray intensity in summer and fall, within the limits of uncertainty imposed by fluctuations estimated at not over one per cent, on the voyage between Los Angeles and Vancouver; (4) but in winter and spring an increase of as much as two or three per cent between Los Angeles and Vancouver. This is interpreted as the atmospheric-temperature effect earlier studied by Hess, Compton, and their respective collaborators.

*Some Engineering Considerations in Loading Circuits.*¹⁰ J. A. PARROTT. This paper describes the various loading arrangements used on toll entrance and intermediate cable circuits and discusses the transmission benefits obtained by loading and some of the important problems in the consideration of loading railroad entrance and intermediate cables. In addition to voice frequency loading, loading for the lower frequency carrier systems such as the Type H is also discussed.

*The Formation of Metallic Bridges between Separated Contacts.*¹¹ G. L. PEARSON. Low resistance bridges were formed between gold, steel and carbon electrodes having separations of $2-70 \times 10^{-6}$ cm by applying voltages less than the minimum sparking potential. For a given pair of electrodes the field required to form the bridges is a constant and is $5-16 \times 10^6$ volts per centimeter. Measurements of the temperature coefficient of resistance of the bridges identify them as consisting of the material of the electrodes. A study of their resistance as a function of the displacement of one of the electrodes shows that they may be pulled out as well as crushed. At voltages

⁹ *Phys. Rev.*, September 15, 1939.

¹⁰ *Proc., Assoc. Amer. R.R., Telegraph and Telephone Section*, April 1939.

¹¹ *Phys. Rev.*, September 1, 1939.

less than those required to form the bridges, field currents exist. These increase rapidly as the field is raised and attain a value around 10^{-10} ampere before the bridges are formed. Calculation of the maximum electrostatic stress on the electrodes at the time of breakdown gives a value 0.05 to 0.0005 times the tensile strength of the electrode material at room temperature. The field is locally higher than that calculated because of surface roughness and the tensile strength is probably lowered by the local heating known to accompany field currents. The data therefore indicate that electrostatic force pulls material from the electrodes to bridge the gap.

*Measuring Transmission Speed of the Coaxial Cable.*¹² J. F. WENTZ. Time of transmission of carrier currents over high speed lines is discussed. A method of measuring this time delay as used on the 1000-kc system of the New York-Philadelphia coaxial cable is described and the results are given for the television band transmitted over it experimentally.

¹² *Bell Labs. Record*, June 1939.

Contributors to this Issue

M. J. AYKROYD, B.S. in Civil Engineering, Queen's University, 1913. Bitulithic and Contracting Company, Edmonton, 1913-14; Department of Public Works, Ottawa, 1915-16; Imperial Ministry of Munitions, Montreal and New York, 1916-18; Manager and Director, Export and Import Company, Montreal and London, England, 1919-23. Bell Telephone Company of Canada, Plant Department, Montreal and Toronto, 1923-34; Outside Plant Engineer, Bell Telephone Company of Canada, Toronto, 1934-.

D. G. GEIGER, Queen's University: B.S. in Electrical Engineering, 1922; B.S. in Mechanical Engineering, 1923; Department of Electrical Engineering, 1923-24. Bell Telephone Company of Canada, Montreal, Transmission Division of General Engineering Department, 1924-26 and 1928-29. Lecturer in Electrical Engineering, Queen's University, 1926-28. Transmission Engineer, Bell Telephone Company of Canada, Toronto, 1930-.

EGINHARD DIETZE, B.S. in Electrical Engineering, University of Michigan, 1917. American Telephone and Telegraph Company, 1917-34; Bell Telephone Laboratories, 1934-. Mr. Dietze has been engaged in transmission studies of telephone stations, including room noise conditions at telephone locations. He is co-author of "Indicating Meter for Measurement and Analysis of Noise."

WALTER D. GOODALE, JR., E.E., Lehigh University, 1928; M.E.E., Polytechnic Institute of Brooklyn, 1937. American Telephone and Telegraph Company, Department of Development and Research, 1928-34; Bell Telephone Laboratories, 1934-. Mr. Goodale has been engaged in studies of various factors affecting telephone station transmission.

W. B. BEDELL, Ohio Northern University, U. S. Army, 1917-1919 (Lieutenant, Infantry); American Telephone and Telegraph Company, Long Lines Engineering Department, 1919-. Mr. Bedell is engaged in equipment engineering work involving broad-band carrier telephone systems.

GLEN B. RANSOM, B.S. in Electrical Engineering, University of Minnesota, 1921. American Telephone and Telegraph Company, Long Lines Department, at Chicago, 1922-27; District Engineer at Indianapolis, 1927-28; Division Transmission Engineer at Cleveland, 1928-30; Long Lines Engineering Department, Transmission Branch, 1930-. Appointed to present position, Circuit Layout Engineer, 1932.

W. A. STEVENS, B.S. in Electrical Engineering, Bucknell University, 1925. New York Telephone Company, Engineering Department, 1925-28. American Telephone and Telegraph Company, Department of Operation and Engineering, Transmission Engineering Group, 1928-. Mr. Stevens' work has largely dealt with toll transmission matters.

B. D. HOLBROOK, A.B., Stanford University, 1924; A.M., 1925. University of Chicago, 1926-30. Bell Telephone Laboratories, 1930-. Mr. Holbrook has engaged in research on broad-band carrier telephony and on signaling systems.

J. T. DIXON, B.E. in Electrical Engineering, Johns Hopkins University, 1924; S.M. in Electrical Engineering, Massachusetts Institute of Technology, 1926. American Telephone and Telegraph Company, Department of Development and Research, 1926-34; Bell Telephone Laboratories, 1934-. Mr. Dixon has been engaged in development work on carrier systems.

W. SHOCKLEY, B.Sc., California Institute of Technology, 1932; Ph.D., Massachusetts Institute of Technology, 1936. Bell Telephone Laboratories, 1936-. Dr. Shockley's work in the Laboratories has been concerned with problems in electronics.

C. F. WIEBUSCH, B.A., 1924, M.A., 1925, University of Texas. Instructor, Department of Physics, University of Texas, 1925-1927. Bell Telephone Laboratories, 1927-. Mr. Wiebusch was involved for a number of years in the development of disc recording and reproducing apparatus and loud speakers and for the past few years has been engaged in the development of telephone signaling and registering apparatus.

