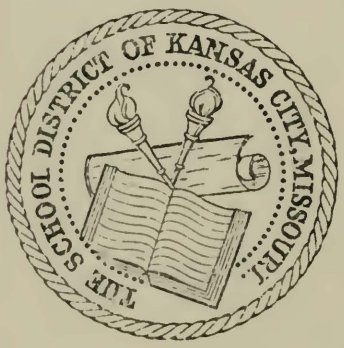
Bound Periodical 864171

Kansas City
Public Library



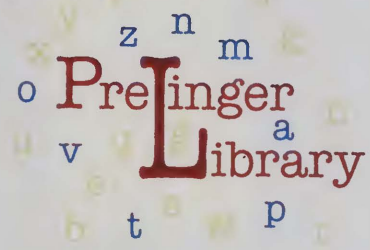
This Volume is for

REFERENCE USE ONLY

11-34 —6m—P

PUBLIC LIBRARY
KANSAS CITY
MO

From the collection of the



San Francisco, California
2008

PUBLIC LIBRARY
KANSAS CITY
MO

VOLUME XIV JANUARY, 1935 NUMBER 1

PUBLIC LIBRARY
KANSAS CITY
MO

THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

Wide Band Transmission Over Balanced Circuits—*A. B Clark* 1

A Survey of Magnetic Materials in Relation to Structure—*W. C. Ellis and Earle E. Schumacher* 8

Theory of Multi-Electrode Vacuum Tubes—*H. A. Pidgeon* 44

Limits to Amplification—*J. B. Johnson and F. B. Llewellyn* 85

Vacuum Tubes as High-Frequency Oscillators—*M. J. Kelly and A. L. Samuel* 97

Horizontal Rhombic Antennas—*E. Bruce, A. C. Beck and L. R. Lowry* 135

Extraneous Frequencies Generated in Air Carrying Intense Sound Waves—*A. L. Thuras, R. T. Jenkins and H. T. O'Neil* 159

Abstracts of Technical Papers 173

Contributors to this Issue 176

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

50c per Copy *$1.50 per Year*

PUBLIC LIBRARY
KANSAS CITY
MO

THE BELL SYSTEM TECHNICAL JOURNAL

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

EDITORIAL BOARD

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

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

SUBSCRIPTIONS

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

Copyright, 1935

PRINTED IN U. S. A.

THE BELL SYSTEM TECHNICAL JOURNAL

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

EDITORIAL BOARD

BANCROFT GHERARDI H. P. CHARLESWORTH F. B. JEWETT
A. F. DIXON E. H. COLPITTS O. B. BLACKWELL
D. LEVINGER O. E. BUCKLEY H. S. OSBORNE
R. W. KING, *Editor* J. O. PERRINE, *Associate Editor*

TABLE OF CONTENTS AND INDEX

VOLUME XIV

1935

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

THE BELL SYSTEM TECHNICAL JOURNAL

VOLUME XIV, 1935

Table of Contents

JANUARY, 1935

Wide Band Transmission Over Balanced Circuits— *A. B. Clark* .. 1
A Survey of Magnetic Materials in Relation to Structure— *W. C. Ellis and Earle E. Schumacher* 8
Theory of Multi-Electrode Vacuum Tubes—*H. A. Pidgeon* 44
Limits to Amplification—*J. B. Johnson and F. B. Llewellyn* 85
Vacuum Tubes as High-Frequency Oscillators— *M. J. Kelly and A. L. Samuel* 97
Horizontal Rhombic Antennas— *E. Bruce, A. C. Beck and L. R. Lowry* 135
Extraneous Frequencies Generated in Air Carrying Intense Sound Waves—*A. L. Thuras, R. T. Jenkins and H. T. O'Neil* 159

APRIL, 1935

Cable Crosstalk—Effect of Non-Uniform Current Distribution in the Wires—*R. N. Hunter and R. P. Booth* 179
Experiments with Directivity Steering for Fading Reduction— *E. Bruce and A. C. Beck* 195
A General Theory of Electric Wave Filters—*H. W. Bode* 211
Ideal Wave Filters—*H. W. Bode and R. L. Dietzold* 215
Ultra-Short-Wave Propagation: Mobile Urban Transmission Characteristics—*C. R. Burrows, L. E. Hunt and A. Decino* .. 253
An Application of Number Theory to the Splicing of Telephone Cables—*H. P. Lawther, Jr.* 273
Contemporary Advances in Physics, XXIX—The Nucleus, Fourth Part—*Karl K. Darrow* 285
Ferromagnetic Distortion of a Two-Frequency Wave— *Robert M. Kalb and William R. Bennett* 322

Bound Periodical
JA 7 '36
864171

JULY, 1935

Further Results of a Study of Ultra-Short-Wave Transmission Phenomena—
C. R. Englund, A. B. Crawford and W. W. Mumford 369
Acoustical Instruments—*E. C. Wente* 388
Thermionic Electron Emission—*J. A. Becker* 413
Radio Propagation Over Spherical Earth—*Chas. R. Burrows* 477
A Single-Sideband Short-Wave System for Transatlantic Telephony—*F. A. Polkinghorn and N. F. Schlaack* 489
Mutual Impedances of Parallel Wires—
Ray S. Hoyt and Sallie Pero Mead 509
An Unattended Ultra-Short-Wave Radio Telephone System—
N. F. Schlaack and F. A. Polkinghorn 534
Around the World by Telephone 542

OCTOBER, 1935

Dr. George A. Campbell—*F. B. Jewett* 553
Dr. Campbell's Memoranda of 1907 and 1912 558
Some Aspects of Low-Frequency Induction Between Power and Telephone Circuits—*H. R. Huntley and E. J. O'Connell* 573
Circulating Currents and Singing on Two-Wire Cable Circuits—
L. G. Abraham 600
Operation of Ultra-High-Frequency Vacuum Tubes—
F. B. Llewellyn 632
Further Extensions of Theory of Multi-Electrode Vacuum Tube Circuits—*S. A. Levin and L. C. Peterson* 666
Transatlantic Long-Wave Transmission—
Austin Bailey and H. M. Thomson 680
Technical Digests—
1. Superiorities of Lead-Calcium Alloys for Storage Battery Construction—*H. E. Haring, U. B. Thomas, E. E. Schumacher and G. S. Phipps* 699
2. Marine Radio Telephone Service for Boston Harbor—
F. A. Gifford and R. B. Meader 702
3. Harbor Craft Ship-to-Shore Radio Telephone Service in Puget Sound Area—*E. B. Hansen* 708
4. Ship Sets for Harbor Ship-to-Shore Service—
H. N. Willets 713
5. An Electromechanical Representation of a Piezoelectric Crystal—*W. P. Mason* 718

Index to Volume XIV

A

Abraham, L. G., Circulating Currents and Singing on Two-Wire Cable Circuits, page 600.
Acoustical Instruments, *E. C. Wente*, page 388.
Alloys, Lead-Calcium, Superiorities of for Storage Battery Construction (a Digest), *H. E. Haring, U. B. Thomas, E. E. Schumacher and G. S. Phipps*, page 699.
Amplification, Limits to, *J. B. Johnson and F. B. Llewellyn*, page 85.
Antennas, Horizontal Rhombic, *E. Bruce, A. C. Beck and L. R. Lowry*, page 135.

B

Bailey, Austin and H. M. Thomson, Transatlantic Long-Wave Transmission, page 680.
Balanced Circuits, Wide Band Transmission Over, *A. B. Clark*, page 1.
Battery Construction, Storage, Superiorities of Lead-Calcium Alloys for (a Digest), *H. E. Haring, U. B. Thomas, E. E. Schumacher and G. S. Phipps*, page 699.
Beck, A. C. and E. Bruce, Experiments with Directivity Steering for Fading Reduction, page 195.
Beck, A. C., E. Bruce and L. R. Lowry, Horizontal Rhombic Antennas, page 135.
Becker, J. A., Thermionic Electron Emission, page 413.
Bennett, William R. and Robert M. Kalb, Ferromagnetic Distortion of a Two-Frequency Wave, page 322.
Bode, H. W., A General Theory of Electric Wave Filters, page 211.
Bode, H. W. and R. L. Dietzold, Ideal Wave Filters, page 215.
Booth, R. P. and R. N. Hunter, Cable Crosstalk—Effect of Non-Uniform Current Distribution in the Wires, page 179.
Boston Harbor, Marine Radio Telephone Service for (a Digest), *F. A. Gifford and R. B. Meader*, page 702.
Bruce, E. and A. C. Beck, Experiments with Directivity Steering for Fading Reduction, page 195.
Bruce, E., A. C. Beck and L. R. Lowry, Horizontal Rhombic Antennas, page 135.
Burrows, Chas. R., Radio Propagation Over Spherical Earth, page 477.
Burrows, C. R., L. E. Hunt and A. Decino, Ultra-Short-Wave Propagation: Mobile Urban Transmission Characteristics, page 253.

C

Cable Circuits, Two-Wire, Circulating Currents and Singing on, *L. G. Abraham*, page 600.
Cable Crosstalk—Effect of Non-Uniform Current Distribution in the Wires, *R. N. Hunter and R. P. Booth*, page 179.
Cables, Telephone, An Application of Number Theory to the Splicing of, *H. P. Lawther, Jr.*, page 273.
Campbell, Dr. George A., *F. B. Jewett*, page 553.
Campbell's Memoranda of 1907 and 1912, Dr., page 558.
Clark, A. B., Wide Band Transmission Over Balanced Circuits, page 1.
Contemporary Advances in Physics, XXIX—The Nucleus, Fourth Part, *Karl K. Darrow*, page 285.
Crawford, A. B., C. R. Englund and W. W. Mumford, Further Results of a Study of Ultra-Short-Wave Transmission Phenomena, page 369.
Crosstalk, Cable—Effect of Non-Uniform Current Distribution in the Wires, *R. N. Hunter and R. P. Booth*, page 179.
Crystal, Piezoelectric, An Electromechanical Representation of a (a Digest), *W. P. Mason*, page 718.
Current Distribution in the Wires, Non-Uniform, Effect of—Cable Crosstalk, *R. N. Hunter and R. P. Booth*, page 179.

D

Darrow, Karl K., Contemporary Advances in Physics, XXIX—The Nucleus, Fourth Part, page 285.
Decino, A., C. R. Burrows and L. E. Hunt, Ultra-Short-Wave Propagation: Mobile Urban Transmission Characteristics, page 253.
Dietzold, R. L. and H. W. Bode, Ideal Wave Filters, page 215.
Directivity Steering for Fading Reduction, Experiments with, *E. Bruce and A. C. Beck*, page 195.

E

Ellis, W. C. and Earle E. Schumacher, A Survey of Magnetic Materials in Relation to Structure, page 8.
Englund, C. R., A. B. Crawford and W. W. Mumford, Further Results of a Study of Ultra-Short-Wave Transmission Phenomena, page 369.

F

Fading Reduction, Experiments with Directivity Steering for, *E. Bruce and A. C. Beck*, page 195.
Ferromagnetic Distortion of a Two-Frequency Wave, *Robert M. Kalb and William R. Bennett*, page 322.
Filters, Electric Wave, A General Theory of, *H. W. Bode*, page 211.
Filters, Ideal Wave, *H. W. Bode and R. L. Dietzold*, page 215.
Frequency Vacuum Tubes, Ultra-High-, Operation of, *F. B. Llewellyn*, page 632.
Frequency: Extraneous Frequencies Generated in Air Carrying Intense Sound Waves, *A. L. Thuras, R. T. Jenkins and H. T. O'Neil*, page 159.
Frequency: Wide Band Transmission Over Balanced Circuits, *A. B. Clark*, page 1.
Frequency: Vacuum Tubes as High-Frequency Oscillators, *M. J. Kelly and A. L. Samuel*, page 97.

G

Gifford, F. A. and R. B. Meader, Marine Radio Telephone Service for Boston Harbor (a Digest), page 702.
Gifford, W. S.—Around the World by Telephone, page 542.

H

Hansen, E. B., Harbor Craft Ship-to-Shore Radio Telephone Service in Puget Sound Area (a Digest), page 708.
Haring, H. E., U. B. Thomas, E. E. Schumacher and G. S. Phipps, Superiorities of Lead-Calcium Alloys for Storage Battery Construction (a Digest), page 699.
High-Frequency Vacuum Tubes, Ultra-, Operation of, *F. B. Llewellyn*, page 632.
Hunt, L. E., C. R. Burrows and A. Decino, Ultra-Short-Wave Propagation: Mobile Urban Transmission Characteristics, page 253.
Hunter, R. N. and R. P. Booth, Cable Crosstalk—Effect of Non-Uniform Current Distribution in the Wires, page 179.
Huntley, H. R. and E. J. O'Connell, Some Aspects of Low-Frequency Induction Between Power and Telephone Circuits, page 573.
Hoyt, Ray S. and Sallie Pero Mead, Mutual Impedances of Parallel Wires, page 509.

I

Impedances, Mutual, of Parallel Wires, *Ray S. Hoyt and Sallie Pero Mead*, page 509
Induction, Low-Frequency, Some Aspects of Between Power and Telephone, Circuits, *H. R. Huntley and E. J. O'Connell*, page 573.
Instruments, Acoustical, *E. C. Wente*, page 388.
Intense Sound Waves, Extraneous Frequencies Generated in Air Carrying, *A. L. Thuras, R. T. Jenkins and H. T. O'Neil*, page 159.

J

Johnson J. B. and F. B. Llewellyn, Limits to Amplification, page 85.
Jenkins, R. T., A. L. Thuras and H. T. O'Neil, Extraneous Frequencies Generated in Air Carrying Intense Sound Waves, page 159.
Jewett, F. B., Dr. George A. Campbell, page 553.

K

Kalb, Robert M. and William R. Bennett, Ferromagnetic Distortion of a Two-Frequency Wave, page 322.
Kelly, M. J. and A. L. Samuel, Vacuum Tubes as High-Frequency Oscillators, page 97.

L

Lawther, H. P., An Application of Number Theory to the Splicing of Telephone Cables, page 273.
Levin, S. A. and L. C. Peterson, Further Extensions of Theory of Multi-Electrode Vacuum Tube Circuits, page 666.
Llewellyn, F. B., Operation of Ultra-High-Frequency Vacuum Tubes, page 632.
Llewellyn, F. B. and J. B. Johnson, Limits to Amplification, page 85.
Lowry, L. R., E. Bruce and A. C. Beck, Horizontal Rhombic Antennas, page 135.

M

Magnetic Materials in Relation to Structure, A Survey of, *W. C. Ellis and Earle E. Schumacher*, page 8.
Magnetism: Ferromagnetic Distortion of a Two-Frequency Wave, *Robert M. Kalb and William R. Bennett*, page 322.
Mason, W. P., An Electromechanical Representation of a Piezoelectric Crystal (a Digest), page 718.
Materials, Magnetic, A Survey of in Relation to Structure, *W. C. Ellis and Earle E. Schumacher*, page 8.
Mead, Sallie Pero, and Ray S. Hoyt, Mutual Impedances of Parallel Wires, page 509.
Meader, R. B. and F. A. Gifford, Marine Radio Telephone Service for Boston Harbor (a Digest), page 702.
Multi-Electrode Vacuum Tubes, Theory of, *H. A. Pidgeon*, page 44.
Mumford, W. W., C. R. Englund and A. B. Crawford, Further Results of a Study of Ultra-Short-Wave Transmission Phenomena, page 369.
Mutual Impedances of Parallel Wires, *Ray S. Hoyt and Sallie Pero Mead*, page 509.

O

O'Connell, E. J. and H. R. Huntley, Some Aspects of Low-Frequency Induction Between Power and Telephone Circuits, page 573.
O'Neil, H. T., A. L. Thuras and R. T. Jenkins, Extraneous Frequencies Generated in Air Carrying Intense Sound Waves, page 159.
Oscillators, High-Frequency, Vacuum Tubes as, *M. J. Kelly and A. L. Samuel*, page 97.

P

Peterson, L. C. and S. A. Levin, Further Extensions of Theory of Multi-Electrode Vacuum Tube Circuits, page 666.
Phipps, G. S., H. E. Haring, U. B. Thomas and E. E. Schumacher, Superiorities of Lead-Calcium Alloys for Storage Battery Construction (a Digest), page 699.
Physics, XXIX, Contemporary Advances in—The Nucleus, Fourth Part, *Karl K. Darrow*, page 285.
Pidgeon, H. A., Theory of Multi-Electrode Vacuum Tubes, page 44.
Piezoelectric Crystal, An Electromechanical Representation of a (a Digest), *W. P. Mason*, page 718.
Polkinghorn, F. A. and N. F. Schlaack, A Single-Sideband Short-Wave System for Transatlantic Telephony, page 489.
Polkinghorn, F. A. and N. F. Schlaack, An Unattended Ultra-Short-Wave Radio Telephone System (a Digest), page 534.
Propagation, Radio, Over Spherical Earth, *Chas. R. Burrows*, page 477.
Propagation, Ultra-Short-Wave: Mobile Urban Transmission Characteristics, *C. R. Burrows, L. E. Hunt and A. Decino*, page 253.
Puget Sound Area, Harbor Craft Ship-to-Shore Radio Telephone Service in (a Digest), *E. B. Hansen*, page 708.

R

Radio Propagation Over Spherical Earth, *Chas. R. Burrows*, page 477.
Radio Telephone Service for Boston Harbor, Marine (a Digest), *F. A. Gifford and R. B. Meader*, page 702.
Radio Telephone Service in Puget Sound Area, Harbor Craft Ship-to-Shore (a Digest), *E. B. Hansen*, page 708.
Radio Telephone System, An Unattended Ultra-Short-Wave (a Digest), *N. F. Schlaack and F. A. Polkinghorn*, page 534.
Radio: Ultra-Short-Wave a Propagation: Mobile Urban Transmission Characteristics, *C. R. Burrows, L. E. Hunt and A. Decino*, page 253.
Radio: Further Results of a Study of Ultra-Short-Wave Transmission Phenomena, *C. R. Englund, A. B. Crawford and W. W. Mumford*, page 369.
Radio: A Single-Sideband Short-Wave System for Transatlantic Telephony, *F. A. Polkinghorn and N. F. Schlaack*, page 489.
Radio: Transatlantic Long-Wave Transmission, *Austin Bailey and H. M. Thomson*, page 680.
Rhombic Antennas, Horizontal, *E. Bruce, A. C. Beck and L. R. Lowry*, page 135.

S

Samuel, A. L. and M. J. Kelly, Vacuum Tubes as High-Frequency Oscillators, page 97.
Schlaack, N. F. and F. A. Polkinghorn, A Single-Sideband Short-Wave System for Transatlantic Telephony, page 489.
Schlaack, N. F. and F. A. Polkinghorn, An Unattended Ultra-Short-Wave Radio Telephone System (a Digest), page 534.
Schumacher, Earle E. and W. C. Ellis, A Survey of Magnetic Materials in Relation to Structure, page 8.
Schumacher, E. E., H. E. Haring, U. B. Thomas and G. S. Phipps, Superiorities of Lead-Calcium Alloys for Storage Battery Construction (a Digest), page 699.
Ship-to-Shore Radio Telephone Service in Puget Sound Area, Harbor Craft (a Digest), *E. B. Hansen*, page 708.
Ship-to-Shore Service, Harbor, Ship Sets for (a Digest), *H. N. Willets*, page 713.
Short-Wave Propagation, Ultra: Mobile Urban Transmission Characteristics: *C. R. Burrows, L. E. Hunt, and A. Decino*, page 253.
Short-Wave System, A Single-Sideband, for Transatlantic Telephony, *F. A. Polkinghorn and N. F. Schlaack*, page 489.
Short-Wave, Ultra-, Radio Telephone System, An Unattended (a Digest), *N. F. Schlaack and F. A. Polkinghorn*, page 534.
Short-Wave Transmission Phenomena, Ultra, Further Results of a Study of, *C. R. Englund, A. B. Crawford and W. W. Mumford*, page 369.
Singing on Two-Wire Cable Circuits, Circulating Currents and, *L. G. Abraham*, page 600.
Single-Sideband Short-Wave System for Transatlantic Telephony, *F. A. Polkinghorn and N. F. Schlaack*, page 489.
Sound Waves, Intense, Extraneous Frequencies Generated in Air Carrying, *A. L. Thuras, R. T. Jenkins and H. T. O'Neil*, page 159.
Splicing of Telephone Cables, An Application of Number Theory to the, *H. P. Lawther, Jr.*, page 273.
Storage Battery Construction, Superiorities of Lead-Calcium Alloys for (a Digest), *H. E. Haring, U. B. Thomas, E. E. Schumacher and G. S. Phipps*, page 699.
Structure, A Survey of Magnetic Materials in Relation to, *W. C. Ellis and Earle E. Schumacher*, page 8.

T

Telephone, Around the World by, page 542.
Telephony, Transatlantic, A Single-Sideband Short-Wave System for, *F. A. Polkinghorn and N. F. Schlaack*, page 489.
Thermionic Electron Emission, *J. A. Becker*, page 413.
Thomas, U. B., H. E. Haring, E. E. Schumacher, and G. S. Phipps, Superiorities of Lead-Calcium Alloys for Storage Battery Construction (a Digest), page 699.
Thomson, H. M. and Austin Bailey, Transatlantic Long-Wave Transmission, page 680.

Thuras, A. L., R. T. Jenkins and H. T. O'Neil, Extraneous Frequencies Generated in Air Carrying Intense Sound Waves, page 159.
Transatlantic Telephony, A Single-Sideband Short-Wave System for, *F. A. Polkinghorn and N. F. Schlaack*, page 489.
Transatlantic Long-Wave Transmission, *Austin Bailey and H. M. Thomson*, page 680.
Transmission, Transatlantic Long-Wave, *Austin Bailey and H. M. Thomson*, page 680.
Transmission Phenomena, Ultra-Short-Wave, Further Results of a Study of, *C. R. Englund, A. B. Crawford and W. W. Mumford*, page 369.
Tubes, Ultra-High-Frequency Vacuum, Operation of, *F. B. Llewellyn*, page 632.

U

Ultra-High-Frequency Vacuum Tubes, Operation of, *F. B. Llewellyn*, page 632.
Ultra-Short-Wave Propagation: Mobile Urban Transmission Characteristics, *C. R. Burrows, L. E. Hunt and A. Decino*, page 253.
Ultra-Short-Wave Radio Telephone System, An Unattended (a Digest), *N. F. Schlaack and F. A. Polkinghorn*, page 534.
Ultra-Short-Wave Transmission Phenomena, Further Results of a Study of, *C. R. Englund, A. B. Crawford and W. W. Mumford*, page 369.

V

Vacuum Tubes, Multi-Electrode, Theory of, *H. A. Pidgeon*, page 44.
Vacuum Tubes as High-Frequency Oscillators, *M. J. Kelly and A. L. Samuel*, page 97.
Vacuum Tubes, Ultra-High-Frequency, Operation of, *F. B. Llewellyn*, page 632.

W

Wave Propagation, Ultra-Short: Mobile Urban Transmission Characteristics, *C. R. Burrows, L. E. Hunt and A. Decino*, page 253.
Wave System, A Single-Sideband Short-, for Transatlantic Telephony, *F. A. Polkinghorn and N. F. Schlaack*, page 489.
Wave, Ultra-Short-, Radio Telephone System, An Unattended (a Digest), *N. F. Schlaack and F. A. Polkinghorn*, page 534.
Wave, Long-, Transmission, Transatlantic, *Austin Bailey and H. M. Thomson*, page 680.
Wave, Ultra-Short-, Transmission Phenomena, Further Results of a Study of, *C. R. Englund, A. B. Crawford and W. W. Mumford*, page 369.
Waves, Intense Sound, Extraneous Frequencies Generated in Air Carrying, *A. L. Thuras, R. T. Jenkins and H. T. O'Neil*, page 159.
Wente, E. C., Acoustical Instruments, page 388.
Wide Band Transmission Over Balanced Circuits, *A. B. Clark*, page 1.
Willets, H. N., Ship Sets for Harbor Ship-to-Shore Service (a Digest), page 713.

PRINTED IN U. S. A.

The Bell System Technical Journal

January, 1935

Wide Band Transmission Over Balanced Circuits *

By A. B. CLARK

In a recent paper [1] amplifiers capable of handling frequency band widths of the order of 1,000 kc. or more are described, together with terminal apparatus for effectively utilizing these wide bands for telephone, telegraph and television purposes. The paper confines itself to the coaxial line structure for transmitting these wide frequency ranges but points out that broad-band transmission is also applicable to balanced conductor systems. The present paper discusses briefly some of the possibilities of the more familiar balanced circuits, circuits more or less as they now exist in the present plant being first considered, following which are circuits obtained by new construction. Wide-band transmission over balanced circuits offers interesting possibilities both for circuits in the present plant and for new construction.

A HIGH degree of *electrical balance* [2] has been for a long time a fundamental requirement of telephone transmission lines. This has been required not only to prevent interference entering into telephone circuits from other types of electrical circuits but also to prevent mutual interference between the closely adjacent telephone circuits on open-wire lines or in cables.

In the central offices, to be sure, unbalance in apparatus has been frequently employed for simplicity and convenience. In toll circuits, however, such office unbalance has been electrically separated from the outside plant by the use of repeating coils or otherwise. In local circuits the high standard of balance required for toll circuits has not generally been necessary since the exposures to interfering fields are less severe and the range of speech levels is much smaller. However, when local circuits are connected to toll circuits the unbalances are kept electrically separated from the toll circuits by the use of repeating coils. In recent years the tendency to the use of higher frequencies in com-

* Published in January, 1935 issue of *Electrical Engineering*. Scheduled for presentation at Winter Convention of A. I. E. E., New York City, January 22–25, 1935.

[1] "Systems for Wide-Band Transmission Over Coaxial Lines" by L. Espenschied and M. E. Strieby, *Elec. Engg.*, October, 1934; *Bell Sys. Tech. Jour.*, October, 1934.

[2] As used here, the term *electrical balance* refers to the two sides of a telephone circuit. To secure such balance, the aim has been to construct the go and return conductors of each circuit of the same gauge and material and to locate them symmetrically with respect to earth and to surrounding conductors. The aim has also been to apply to each circuit terminal apparatus symmetrical with respect to its series impedances and shunt admittances to ground.

munication circuits, the increase in the strength of interfering fields and the development of highly efficient amplifiers has led to constantly more exacting requirements in electrical balance of telephone circuits.

The development of multichannel systems by carrier methods employing constantly increasing frequency ranges has placed particularly exacting requirements on such electrical balance. In a recent paper, "Carrier in Cable," [3] are described the balancing methods which have been developed to permit the use of an increased frequency range in such cables. The recently published paper, "Systems for Wide Band Transmission Over Coaxial Lines" [1] points out the possibilities and possible requirements for very much wider frequency ranges. In that paper, coaxial lines are proposed which are particularly interesting in that they abandon electrical balance altogether and depend entirely on metallic shielding.

For such wide frequency range transmission, very interesting and important questions are raised, first as to the extent to which such wide bands can be placed on existing types of structure which are based on balance and, second, as to whether new construction designed particularly for such wide bands should depend on balance or shielding alone or a combination of the two. It is the purpose of this present paper to discuss these questions.

As noted in the Espenschied-Strieby paper, the apparatus described for broad-band transmission on concentric structures would also serve for other types of line structures. There would, of course, be problems in either balancing the apparatus or isolating its unbalances from the line structure. For the purposes of the present paper it is assumed that there will be no important reaction from the apparatus standpoint on this consideration of line balance and shielding.

Existing Cables

The attenuation of pairs in existing cables has a characteristic with respect to frequency generally similar to that of coaxial conductors but is, naturally, considerably higher because of the smaller physical dimensions and higher dielectric losses of the cable pair. For example, at a million cycles an ordinary 19-guage cable pair has an attenuation of about 18 db per mile, an ordinary 16-gauge pair about 14 db per mile, while a small-sized coaxial structure has an attenuation of about 6 db per mile. This means more repeaters for the cable pairs, three times as many for 19-gauge and a little more than twice as many for 16-gauge. Also, it means more difficulty in maintaining stability of transmission,

[3] A. B. Clark and B. W. Kendall, *Elec. Engg.*, July, 1933; *Bell Sys. Tech. Jour.*, July, 1933.

including overcoming of the variations due to the effect of temperature changes.

Stable and highly linear repeater gain can be produced so readily now-a-days, thanks to the negative feedback amplifier invented by Mr. H. S. Black,[4] that the idea of such high attenuations is no longer appalling even though on 16-gauge pairs it means repeaters spaced only about four miles apart. Overcoming the transmission variations due to temperature with automatic regulators introduces no fundamentally new problems, but, of course, the complexity and precision of regulation must be considerably higher due to the considerably larger variations.

Crosstalk, of course, must be given special consideration. First of all, it is necessary to restrict transmission of a given high-frequency band to only one direction in a single cable; the other direction must be supplied by another cable [3] or other separate transmission medium.

Considering transmission in one direction only, if only one pair in a cable is set aside for high-frequency transmission of, say, a million-cycle band, most, but not all, of the crosstalk difficulty can be avoided. However, the fact must be reckoned with that if one pair in a cable is singled out and an amplifier is applied having 60 db or more gain at a point intermediate between voice-frequency repeater stations, the amplifier will have a strong tendency to sing due to crosstalk between the pairs connected to the input and output and the other pairs in the cable. If two cables are available, this difficulty can be avoided by jumping from one cable to another every time the high-frequency amplification is introduced. If two cables are not available, overcoming the difficulty may call for the insertion of high-frequency choking devices in some or all of the low-frequency cable conductors at points where high-frequency amplification is introduced.

Considering now crosstalk between two circuits in a cable transmitting in the same direction, assuming the amplifier difficulty to have been overcome, tests have been made on various cables in the field from which these conclusions have been drawn. For telephone message purposes it is probably uneconomical to apply million-cycle frequency ranges to more than a single pair in an existing cable. However, with television, the crosstalk requirements are much less exacting. This is because the range of intensities necessary for a good television image is much less than is needed to accommodate message telephone subscribers and, therefore, a considerably larger ratio of extraneous current to maximum signal current can be tolerated. Tests indicate that two or more television channels, each a million

[4] "Stabilized Feedback Amplifiers," H. S. Black, *Elec. Engg.*, January, 1934; *Bell Sys. Tech. Jour.*, January, 1934.

cycles wide (possibly wider), can be transmitted over separate properly arranged pairs in the same direction in a single existing cable without serious disturbance due to crosstalk.

With respect to noise in existing cables, the matter of principal concern is noise produced in telephone offices by apparatus working on other circuits, since the natural shielding afforded by the cable sheath largely eliminates noise from outside sources. Two methods are available for control of the noise produced in telephone offices: (1) Introduction of high-frequency choking devices in all wires not assigned to carrier service at the points where the cables enter offices in which noise is produced; (2) Attack on noise at points where it is produced by introduction of spark-killers and individual high-frequency choking devices. The lenient noise-to-signal ratio requirement for television mentioned above makes high-frequency television application much easier than message telephone.

While a million-cycle frequency range over more than one pair in an existing cable seems unlikely for telephone message purposes, there are interesting possibilities in the use of lower maximum frequencies. For example, it seems likely that twelve same-directional telephone channels may be obtained from each one of a large fraction of the pairs in existing toll cables and that the crosstalk between the pairs may be kept within proper bounds by simple balancing methods previously described.[3]

Open Wire

With open wire, conditions are just about the reverse of those with cable. A mile of open wire has an attenuation of only about 1 db at a million cycles as compared to 6 db for the small-sized coaxial. However, overcoming crosstalk between different pairs of wires on a pole line presents very formidable problems while avoiding interference from and to radio systems may be even more formidable.

The attenuation of open-wire pairs has been checked up to several million cycles and it has been found that it behaves as expected. When there is little crosstalk from the high-frequency band to other wires in the lead the attenuation-frequency characteristic is smooth; when there is severe crosstalk, the characteristic is bumpy. For a given length of circuit the attenuation is small compared with that of the small-sized coaxial, and the variation due to changing weather conditions is also small—about one-third that of the coaxial. It is interesting to note, however, that the percentage change in attenuation for the coaxial is less than half that for the open-wire line. While it is, of course, evident that the open-wire transmission variations depend in part on changes in the series resistance of the wires due to changing

temperature and in part to changes in leakage and capacitance due to varying weather conditions, automatic transmission regulating systems similar in general principles to those already developed for other purposes should be adequate to maintain the required stability.

Crosstalk between different circuits becomes so severe at high frequencies that special transposition treatment or respacing of the wires becomes necessary. Minimizing interference from and to radio systems calls for a high degree of balance which may or may not dictate changes in the wire configuration. Here again it is necessary to distinguish between the requirements for television and for message telephone. Tests indicate that, in view of the more lenient television requirements, several million-cycle television channels can be transmitted over different pairs of a single open-wire line without serious disturbance and that to do this it will not be necessary to make radical changes in present wire configurations.

New Cables

For new construction, if television is not considered, effective carrier telephone systems may be set up by various methods. One might be a very broad-band method, a good example of which is given in the Espenschied-Strieby paper.[1] Another might be a much narrower band method using conductors similar to those in an ordinary cable. In the one case many telephone channels are obtained from a single pair by dividing up a frequency range, say one million cycles wide, into somewhat more than 200 channels. In the other case only 20 odd channels are obtained per pair of wires and use is made of 10 pairs of wires to obtain the same total number. It is too early to say which of these plans might be best under various practical conditions.

To meet future television needs, it may be necessary to provide for transmission of continuous frequency bands a million cycles in width or wider. It is interesting to compare the coaxial with balanced pairs surrounded by individual shields for such transmission.

For 6 db loss per mile at a million cycles, it works out that a solid copper coaxial unit with the rubber disc insulation described in the Espenschied-Strieby paper [1] has an internal diameter (inside of shield) of about 0.25 inch. For the same attenuation a pair of wires, each the same size as the central wire in the coaxial unit (70 mils diameter), insulated with rubber discs and with a copper shield, will have an inside diameter under the shield of about 0.4 inch. For outside conductors or shields made of lead the inside diameters become about 0.4 inch for the coaxial and 0.5 inch for the balanced pair. Therefore, if the thickness of the outer conductor is determined by mechanical considerations

rather than outside interference, the coaxial is smaller and cheaper. As the frequency is made higher, the shielding from outside interference afforded by the surrounding cylinder increases, so that at very high frequencies mechanical considerations alone control and the coaxial is clearly cheaper than the balanced shielded pair.

In the frequency range up to about a million cycles, however, interference from outside sources, including natural static and radio, must be considered in determining the thickness of the surrounding cylinder and the cost comparison is not so clear. It will be evident that as regards shielding, the balanced pair is at a large advantage because the two sides of the circuit are designed to be electrically similar. By proper care in manufacture this balance can readily be made sufficient to insure adequate shielding with a surrounding lead tube of thickness determined solely by mechanical considerations.

With a coaxial structure, however, it appears likely that to keep interference within proper bounds, a simple lead tube must be made considerably thicker than required by mechanical considerations, so that such a structure would probably be more expensive than a lead shielded pair. However, by adding other materials an adequately shielded coaxial unit can be constructed which will have a considerably thinner outside wall.

For example, there is described in the Espenschied-Strieby paper [1] a coaxial unit in which the inside diameter is minimized by first using copper tapes, the thickness of wall is minimized by adding thin iron tapes and the whole is made waterproof by a thin surrounding lead tube. This results in a unit of smaller inside and outside diameters than those of a lead tube surrounding a balanced pair of like attenuation. Since, however, the wall of the coaxial is thicker and the structure more complicated, the costs of the two units are estimated to be not greatly different when they are designed for the frequency range up to a million cycles. A minor advantage for the balanced pair remains, however, in that, whatever may be the top frequency, there is no limitation as to the lowest frequency permissible for interference reasons.

In the above discussion of new cable construction the amplifiers and transmission regulators required have not been mentioned. If similar conducting and insulating materials are used, shielded balanced pairs and coaxials have similar transmission-frequency characteristics. The variations with temperature are also similar. The factors which limit the overall amplifications are also the same. There is only one important point of difference between the amplifiers required for the two systems. This is the necessity for input and output transformers to be

balanced to ground with the balanced pairs. The excellence of balance required, of course, depends on the extent to which balance is relied on to reduce the required thickness of sheath. In view of the fact that very thin sheaths are impracticable for mechanical reasons it appears probable that only very modest requirements as to balance need be imposed on the design of these transformers.

To provide several circuits in new cables for meeting wide-band television needs, another method may be considered, that is, to provide balanced pairs considerably larger in size than ordinary pairs and with the rubber disc form of construction, or other form giving low dielectric losses, but with no shields at all around the individual pairs. Shielding from outside disturbances would be adequately provided by the outside lead sheath of the cable. Crosstalk between different pairs would be the principal concern. If all of the high-frequency pairs were to be used for television transmission, the crosstalk requirements, as already mentioned, would not be severe, so that by careful design the crosstalk could readily be kept within proper bounds—of course, restricting transmission of all wide-frequency bands to a single direction within a single sheath. Such high-frequency balanced pairs might prove suitable for telephone message circuits also. If not, the high-frequency pairs would be restricted to television only, and other pairs, worked at lower frequencies, would be provided for telephone message service.

Summary

It appears feasible under certain conditions to transmit continuous frequency ranges of 1,000,000 cycles or more over conductors in the existing telephone plant. This may some day prove very important, particularly if the art of television develops to the point of calling for such wide frequency range circuits to carry television impulses around the country as sound programs are now carried.

For new construction, the balanced type of circuit, as well as the unbalanced coaxial circuit, offers many interesting possibilities.

A Survey of Magnetic Materials in Relation to Structure*

By W. C. ELLIS and EARLE E. SCHUMACHER

The structure dependence of magnetic characteristics of the more important magnetic materials is discussed. The natural grouping into soft magnetic materials and hard or permanent magnet materials is used. In the first group are: magnetic iron, silicon steel and the alloys of iron, nickel and cobalt. The latter group includes the well-known carbon and alloy steels, and also the newly developed precipitation hardening α-solid solution types. Examples of the latter are the iron-nickel-aluminum and the iron-cobalt-molybdenum alloys.

In discussing the properties, the relationship to structure is emphasized. With the soft magnetic materials purification and the development of proper structural configuration are important. The special metallurgical control required to produce certain magnetic materials in dust form is described. The generalization is made that permanent magnet characteristics are due to precipitation effects resulting from decomposition of supersaturated solid solutions.

Introduction

MAGNETIC materials may be classified according to properties into two groups: (1) soft magnetic materials, and (2) hard or permanent magnet materials. The differences in properties of the two groups are illustrated in Fig. 1, which shows typical magnetization cycles for the two types. The dashed line is a normal magnetization curve while the closed full line is known as the hysteresis loop.

A soft magnetic material is, in general, characterized by a steeply ascending magnetization curve; that is, large values of flux density are produced by small magnetizing forces. For certain applications where the flux density is low, the initial portion of the curve is important. For intermediate flux density applications the steeply ascending portion is of paramount interest while for higher densities, the upper portion is of prime importance.

Another term which is widely used in discussing soft magnetic materials is permeability. The permeability at a given flux density is the slope of the straight line joining that point on the magnetization curve with the origin. Algebraically, it is the ratio of the flux density, B, to the magnetizing force, H. It varies with the flux density, and for soft magnetic materials at low magnetizing forces, is a large quantity. The permeabilities usually reported in discussing magnetic properties are the initial permeability, μ_0, and the maximum permeability, $\mu_{max.}$. The initial permeability is the initial slope of the magnetization curve. The

* Published in *Metals and Alloys*, first part in December 1934 issue, second and concluding part in January 1935 issue.

maximum permeability is the maximum value that the ratio B/H attains.

Hysteresis is also of importance, especially in alternating current applications. The phenomenon of magnetic hysteresis results in a loss of energy in a magnetic material when the material is carried through a magnetization cycle. The loss of energy is proportional to the area of the hysteresis loop. In soft magnetic materials, the aim, in practically all cases, is to keep this loss a minimum.

A hard or permanent magnet material is characterized by a gradually ascending magnetization curve; that is, the material at all magnet-

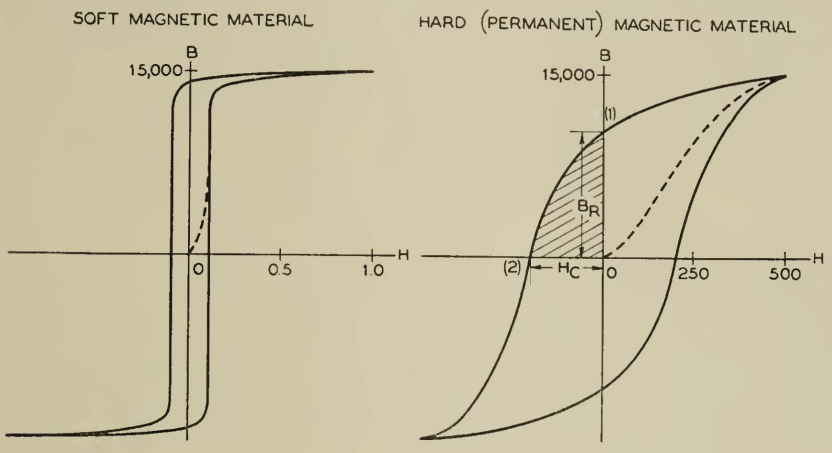


Fig. 1—Schematic representation comparing characteristics of soft and permanent magnet materials.

izing forces has a relatively low permeability. To attain practical saturation is difficult, requiring in some alloys magnetizing forces of thousands of oersteds. The important properties are associated with the demagnetization curve, that portion of the hysteresis loop between points (1) and (2) in Fig. 1. In general the hysteresis loop in its entirety is of little importance. The constants usually determined in permanent magnet investigations are the residual induction, B_r, and the coercive force, H_c. The product of these two quantities ($B_r H_c$), which is roughly proportional to $(BH)_{max.}$ on the demagnetization curve, has been considered by some [1] as a quality index for permanent magnet materials.

Summarizing the general properties of soft and of hard magnetic materials, in the former, high maximum permeability is associated with low coercive force and a hysteresis loop of small area. In the latter, a low permeability is associated with a high coercive force and a large hysteresis loop. The range of properties which can be obtained in magnetic materials by alloying and other metallurgical control is most remarkable. Permeabilities from 1 to over 600,000 and coercive forces from 0.012 to 600 oersteds, are attainable.

In manufacturing soft magnetic materials, the metallurgist strives constantly to eliminate those chemical elements (impurities) which broaden the hysteresis loop, while in the case of the permanent magnet materials he intentionally adds certain of the same, or other elements, and resorts to heat treatments which broaden the loop. Some details of these procedures with well known materials, and also with some more recently developed, are described in the following pages.

Soft Magnetic Materials

Magnetic Iron

Of the ferromagnetic elements iron, nickel, and cobalt, iron, at the present time, is the only one of industrial importance as a soft magnetic material in the unalloyed condition. Throughout the last half century, a continual improvement in the magnetic quality of iron has been effected until at present, laboratory samples have been prepared with permeabilities [2, 11] of over 200,000. Careful analysis of the data shows that this improvement has paralleled closely increases in the purity of the iron.

Since the production of iron of high magnetic quality is of considerable interest at this time, it seems worth while to examine, in some detail, the methods which have been used for producing high purity iron. Methods which have been developed for the purification of iron, have, in each instance, removed from the iron, elements (impurities) which are harmful to easy magnetization.

The elements which enter into iron as impurities may be grouped, structurally, into two classes: substitutional elements and interstitial elements. The substitutional elements exist in the iron structure at lattice points; that is, they substitute for iron atoms in the lattice shown in Fig. 2. These elements include most of the metals, for example: nickel, cobalt, manganese and silicon. The interstitial elements enter into the iron lattice at intermediate points; that is, they take positions in the structure between the iron atoms. Elements included in this classification are carbon, oxygen, and nitrogen. It is these latter elements which, although present in small quantities, cause severe

strains in the lattice, and are believed, on this account, to be particularly harmful to the magnetic softness.

The solubilities of the interstitial elements at room temperature have been determined by a number of investigators. The most reliable values, which, because of the difficulties involved, probably should be considered tentative, place the room temperature solubility of carbon at approximately 0.008 per cent;[3] the solubility of oxygen at 0.01 per cent;[4] of nitrogen at 0.015 per cent;[5] and of sulfur at 0.015 per cent.[6]

The solubilities of the interstitial elements vary with the temperature, increasing, in general, at higher temperatures. Varying solubility provides the possibility of precipitation hardening, which, in the case of iron, results in magnetic hardening. It is this latter type of

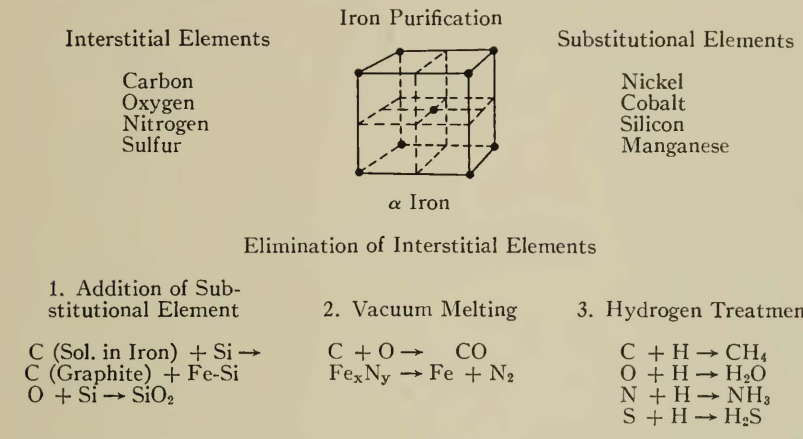


Fig. 2—Schematic representation of methods of purifying iron for magnetic purposes. The lattice structure shown is that for α-iron. The iron atoms are located at the cube corners and at the center and are represented by the circles shown.

hardening to which the magnetic aging of iron is attributed. This is possible because the supersaturated phase will in most of these systems precipitate the interstitial element at room temperature in a form which produces severe strains with deleterious effects on magnetic softness. Because of their harmful effects, the purification of iron for soft magnetic purposes has been in the direction of eliminating these interstitial elements. The three methods of purification which have been used are illustrated in the chart shown in Fig. 2.

The earliest method used was the addition in the molten state of a substitutional element which diminishes the harmful effect of interstitial elements in the iron lattice. Silicon* is an element beneficial in

* Messrs. Barrett, Brown, and Hadfield [7] reported the magnetic properties of iron-silicon alloys in 1900. More recent studies have been reported by Yensen.[7]

this respect. Its action may be explained by postulating that its presence decreases the solubility of carbon in iron. The silicon enters the iron lattice as a substitutional element, and as such, is less harmful to the magnetic softness than the interstitial element, carbon, which is removed. Since silicon is also a strong deoxidizer, it will remove oxygen from the iron. These reactions do not completely eliminate the carbon and oxygen, and consequently, although a noticeable increase in magnetic softness is obtained, the high values of permeability achieved with the more efficient methods of purification are not

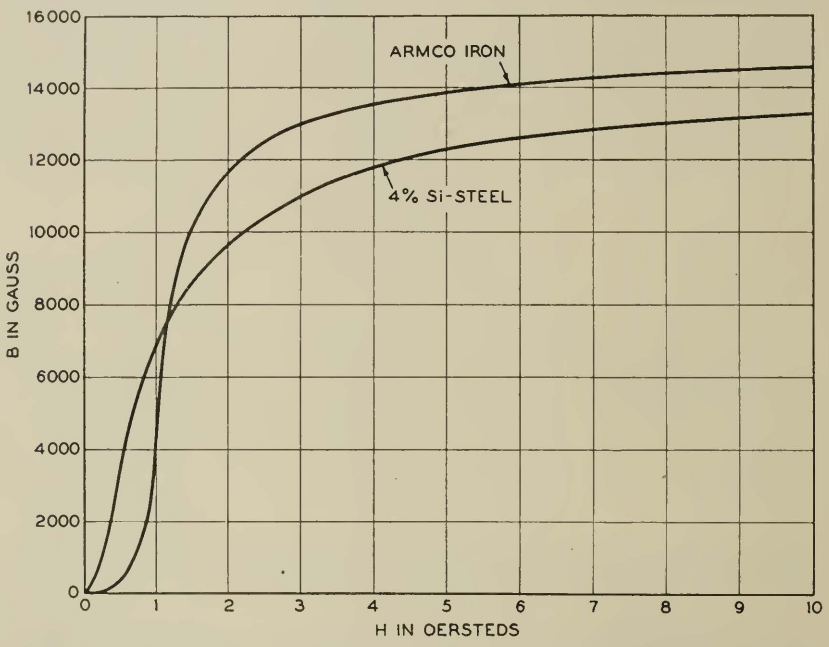


Fig. 3—A comparison of the magnetization curve of commercial 4 per cent silicon steel[8] with that of Armco iron[17] showing the superiority of the former at low induction densities.

reached. The improvement attained by this treatment is illustrated in Fig. 3 by comparison of typical magnetization curves for armco iron [17] and for a commercial 4 per cent silicon steel.[8]

The second method of producing a purer iron consists of eliminating some of the harmful elements by melting in a high vacuum.[9] Under these conditions, the residual carbon and oxygen in the iron will combine to form carbon monoxide, which is removed by the exhausting pumps. At the same time, due to the dissociation of the nitrides at the high temperature, the nitrogen is similarly removed. Again, the

reactions do not go to completion, and residual impurities remain. Another obvious disadvantage is the fact that it would be only accidental if oxygen and carbon were present in the iron in the right proportions to eliminate one another. If the oxygen and carbon contents of the raw materials are known, adjustments can theoretically be made by adding either carbon or iron oxide, but if such a process is in commercial operation, the writers are not aware of it.

Yensen [9] reports a laboratory experiment in which this method of purification was used. A number of samples were prepared in high vacuum from electrolytic iron. To successive samples, increasing amounts of carbon were added to a maximum of 0.5 per cent. The maximum permeability increased with additions of carbon from approximately 25,000 for the sample with no addition to a maximum of 61,000 for a 0.06 per cent addition. From this value, the permeability dropped rapidly with further additions of carbon. The maximum value of 61,000 is explained as due to the fortuitous coincidence that the correct amounts of carbon and oxygen were present for optimum elimination.

Vacuum melting and vacuum purification present some interesting possibilities in connection with the preparation of magnetic materials. Commercial vacuum melting has been developed in Germany at the Heraeus plant,[10] where furnaces with capacities of 5 tons are in operation. It is stated that capacities of 20 to 30 tons can be achieved without large departures from the present designs. Vacuum melted metal is mechanically softer and works more readily than metal prepared by ordinary melting procedures. For magnetic alloys, the combination of vacuum melting and casting provides facilities for preparing material without contamination by the atmosphere. There seems to be no reason why one cannot go further and actually carry on refining operations in the furnace. The future of vacuum melting and casting appears extremely rich in possibilities, particularly for producing magnetic alloys where high purity is a primary consideration.

The third method of purification of iron is by treatment with an element which reacts with the interstitial elements to form gases which are removed. Hydrogen has been used for this purpose by Cioffi,[11] and the method has been described recently. The treatment is carried out at high temperatures between 1300° C. and the melting point of iron. Cioffi states that ordinary amounts of carbon, oxygen, nitrogen, sulfur and phosphorus are reduced to very small quantities by hydrogen treatment. The excess hydrogen is either liberated as the metal cools, or, if it remains in the metal, is without harmful effects on the magnetic characteristics. A value of maximum permeability of 280,000

has been reported [11] for iron prepared in this manner. In Fig. 4 is reproduced a curve from the unpublished work of Cioffi showing the enormous improvement in magnetic softness achieved by the hydrogen treatment.

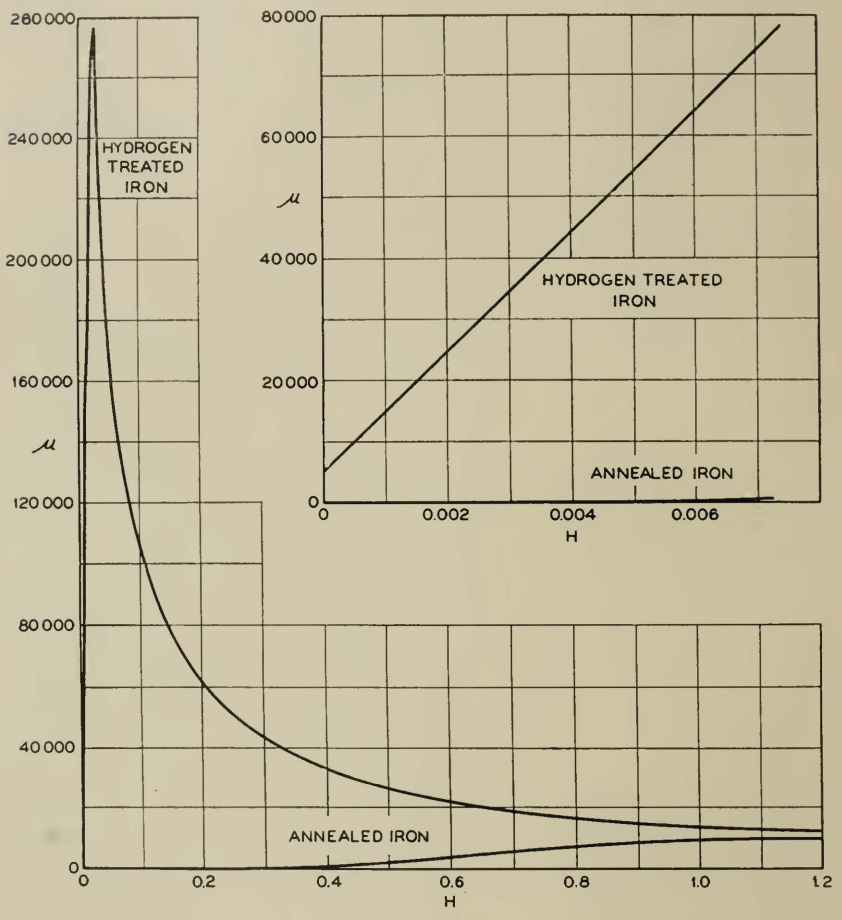


Fig. 4—The permeability of a laboratory sample of hydrogen treated iron compared with that of untreated iron—Cioffi[11].

The mechanism of the purification is outlined by the reactions in Fig. 2. The carbon is eliminated as a hydrocarbon gas, probably methane; the oxygen as water vapor; the nitrogen as ammonia or as nitrogen gas due to thermal dissociation of nitrides; the sulfur as hydrogen sulfide; and the phosphorus probably as phosphine. The most probable explanation to account for the purification is that, at the

high temperatures, the impurities continually diffuse to the surface of the sample where the reaction with hydrogen occurs. Another possibility is that the reactions occur within the body of the sample, in which case the hydrogen must diffuse in and the reaction products out. The first possibility appears the more promising in that the diffusion of reaction products through the iron samples would be difficult, since, in general, the products are rather large molecules.

Another type of high purity iron which is being used for magnetic purposes is carbonyl iron [12] manufactured by the I. G. Farbenindustrie in Germany. It is prepared by first forming iron carbonyl under suitable conditions of temperature and pressure, and subsequently decomposing the carbonyl. For the best quality, the material is further purified by a low temperature hydrogen treatment. This iron is in the form of spherical particles, a few microns in diameter.

One use for this material is in cores of high frequency inductance coils. The cores of such coils are prepared by pressing the insulated magnetic particles into the desired core shape.

The material is also used in preparing iron sheet [13] and as an alloying constituent for the manufacture of iron-nickel alloy sheet.[14] The process of alloying consists of pressing the powders and sintering at a high temperature, followed by working and annealing in hydrogen or vacuum. Typical permeability values for iron sheet prepared in this manner are: $\mu_0 = 2000–3000$; $\mu_{max.} = 15{,}000$. The values are decidedly lower than those for a laboratory sample of hydrogen-treated iron shown in Fig. 4.

Iron Nickel Alloys

No alloy system has been more fruitful in yielding interesting and useful magnetic alloys than that of iron and nickel. In this system, the three regions marked with arrows on the constitutional diagram in Fig. 5 are of principal importance.

In the proximity of 25 per cent of nickel, which is the region of retarded phase change, alloys can be obtained at room temperature in the non-ferromagnetic state. How this is accomplished is evident from a consideration of the constitutional diagram. The γ solid solution above the magnetic transformation is non-ferromagnetic while the α solid solution is ferromagnetic. When alloys in the region of 25 per cent nickel are cooled, the transformation to the ferromagnetic state is not completed until temperatures below room temperature are reached. If the alloy is cooled until the ferromagnetic structure is obtained, the transformation to the non-ferromagnetic state does not occur on heating until the temperature has reached approximately 600° C.

Elmen [17] has investigated alloys containing more than 30 per cent of nickel which are characterized by great magnetic softness and have been called "permalloys," [17] the name suggesting high permeability. These alloys in the region of 40–55 per cent nickel find industrial application as soft magnetic materials. They are used in telephone and radio transformers and telephone relays. An alloy typical of this region is the one containing 45 per cent of nickel. A curve [15] showing

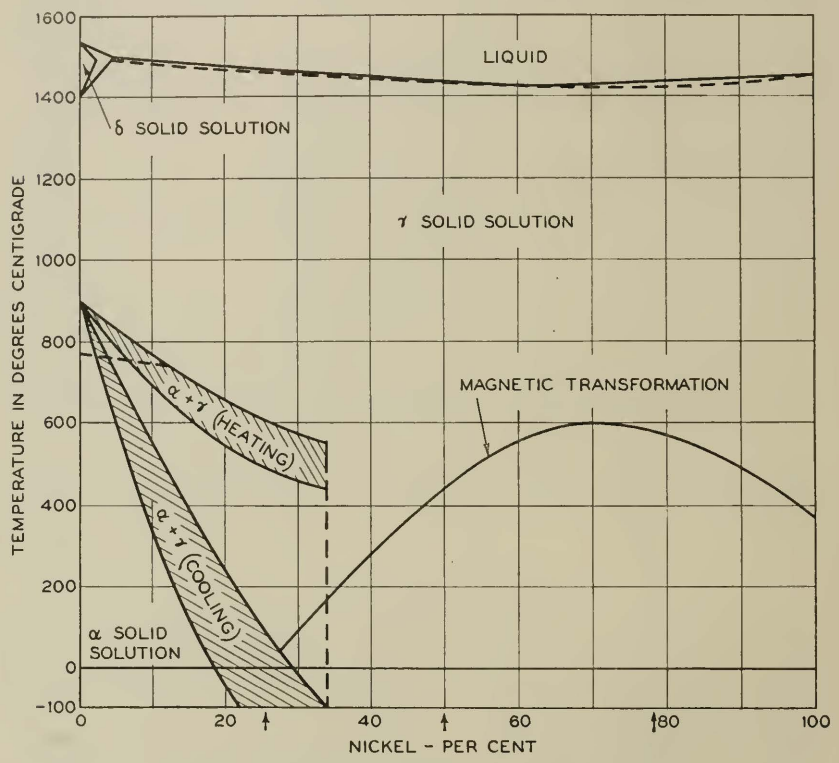


Fig. 5—The constitutional diagram of iron and nickel.

the permeability of this alloy is plotted in Fig. 6. From the data given on this curve, the material exhibits an initial permeability of 2000 and a maximum permeability of 16,000.

Another alloy in this region is the one containing 50 per cent of nickel which has been investigated by Yensen. He has applied the name "Hypernik" [16] to it. Yensen, by giving this material a special heat treatment, obtained an initial permeability of 4000 and maximum permeability of 70,000.

Arnold and Elmen [17] have investigated in great detail the alloys in the region of 70 to 80 per cent nickel. These alloys are used where extremely high initial permeability is required. One especially interesting use is in the loading of submarine telegraph cable. An alloy typical of this region is the one containing 78.5 per cent of nickel. Magnetization curves for this alloy after two different heat-treatments are compared with the curve for annealed armco iron in Fig. 7. In the

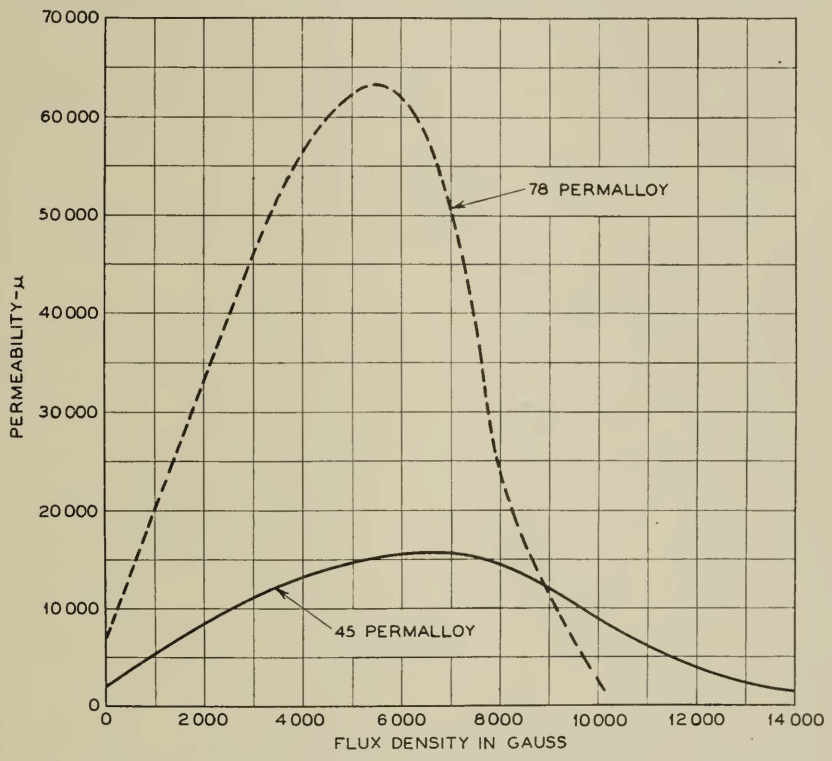


Fig. 6—The permeability of 45 permalloy compared with that of 78 permalloy—Pettit[15].

air-quenched condition, in which the alloy is magnetically the softer, typical values for the initial and maximum permeability are 7000 and 65,000, respectively.

In comparing the properties of 45 per cent nickel and 78 per cent nickel alloys with reference to Fig. 6, the initial permeability is higher for the higher nickel alloy. Therefore, for very low magnetizing forces, the higher nickel alloy is characterized by higher flux densities. For higher magnetizing forces, the 45 per cent alloy has more attractive

properties in that the flux density is higher. In addition, the electrical resistivity of the alloy is approximately three times that of the 78 per cent alloy. The electrical resistivity is important in alternating current applications since the eddy current loss in sheet of the same physical dimensions varies inversely with the electrical resistivity of the material.

To the metallurgist the effects of heat treatment in establishing the magnetic properties of iron-nickel alloys possess a great fascination.

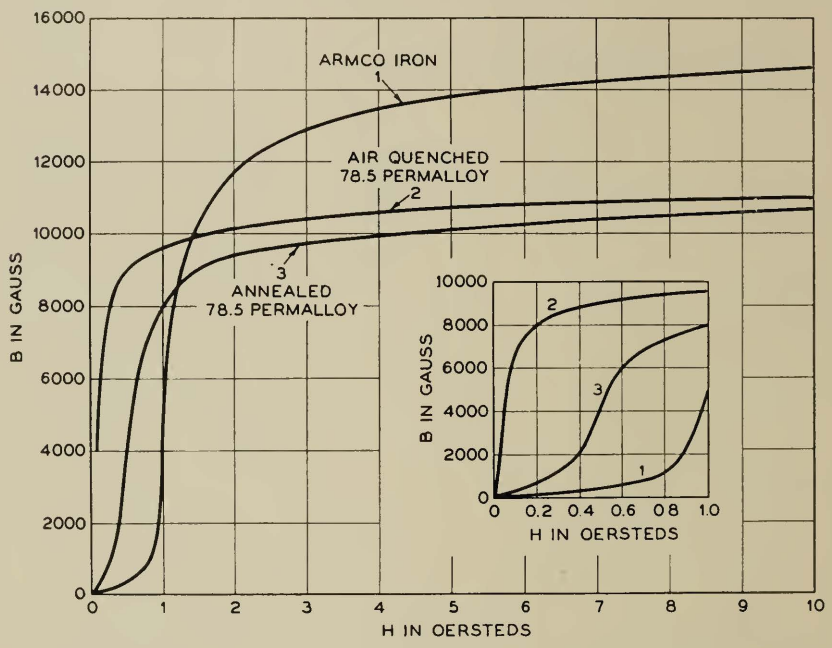


Fig. 7—Magnetization curves for 78.5 permalloy after two heat treatments—Elmen[17].

The effects in question are illustrated by curves shown in Fig. 8. In the region of 78 per cent nickel, the maximum magnetic softness is exhibited after air-quenching, while increases in magnetic hardness are obtained after annealing or baking for extended times at low temperatures. The compositions near 50 per cent nickel are not as much affected by heat treatment.

Elmen [17] attributed the heat treating effect to the separation of the homogeneous phase, stable at high temperatures, into two phases on slow cooling. The two phase structure under these conditions would be less magnetic. More recently in Germany, investigators [18] have advanced the analysis by postulating that in the region of 76 per cent nickel, where the ratio of iron to nickel atoms is 1 : 3, the alloy can exist

in two states: one, a solid solution state in which the atoms of iron and nickel are in a disordered distribution; and two, a state in which the iron and nickel atoms are regularly distributed.

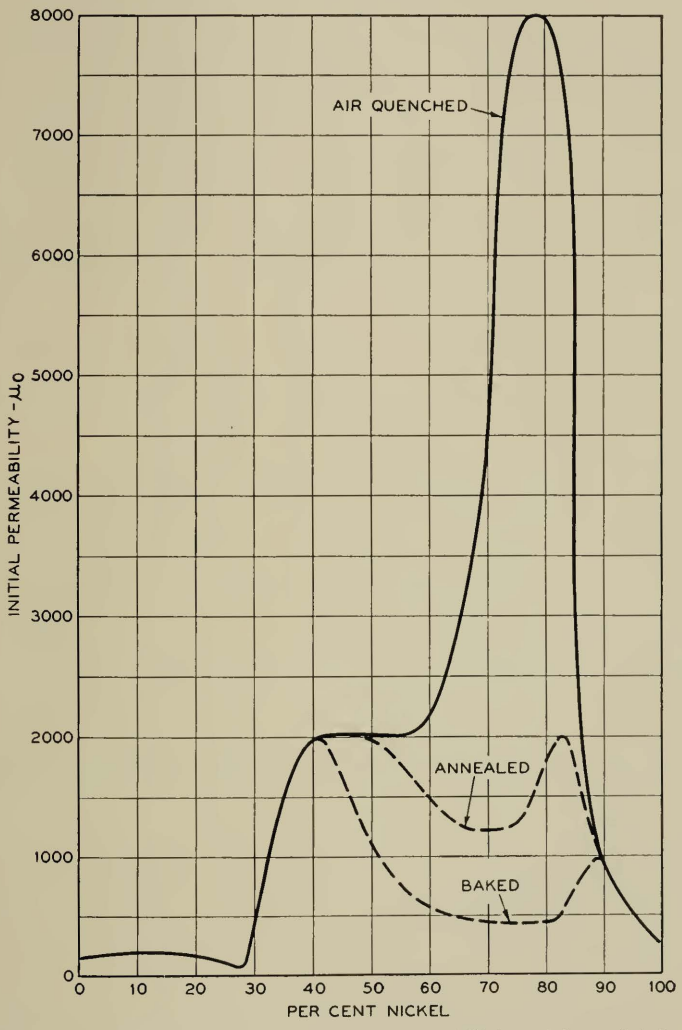


Fig. 8—Effect of heat treatment on initial permeabilities of iron-nickel alloys—Elmen[17].

In the disordered distribution the lattice points are fixed but the atom positions are not; that is, any lattice point may be occupied by either an iron or nickel atom. In the regular or ordered distribution, the lattice points are fixed and also the atom positions. The second

arrangement may be clearer to some if it is considered a "compound," although strictly speaking, the arrangement does not correspond to a compound in the chemical sense. The regular arrangement is now ordinarily called a "superlattice" or a "superstructure," and as such, will be designated in this paper. If it is clearer to the reader, he may substitute for the term "superlattice" or "superstructure," the term "compound" without great sacrifice in the sense of the discussion.

With this picture in mind, the effect of heat treatment is immediately evident. Rapid cooling prevents the formation of the "superlattice" arrangement, which occurs somewhere in the region of 500° C., and a disordered distribution of atoms is retained. This is the distribution possessing the higher magnetic permeabilities. Slow cooling or baking promotes the transformation to a superlattice of lower magnetic quality.

Heat treatment effects are much less in the alloys in the region of 50 per cent nickel, and in fact would not be expected, since a face centered cubic binary alloy with the atomic ratio of 1 : 1 would not be expected to have a special structure (superstructure).

The hypothesis of superlattice formation in the 75 per cent nickel region is supported further by the effect of heat treatment on other physical characteristics, for example the electrical resistivity and the tensile strength. That such special structures are formed in the solid states is well established from detailed studies of the copper-gold system.[19] Unfortunately, in the nickel-iron system where the nickel and iron atoms are so near in atomic number, the detection of superstructure by x-ray methods appears, at the present time, a difficult task.

The effect of heat treatment and the general magnetic softness of iron-nickel alloys in the region of 75 per cent nickel may be explained by another hypothesis which returns for its basis to the considerations set forth in the section on high purity iron. The improvement in magnetic softness may be attributed, first, to the effect of nickel in decreasing the residual quantities of interstitial elements originally present in iron. Whether or not such an effect is present awaits a careful investigation of the interstitial element content of iron-nickel alloys. Secondly, assuming the interstitial elements the responsible factors, the effect [20] of heat treatment on magnetic quality can be explained in that the quench (rapid cooling) retains the residual elements in solution, while slow cooling permits precipitation in a form more deleterious to magnetic quality.

The two hypotheses possess interesting possibilities, but it appears that further investigation will be required before the structural relationships are definitely established.

The permalloys in the region of 75 per cent nickel have been further modified by the addition of third elements [17] which are non-ferromagnetic. Molybdenum and chromium are elements of this nature which confer on the alloys certain characteristics of sufficient interest to warrant discussion. The effects of molybdenum and chromium on the magnetic quality are shown in Fig. 9, which indicates that higher initial permeabilities may be expected with the ternary alloys. The addition of a third element increases the electrical resistivity of the alloy. This is of importance in alternating current applications where eddy current losses are a consideration. In commercial manufacture the third element is particularly valuable, since it decreases the sensi-

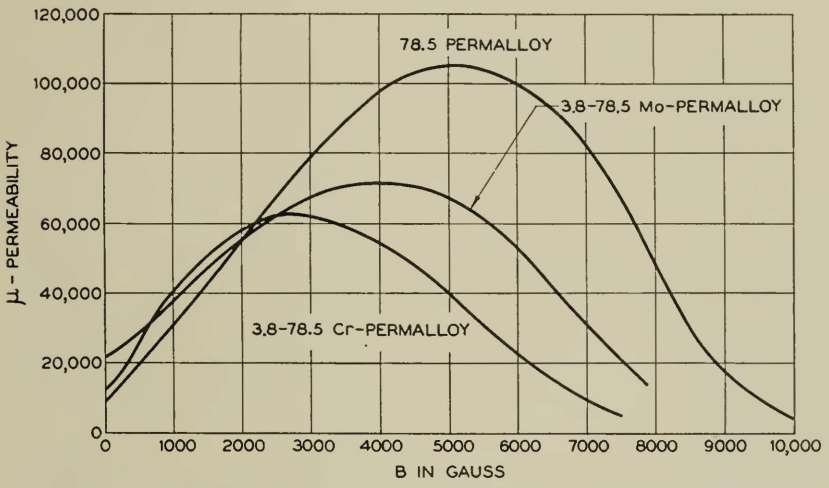


Fig. 9—Permeability curves obtained for laboratories' samples comparing the properties of chromium permalloy and molybdenum permalloy with those of standard permalloy—Elmen [17].

tivity of the magnetic properties of the alloy to the rate of cooling after annealing and, in fact, permits the attaining of high permeabilities with slow cooling. The significance of this from a structural viewpoint is that the third element keeps the iron and nickel atoms in a disordered distribution.

Iron-nickel alloys have a limited field of application in the form of thin tape in the hard rolled condition. A representative material is one containing from 30 to 70 per cent nickel and, in addition, a few per cent of aluminum or copper, which in the hard rolled condition has been called "isoperm." [21] The material is reported to have a low permeability of 50 to 75, which is fairly constant over a range of magnetizing forces of 0–10 oersteds. Laminated cores of this material have prop-

erties similar, but somewhat inferior to, those of magnetic dust cores [27] which are described in a later section of the paper.

Iron-Cobalt Alloys

Iron-cobalt alloys in the region of 50 per cent cobalt are of importance for magnetic purposes because of the high values of flux densities obtained with medium magnetizing forces. The desirable magnetic properties of the alloy containing 50 per cent cobalt were first reported by one [22] of the authors. A very complete investigation of alloys over the whole range of cobalt contents has been described by Elmen [17] from whose paper the curves shown in Fig. 10 are taken. With

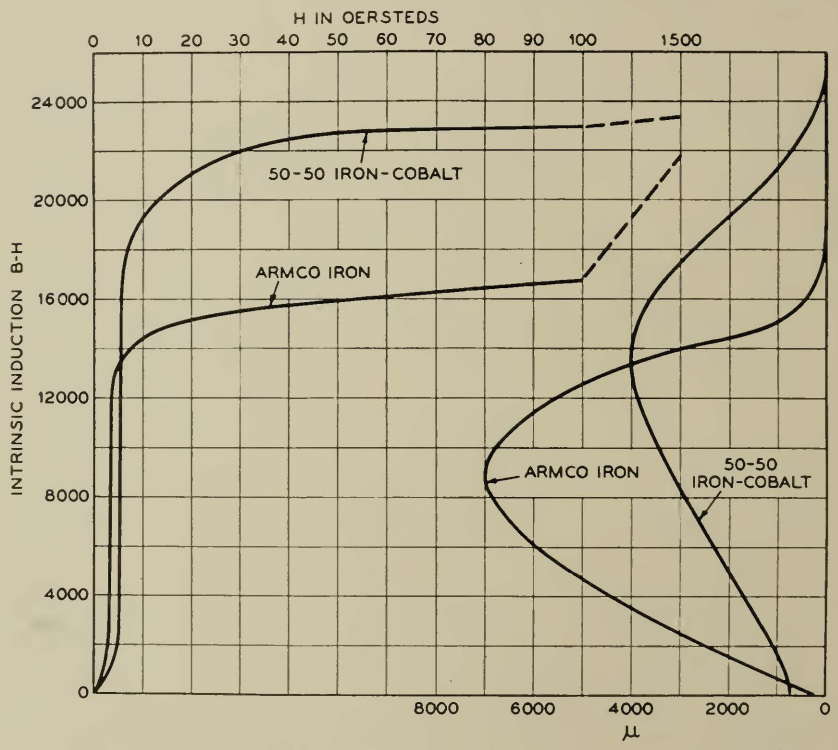


Fig. 10—Magnetization curve and permeability curve for 50–50 iron-cobalt alloy—Elmen[17].

magnetizing forces from 5 to over 1500 oersteds the intrinsic induction is greater than that obtained with iron.

Because of anomalies in properties of the alloys in the region of 50 per cent cobalt, the structural features of the alloys become of interest. In addition to the maximum in intrinsic induction, the electrical con-

ductivity in this region at room temperature is greater than that of either of the components. The anomalous behavior was attributed several years ago by one of the authors,[22] not to compound formation, but to a "peculiar solid solution." Kussman, Scharnow and Schulze [22] in a recent paper attribute the effects to the formation of superstructure in the body centered lattice which exists at this point. This, in a sense, is an extension of the earlier explanation. Here again, it appears that further work is warranted and is required before the structure is completely understood.

In connection with the metallurgy of the 50–50 iron-cobalt alloy, there are some interesting considerations. The binary alloy can be worked hot, but is extremely brittle when cold. This imposes definite limitations where thin sheet produced by cold rolling is required. The limitation can be overcome by adding a few percent of vanadium [23] to the composition. This alloy can be worked hot, and after a quench from a high temperature, also can be cold-rolled to thin sheet. The magnetic characteristics are not greatly affected by the vanadium, provided the addition is small.

The structural changes resulting from the additions of vanadium are not clearly understood. It is known that vanadium retards the transformation which occurs in these alloys at approximately 900° C. and therein lies a possible explanation for its action. The high temperature modification is malleable as shown by satisfactory hot working properties. By quenching, sufficient of this modification may be preserved at room temperature, when vanadium is present, to permit cold rolling.

Iron-Cobalt-Nickel Alloys

In the ternary system, iron-cobalt-nickel, is a region in which the alloys exhibit an unusual and useful property, namely, constant permeability in low fields. Because of this characteristic, alloys in this region have been named perminvars.[24] A typical alloy with marked perminvar characteristics contains 45 per cent nickel, 25 per cent cobalt and 30 per cent iron. The effects of heat treatment are illustrated in Fig. 11. The effect on the magnetization curve of air quenching, annealing, and baking at a low temperature is shown for this typical alloy. The extension of constant permeability to higher magnetizing forces by a low temperature bake is clearly demonstrated.

The constancy of permeability in perminvar has been explained in one hypothesis as due to the presence of two constituents in the alloy, one a soft magnetic material, the other a hard magnetic material. The experimental evidence for this hypothesis is derived largely from the constricted hysteresis loops which are obtained for these alloys after a

baking treatment. A similar type of constricted loop can be obtained in a magnetic sample which is composed of thin parallel sheets of a soft and of a hard magnetic material magnetized longitudinally.

It may be that the two constituents result from a process similar to that described for the iron-nickel alloys in the region of 75 per cent

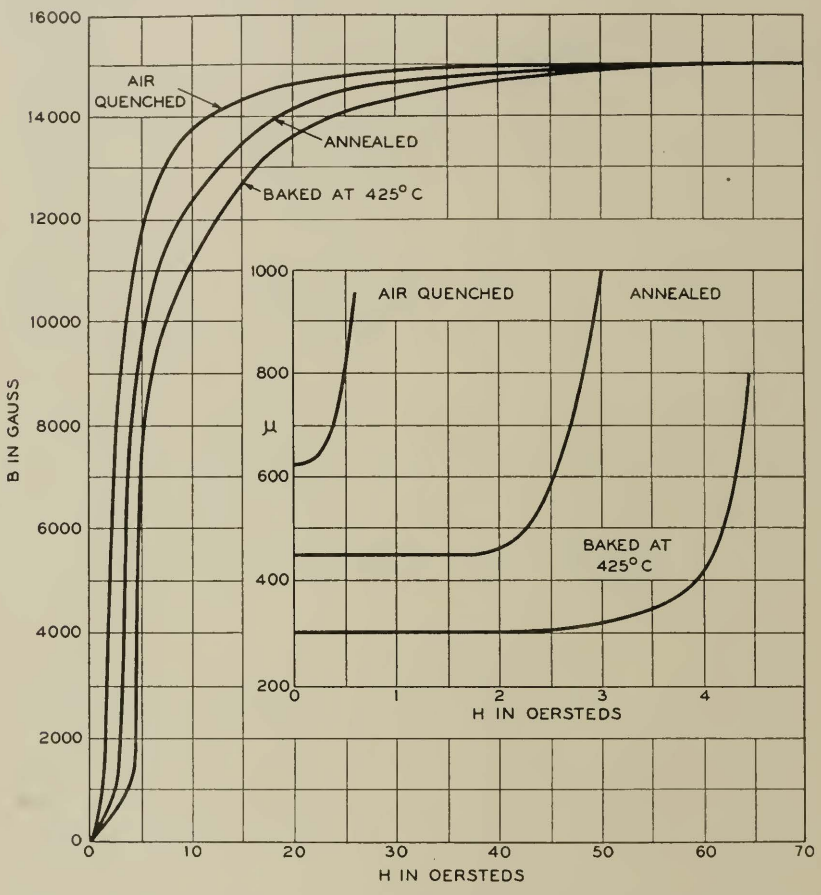


Fig. 11—Magnetization curves for a perminvar containing 45 per cent nickel, 25 per cent cobalt and 30 per cent iron after several heat treatments—Elmen [17].

nickel. In line with this, the baking treatment may form a superstructure [24] which is magnetically hard, but to a greater degree, than the one in the binary iron-nickel alloys. It is probable that the similar effects of heat treatments in the binary and ternary alloys are due to similar structural transformations.

Heat Treatment of Magnetic Material in Magnetic Field [25]

Recently it has been reported that radical changes have been produced in the magnetic properties of some ferro-magnetic alloys by heat treatment in a magnetic field. For example, the maximum permeability of the permalloy containing 65 per cent of nickel has been in-

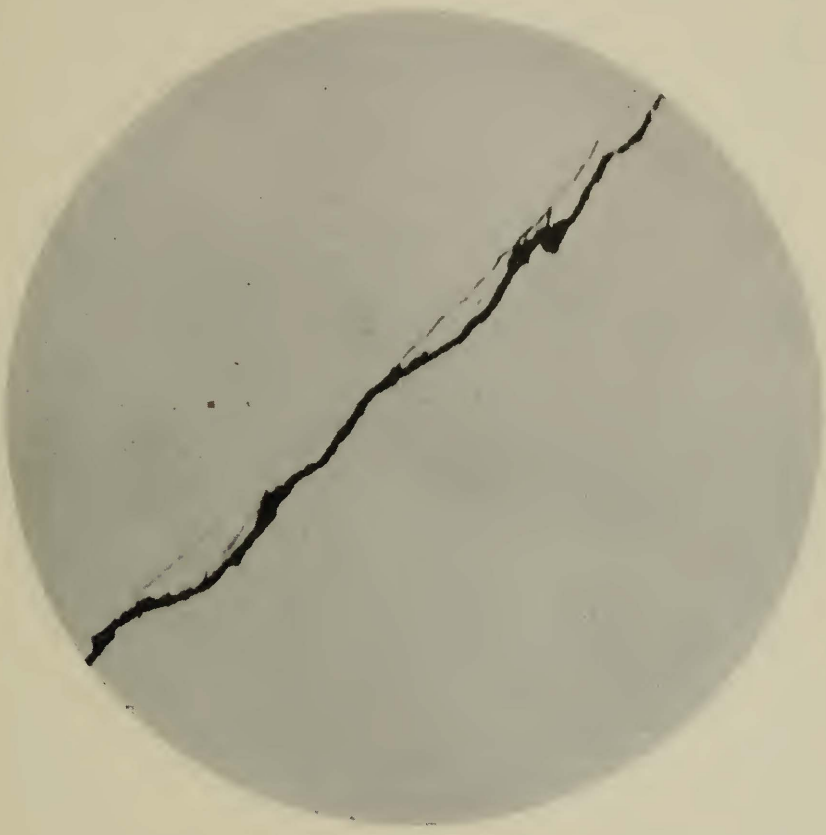


Fig. 12—Structure of an iron-nickel alloy, as cast, containing 80 per cent nickel and 0.030 per cent sulphur. The sulfide constituent exists as a brittle film surrounding the crystallites. Mag. 2000 ×.

creased from 10,000 to 250,000 by heating to 700° C. and cooling slowly in a field of 15 oersteds. The maximum permeability has been further increased to 600,000 by so treating a specimen which had previously been treated in hydrogen at 1400° C. This value of maximum permeability is the highest which has been reported for any ferromagnetic material.

In the iron-cobalt-nickel alloys, high values of maximum permeability are also generally obtained by such a heat treatment, provided the magnetic change point is higher than about 500° C. Another feature in connection with this special heat treatment is that the heat-treated specimens have hysteresis loops which tend to have sharp corners and vertical sides.

An interesting explanation for the effect of heat treatment in a magnetic field is given by Bozorth.[25] According to modern magnetic theory, when a specimen is cooled through the Curie point in the absence of an applied field, small regions or domains are magnetized to saturation in definite directions. The specimen in its entirety exhibits no external magnetic effects since the vector sum of the individual effects would be nearly zero and not detectable. The magnetization of the domains produces internal strains in the material due to lengthening in one direction and contraction at right angles (magnetostriction). By the application of an external magnetic field in the temperature region near to, but below, the Curie point, (1) the magnetic domains are oriented nearly in the direction of the applied field, and (2) the internal strains which were introduced by the local magnetization at the Curie point are removed by plastic flow of the surrounding domains. For this to occur, the temperature of Curie point must be higher than that at which plastic flow begins. Subsequent application of an external magnetic field at room temperature in the direction of, or opposite to, the direction of the magnetization at the higher temperature introduces no strains in the material tending to decrease the ease of magnetization. For that reason relatively higher permeabilities are obtained than with material heat treated in the absence of a field, in which case opposing strains are present.

The theory also offers another explanation for the effects of heat treatment on the permalloys in the region of 75 per cent nickel. This explanation will not be discussed further here except to say that it is concerned with the same magnetostrictive strains as have been mentioned above.

The fundamental idea of the theory proposed to account for the effects of heat treatment in a magnetic field, and the general effects of heat treatment on certain ferromagnetic materials, is that the magnetostriction associated with the local magnetization in the domains is sufficient to cause plastic flow. Experimental data indicate that the magnetostrictive stress is sufficient to cause such plastic flow at about the temperature which is known to be critical for the heat treatment of the permalloys.

Heusler Alloys

One of the most interesting groups of magnetic materials comprises the Heusler alloys [26] which are ferromagnetic although composed of non-ferromagnetic elements. A representative alloy of this type is one containing copper, manganese, and aluminum. The magnetic properties are associated with the composition Cu_2MnAl, since the magnetic intensity of the alloys varies in proportion to the amount of this phase. These properties also depend upon structure.[26] The ferromagnetic condition is obtained by quenching the alloy from approximately 800° C. This treatment produces a body centered cubic structure with a face centered superlattice. Manganese is essential to the composition; copper may be replaced by silver; [26] aluminum by other trivalent or tetravalent elements—for example, tin.

Although possessing no properties of commercial utility, the alloys are extremely important in relation to magnetic theory. For those who may be particularly interested, a rather complete list [26] of recent papers on the subject is given in the references.

Summary of Magnetic Characteristics of Sheet Materials

There are now available for magnetic purposes a large number of materials which are suitable for use over different ranges of magnetizing force and at different flux densities. In summarizing properties, the materials are grouped according to their adaptability at different magnetizing forces.

For low forces, where very high permeabilities are required, as in some signal apparatus, 78-permalloy is a suitable material. Molybdenum or chromium permalloy has properties similar to those of 78-permalloy but has higher initial permeability, is more readily heat-treated and is to be preferred for alternating current applications because of the higher resistivity. Hydrogen-treated iron at low magnetizing forces also has properties similar to those of the 78-permalloy.

For small magnetizing forces where a constant permeability and very low hysteresis loss is required, the heat-treated perminvars are available. The flux densities in cores of this type, if constant permeability and low hysteresis loss are retained, must be kept below 1000 gauss.

For apparatus using higher magnetizing forces, iron-nickel alloys in the range of 40 to 55 per cent nickel, and the silicon steels are most suitable. These two materials are widely used in electric transformers.

In the region of fields of 10 to 50 oersteds and higher, the 50–50 iron-cobalt and vanadium-modified alloy are attractive because of the high intrinsic induction of approximately 22,000 gauss. This superiority exists even in fields of over 1000 oersteds, but in the higher

ranges, the saturation value is approached by that of the less expensive material, magnetic iron. In connection with high flux density applications, the permalloys, the perminvars, and modified alloys of this character are of little interest since they are saturated at comparatively low flux densities.

The data given here are not sufficient to select a material for a specific application since many other detailed properties must be considered in connection with each individual problem. The materials which have been described, however, cover, in a general manner, the entire range of magnetic fields.

Magnetic Alloys in Dust Form

For certain purposes where a substantially constant low permeability is desired, and in particular, for high frequency applications where eddy current losses are of consequence, it is desirable to produce the magnetic material in a fine powder, which is subsequently insulated and pressed into the desired core shape. One alloy prepared in this form is the permalloy [27] containing approximately 80 per cent of nickel. This is a material suited for use at audio frequencies, for example in the cores of loading coils.

The alloy may be prepared in powdered form by a number of methods. In one method, which is essentially metallurgical in nature, advantage is taken of the effect of small amounts of added elements.[27] It has been found that the addition of a few thousandths of a per cent of sulfur to the iron-nickel alloy containing approximately 80 per cent of nickel produces a structure which can be hot-rolled to a small section, but which when cold, is exceedingly brittle and can be pulverized to a fine dust. The presence of other elements in small amounts also affects the properties of the alloy. One element, manganese, has an effect opposite to that of sulfur, and if present in sufficiently large amounts, nullifies the action of sulfur by producing a tough and malleable casting.

The structural behavior of sulfur and manganese in permalloy is interesting in explaining the embrittling action. Sulfur exists in the structure as microscopic films of complex sulfides at the crystallite boundaries as illustrated in Fig. 13. These films are brittle, and when they extend over the greater portion of the crystallite surface, produce an interface of weakness permitting easy pulverization. The satisfactory hot-working properties can be explained in that at the high temperature, either the sulfide film is malleable, or dissolves in the iron-nickel solid solution. If manganese is added to an alloy containing sulfur, the sulfide constituent is blackened, loses its continuous

characteristics, and becomes agglomerated as is well known in the case of iron and of nickel. In this form it has a minimum embrittling effect since contact between metallic crystallites exists over the greater proportion of the interface. The structure is illustrated in Fig. 14. If sufficient manganese is present, alloys containing fairly high percentages of sulfur can not only be hot worked, but are malleable when cold.

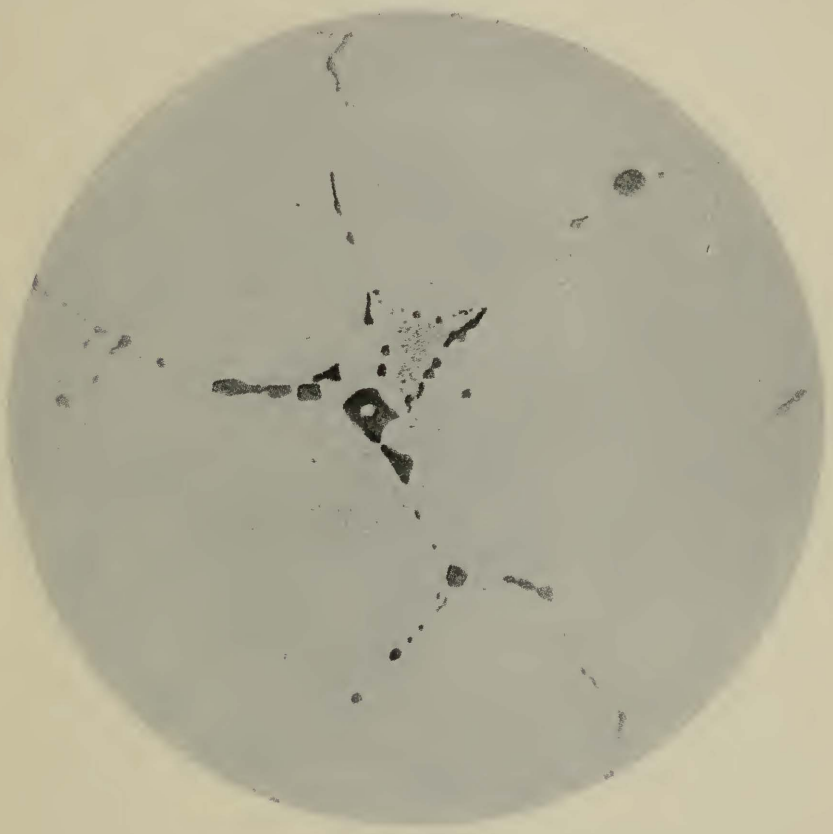


Fig. 13—Structure of an iron-nickel alloy containing 80 per cent nickel to which had been added 0.12 per cent sulphur and 1.05 per cent manganese. The sulfide constituent exists along the crystallite boundaries as grayish rounded inclusions. Mag. 2000 ×.

The production of magnetic dusts is unique as a metallurigcal process since the metallurgist is concerned with adding elements yielding a product which can be pulverized to a fine dust. Generally, the metallurgist exercises his greatest ingenuity to produce ductility and malleability in the end product.

Hard Magnetic Materials (Permanent Magnets)

Permanent magnet materials, as pointed out in the introduction to this paper, are characterized by properties different from those of the soft magnetic alloys. With these materials, a high residual induction and high coercive force are required. Permanent magnets in use

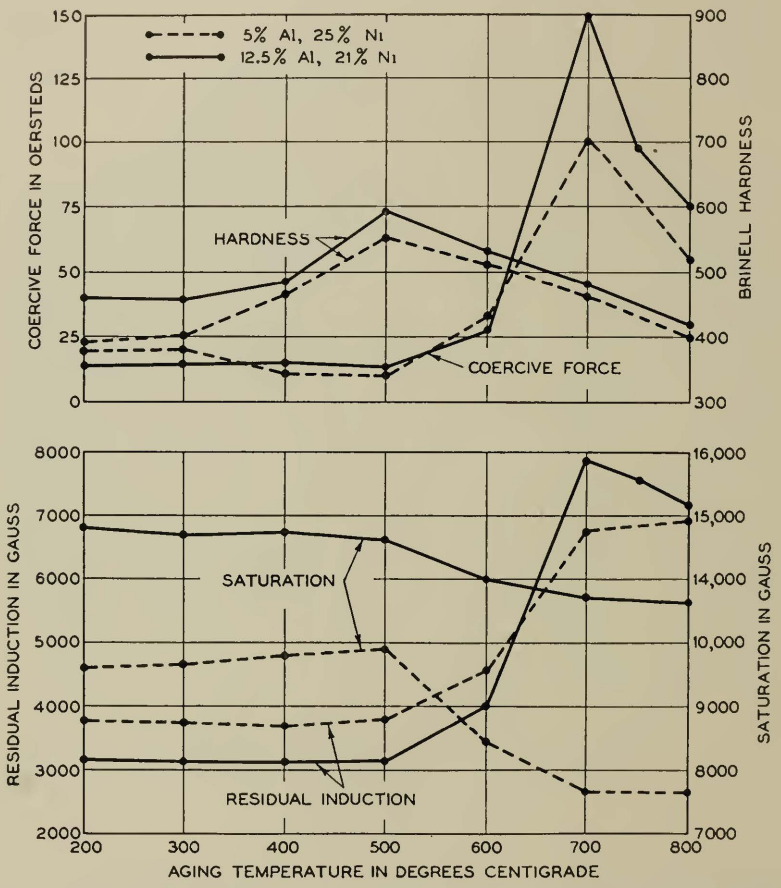


Fig. 14—Effect of aging treatment on typical iron-nickel-aluminum magnet alloys—Köster[31].

commercially at this time are steels. High intrinsic induction is assured by a high proportion of iron in the composition; the magnetic hardness is assured by the addition of carbon, which, when the steel is suitably heat-treated, precipitates as a finely dispersed carbide throughout the matrix. This finely dispersed precipitate keys the

structure and furnishes resistance to change in magnetic condition, a property which is measured as coercive force.

Permanent Magnet Steels

Many types of steel have been developed for magnet applications.[28] The most important commercial ones are carbon-manganese, 1 per cent and higher chrome steels of which 3.5 chrome is typical, tungsten steel, and 35 per cent cobalt steel. Typical magnetic characteristics and required heat treatments are shown in Table I.

TABLE I

THE PROPERTIES OF PERMANENT MAGNET STEELS

Type of Steel	Typical Composition Per Cent	Heat Treatment	Magnetic Characteristics		
			H_c	B_r	$H_c \times B_r \times 10^{-3}$
Carbon-Manganese Steel.....	C—0.60 Mn—0.80 Si—0.20	Quenched 800° C. in water	40	9,000	360
Low Chrome Steel	Cr—0.90 C—0.60	Quenched 800° C. in oil	50	9,500	475
3.5% Chrome Steel	Cr—3.5 C—0.90	Quenched 830° C. in oil	68	9,500	645
Tungsten Steel....	W—5.0 C—0.7	Quenched 845° C. in water	60	10,500	630
35% Cobalt Steel.	Co—35 W—7 Cr—4 C—0.75	Quenched 940° C. in oil	220	9,500	2090

The heat treatment for the permanent magnet steels consists of a quench from a high temperature at which the carbon is in solution. During the period of cooling in the quench, the supersaturated solution precipitates the carbides. This is designated "quench hardening." The quenching treatment varies with the type of steel, but in general, quenching in either oil or water is used. In this connection, it is important that the austenite transformation occur during the quench, since austenite is non-ferromagnetic and its presence will result in low values of residual induction.

The carbon-manganese and 1 per cent chrome steels are the least expensive and enjoy a large use in low cost apparatus where space requirements are liberal. Typical values for coercive force and residual induction for these materials are 40–50 oersteds and 9000 to 9500

gauss. The 1 per cent chrome steel has the advantage over the carbon-manganese steel in that the desirable characteristics are produced by oil quenching while the carbon-manganese steel is ordinarily water-hardened. Water-hardening frequently results in cracked magnets and consequently a higher proportion of rejections.

Of slightly higher quality are the 3.5 per cent chrome and the 5 per cent tungsten steels.* Typical values of coercive force are from 60 to 70 oersteds, and of residual induction from 9500 to 10,500 gauss. In general, tungsten steel in the hardened condition has a higher residual induction than the other magnet steels. Tungsten steel requires water-quenching, while 3.5 chrome steel is oil-quenched. In addition, chrome steel is a somewhat lower cost material. For these reasons, in recent years it has been substituted to some extent in applications where tungsten steel was formerly used.

The best permanent magnet steel in commercial use is the 35 per cent cobalt steel,[29] a complex alloy which contains, in addition to the cobalt, tungsten, chromium and carbon. Typical values of coercive force and residual induction for this material are 220 oersteds and 9500 gauss. Although this material is decidedly superior in properties to the other magnet steels, because of high cost, its use is limited to applications where space curtailment and apparatus requirements eliminate the cheaper steels.

New Developments in Permanent Magnet Alloys

Within the last five years, there have been a number of publications [30, 31] describing new materials which have properties of interest to engineers using permanent magnets. These materials are alloys with no intentional carbon additions, and, hence, are a radical departure in this field. The new magnet alloys solidify as alpha-solid solutions which by suitable heat treatment at a lower temperature decompose precipitating second phases. Contrary to the case with the iron-carbon alloys, the alpha-solid solution undergoes no phase change with decreasing temperature. Consequently the alloys have a coarse grain while the hardened magnet steels have a fine grained structure due to the intermediate phase change.

The permanent magnet qualities, however, result from the same type of metallurgical reaction that occurs in the carbon steels, that is, a precipitation of a second phase which is dispersed throughout the alpha-solid solution. The useful properties are secured by the usual precipitation hardening treatments; either a quench from a high

* A commendable detailed discussion of tungsten magnet steels is given by Gregg in the recent book, Alloys of Iron and Tungsten, p. 212, McGraw Hill, 1934.

temperature during which precipitation occurs, or a quench followed by an aging treatment at a lower temperature to produce the precipitation. Examples of alloys of this type are the iron-nickel-aluminum alloys, which have been described by Mishima,[30] and the iron-cobalt-molybdenum and iron-cobalt-tungsten alloys, which have been described by Seljesater and Rogers [31] and Köster.[31]

Iron-Nickel-Aluminum Alloys

A representative composition of the alloys described by Mishima [30] consists of 65 per cent iron, 25 per cent nickel and 10 per cent aluminum. The composition may be further modified by the addition of manganese, vanadium, cobalt, chromium, tungsten, molybdenum, or copper.

For the simple ternary alloy in the cast condition a coercive force of 240 oersteds and a residual induction of 9600 gauss have been reported by Mishima.[30] By slight modifications in compositions, coercive forces of over 500 oersteds in combination with residual inductions of approximately 9500 gauss are reported. Values for three Mishima alloys, presumably of different composition, have been reported recently by Steinhaus and Kussman [30] and are given below.

Type of Mishima Alloy	Coercive Force Oersteds	Residual Induction Gauss
MK1	660	7,600
MK3	440	9,800
MK5	130	10,800

Köster [30] has investigated the ternary equilibrium conditions for the iron-nickel-aluminum alloys. In the range of compositions of interest for magnet purposes, a surface of solubility varying with the temperature exists. It would be expected, therefore, that these alloys would be amenable to age hardening treatment. This has actually been demonstrated by Köster for the alloys of iron-nickel-aluminum. The curves shown in Fig. 14 are reproduced from his published data and will be recognized as demonstrating typical age hardening phenomena. The optimum aging temperature appears to be 700° C.

The fact that Mishima obtained high permanent magnet quality in specimens in the cast condition can be explained in that the precipitation of the second phase occurs during the simple cooling of the casting. It would be expected, therefore, that the magnet properties obtained would depend upon the casting dimensions and the rate of cooling. If this is true, it might be desirable in some magnet structures to subsequently heat treat the material to obtain uniform and reproducible results.

The fact that precipitation occurs is indicated more completely by examination of the photomicrographs in Figs. 15 and 16 of a typical

iron-nickel-aluminum alloy after quenching, and after quenching followed by aging at 1000° C. for 24 hours. There is no visual evidence of the second phase in the quenched specimen although some precipitation of submicroscopic particles undoubtedly has occurred. In the aged specimen, in Fig. 16, the second phase appears unmistakably.

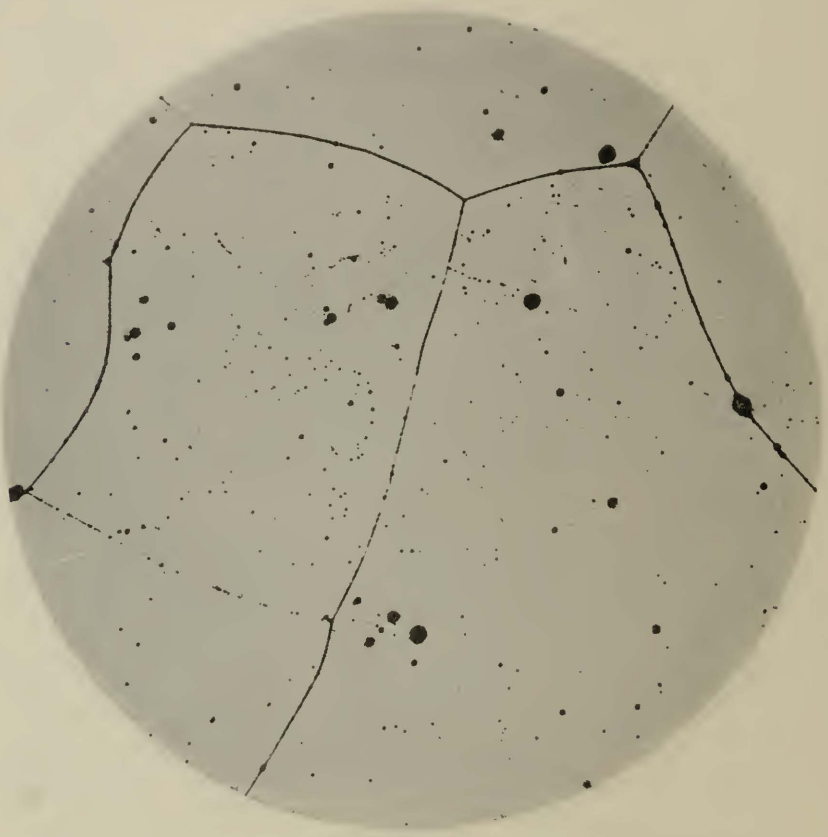


Fig. 15—Typical structure of an iron-nickel-aluminum alloy, containing 25 per cent nickel and 10 per cent aluminum, after quenching from 1200° C. in oil. In this condition the alloy has its optimum permanent magnet quality. Mag. 125 ×.

Iron-Cobalt-Molybdenum and Iron-Cobalt-Tungsten Alloys

The age hardening characteristics of alloys in these systems were established some years ago by Sykes, who developed the alloys for tools. Later, Seljesater and Rogers [31] reported that these alloys possess permanent magnet characteristics. A detailed report on the magnet properties has been made by Köster.[31] Of the two systems, it appears

that appreciably higher coercive forces are available in the one consisting of iron, cobalt and molybdenum. In Fig. 17 are shown the combinations of residual induction and coercive force obtained by Köster in the latter system with variations in molybdenum and cobalt content. Coercive forces of over 200 oersteds, associated with residual inductions of approximately 10,000 gauss, are given.

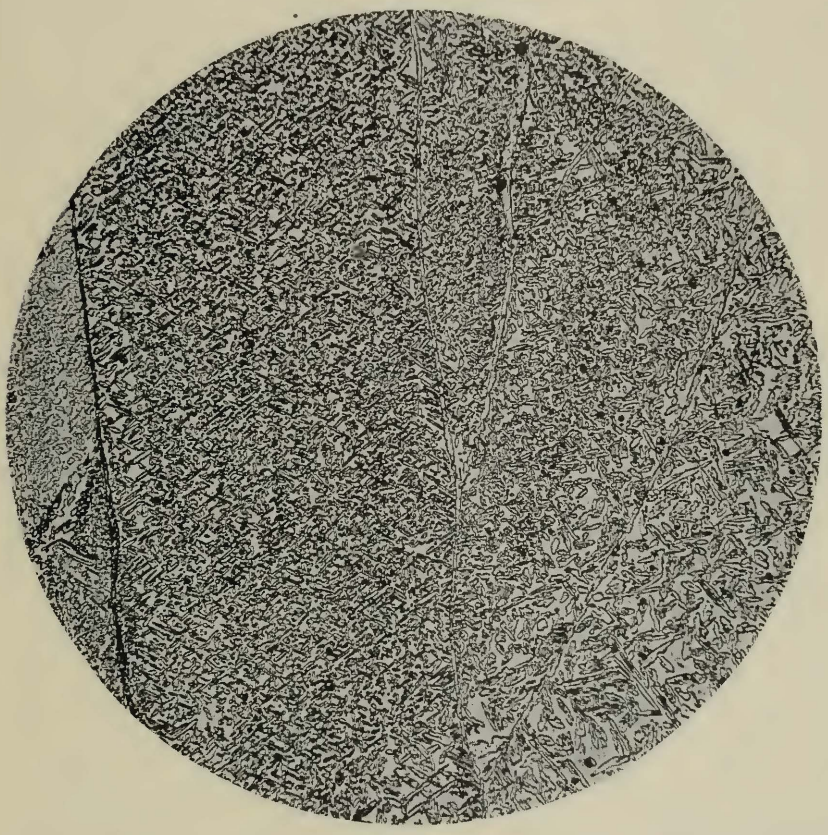


Fig. 16—Structure of the same alloy shown in Fig. 15 after a subsequent aging treatment of 24 hours at 1000° C. This treatment has resulted in the precipitation of a large amount of the second phase. Mag. 125 ×.

The desirable properties of this class of materials are developed by a double heat treatment which consists of a quench from a high temperature followed by aging at a lower temperature. The alloys are precipitation hardening in the same manner as the non-ferrous copper-beryllium alloys or lead-calcium cable sheath alloys. The heat treatment is illustrated in Fig. 18, which shows the effect of aging on the

properties of an alloy containing 15 per cent cobalt and 18 per cent tungsten reproduced from Köster's published data. The alloy was initially quenched from a temperature of 1300° C., and subsequently

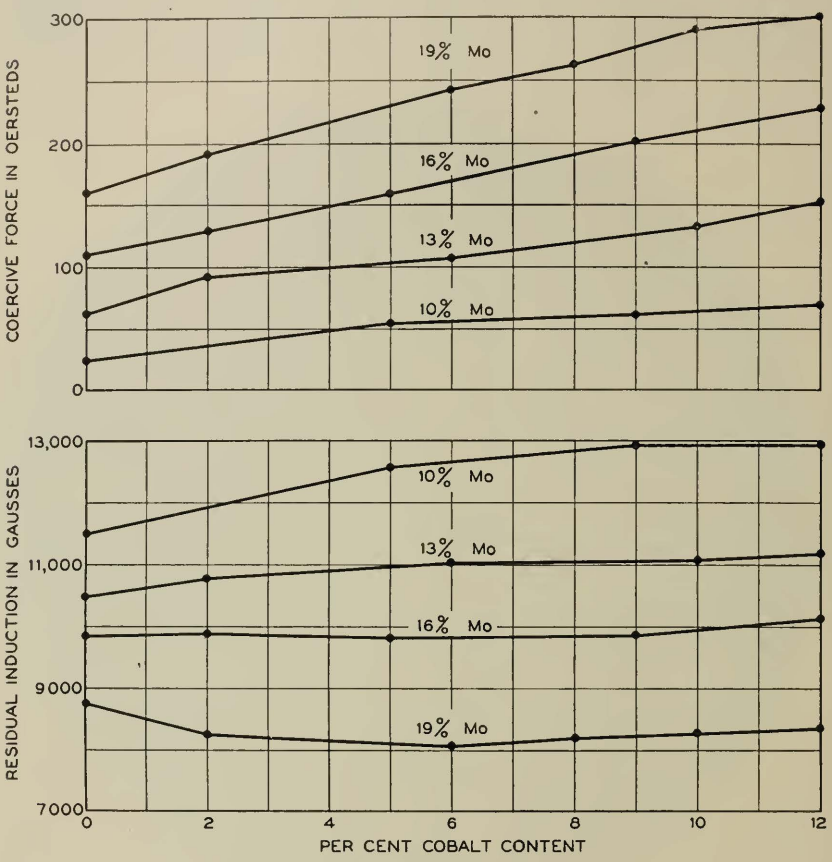


Fig. 17—Effect of different combinations of cobalt and molybdenum on the coercive force and residual induction of iron-cobalt-molybdenum alloys in the heat-treated condition—Köster[31].

aged at temperatures from 300 to 800° C. The best results are obtained by aging at temperatures between 700° C. and 800° C.

The precipitation of the second phase in a typical iron-cobalt-molybdenum alloy is illustrated by a comparison of Figs. 19 and 20. The section in Fig. 19, which is for the alloy after quenching from 1300° C., shows a typical solid solution structure with a small number of rounded inclusions resulting either from the deoxidation of the melt or from incomplete solution at the high temperature. In Fig. 20, a section

of the same alloy after quenching and subsequent aging, the precipitate is visible as rounded particles within the grains and as elongated, needle-like structures within and along grain boundaries.

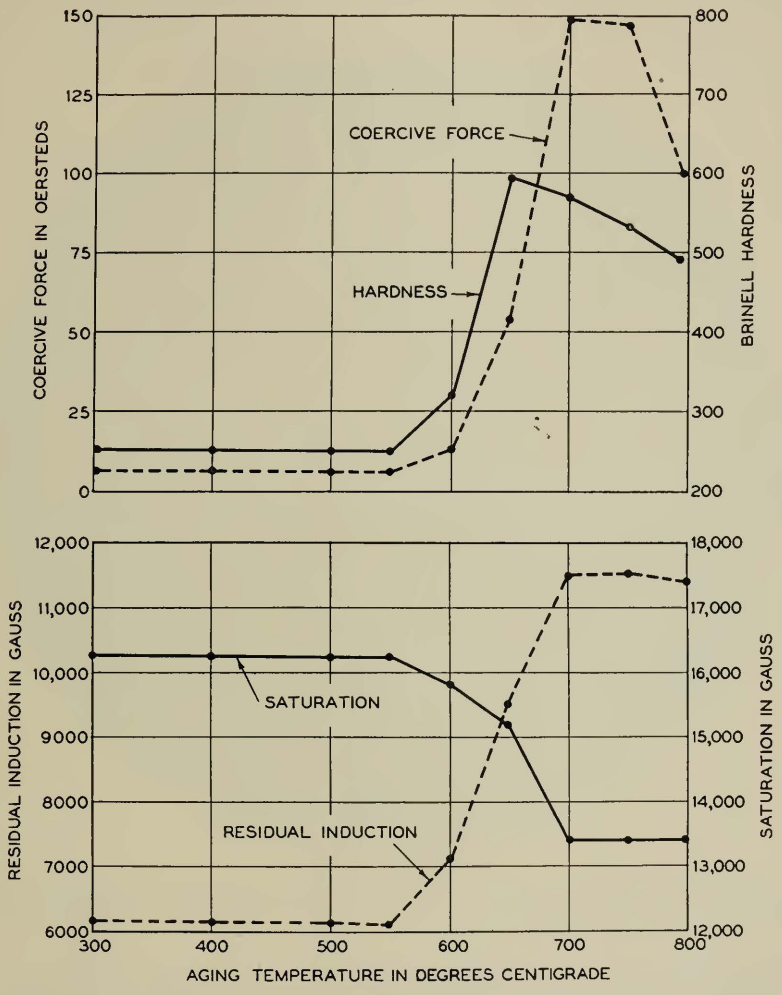


Fig. 18—Effect of aging treatment showing typical age-hardening phenomena for an iron-cobalt-tungsten magnet alloy—Köster[31].

Köster [31] reports that these alloys can be rolled or otherwise shaped at sufficiently high temperatures and can be readily stamped at red heat. The alloys machine readily in the quenched condition but on account of the coarse grained structure and consequent brittleness they must be machined without shock.

Oxide Magnets

In this class of materials are included the metallic oxide magnets recently described in the literature by Kato and Tokei.[32] The metallic oxide magnets open a new field of permanent magnet materials.

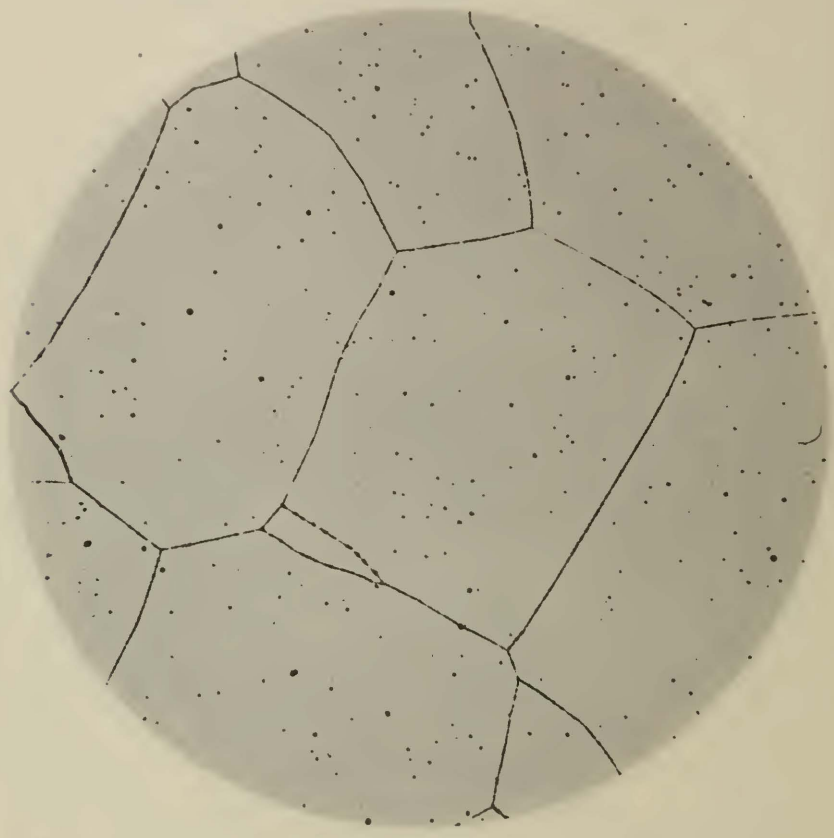


Fig. 19—Typical structure of an age-hardening iron-cobalt-molybdenum alloy, containing 12 per cent cobalt and 16 per cent molybdenum, after quenching from 1300° C. Mag. 125 ×.

These magnets are composed, according to the Japanese authors, in one instance of a solid solution of cobalt ferrite ($CoFe_2O_4$) in magnetic iron oxide ($FeFe_2O_4$). Although the method of manufacture and exact compositions are not completely disclosed, it appears that the preparation involves the powdering of the metallic oxides, compressing in a suitable die to the desired shape, and subsequent heating.

Cobalt ferrite magnets, it is reported, are not easily magnetized at room temperature; but if the temperature is raised, for example, to

300° C. the magnetization is readily accomplished. In this procedure the magnet is cooled in the magnetizing field. If the magnet has once been magnetized at the high temperature, succeeding magnetizations can be carried out at room temperature.

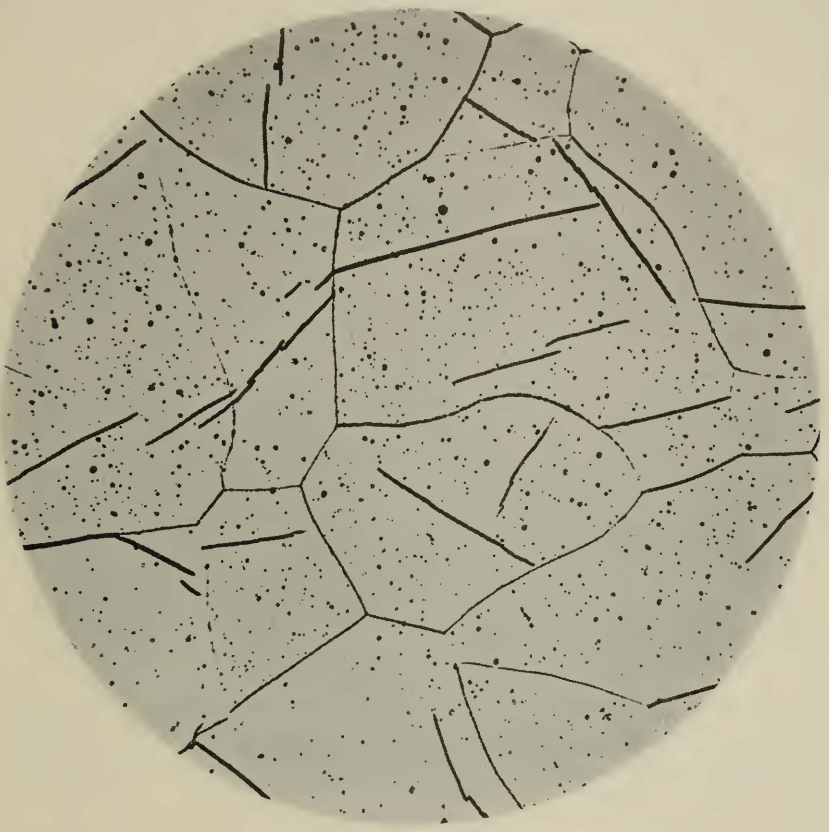


Fig. 20—The same alloy as shown in Fig. 19 after a subsequent aging treatment of 5 hours at 725° C. Mag. 125 ×.

The magnetic characteristics of the material are illustrated in Fig. 21, in which a hysteresis loop for a metallic oxide magnet is compared with one for carbon steel. The coercive force of the new material is approximately 600 oersteds and the residual intensity of magnetization, 320 units, corresponding to a residual induction of approximately 4000 gauss.

The high coercive force of the oxide magnet makes possible the more exact location of poles and permits the utilization of the material in

short magnets. Because of the lower density of the oxide compared with that of iron, the oxide magnets have approximately the same induction flux as metallic magnets on a weight basis.

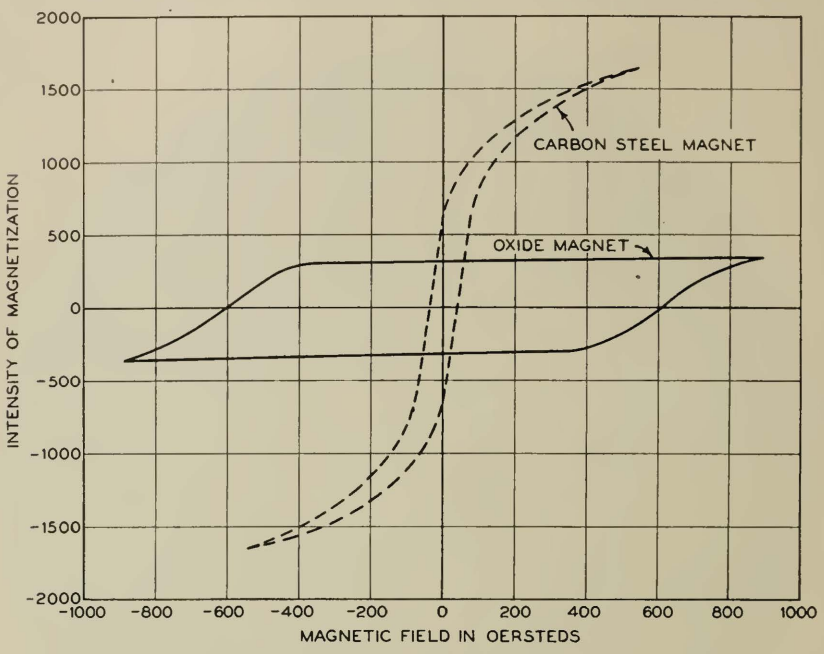


Fig. 21—Comparison of some magnetic characteristics of an oxide magnet with those of carbon steel from Kato and Tokei[32].

The description of the materials given here has been obtained from the Japanese publication.[32] So far as is known, no investigation of the preparation or of the properties of the material has been reported in this country.

Summary of New Permanent Magnet Materials

The properties of the new permanent magnet materials discussed are summarized in Table II.

TABLE II

THE PROPERTIES OF SOME NEW PERMANENT MAGNET MATERIALS

Materials	H_c-Oersteds	B_r-Gauss	$H_c \times B_r \times 10^{-3}$
Iron-Cobalt-Molybdenum Alloys	50–300	12,000–8,500	600–2,550
Iron-Cobalt-Tungsten Alloys	150	11,500	1,720
Iron-Nickel-Aluminum Alloys	130–660	10,800–7,600	1,400–5,000
Oxide Magnets	600	4,000	2,400

An important consideration in connection with the new materials is the broadening of the field of possible new magnet alloys. Previously it had been generally considered that the field of useful permanent magnet materials was confined to the plain and alloy steels. The new metallic materials are systems which are precipitation hardening, and suggest possibilities for developing useful alloys in other similar systems of iron.

References

1. Scott, K. L. "Magnet Steels and Permanent Magnets," *Elec. Engg.*, **51**, 320 (1932).
   Evershed, S. "Permanent Magnets in Theory and Practice," Jour. I. E. E., **63**, 725 (1925).
2. Messkin-Kussmann. "Die Ferromagnetischen Legierungen," p. 308, Julius Springer, Berlin (1932).
3. Scott, H. "Solubility of Carbon in Alpha Iron," *Chem. and Met. Engg.*, **27**, 1156 (1922).
   Whiteley, J. H. "Solution of Cementite, in α-Iron and Its Precipitation," *Jour. Iron and Steel Inst.*, **116**, 293 (1927).
   Bates, A. A. "Aging in Low Carbon Steels," *Trans. Amer. Soc. Steel Treat.*, **19**, 449 (1932).
4. Tritton, F. S. and Henson, D. "Ferrous Alloys Research—Pt. II—Iron and Oxygen," *Jour. Iron and Steel Inst.*, **110**, 90 (1924).
   Jordan, L. and Eckman, J. R. "Gases in Metals II—Determination of Oxygen and Hydrogen in Metals by Fusion in Vacuum," *Sci. Papers, Bu. Stds.*, Paper No. 514, **20**, 445 (1925).
5. Epstein, S. "Observations on the Iron-Nitrogen System," *Trans. Amer. Soc. Steel Treat.*, **16**, 19 (Nitriding Symposium), (1929).
6. Benedicks, C. and Löfquist, H. "Nonmetallic Inclusions in Iron and Steel," p. 13, Chapman and Hall, London.
7. Barrett, W. F., Brown, W. and Hadfield, R. A. "On the Electrical Conductivity and Magnetic Permeability of Various Alloys of Iron," Parts I and II, *Sci. Trans., Roy. Dublin Soc.*, Ser. 2, **7**, 67 (1900).
   Yensen, T. D. "Magnetic and other Properties of Iron Silicon Alloys Melted in Vacuo," *Illinois Univ. Eng. Expt. Stat. Bull.* No. 83 (1915).
8. Messkin-Kussmann. "Die Ferromagnetischen Legierungen," p. 312, Julius Springer, Berlin (1932).
9. Yensen, T. D. "Pure Iron and Allotropic Transformations," *Trans. Iron and Steel Div., A. I. M. E.*, p. 320 (1929).
10. Rohn, W. "Die Entwicklung der Heraeus—Vacuum Schmelze A-G. von 1923–1933," Heraeus Vacuum-Schmelze, 1923–1933, p. 1, Hanau, 1933.
11. Cioffi, P. P. "Hydrogenized Iron," *Physical Review*, **39**, 363 (1932).
    Cioffi, P. P. "New High Permeabilities in Hydrogen-Treated Iron," *Physical Review*, [2] **45**, 742 (1932).
12. Iron Carbonyl Patents: British No. 244,895; German No. 518,387; U. S. Nos. 1,783,744, 1,759,268, 1,725,619.
    Fieldner, A. C. and Jones, G. W. "Iron Carbonyls: Their Physical and Chemical Properties," *Amer. Gas Assoc. Monthly*, **6**, 439 (1924).
13. Buddenberg, O., Duftschmid, F., u. Schlecht, L. "Reines Carbonyleisen als hochpermeabler magnetischer Werkstoff," Heraeus Vacuum—Schmelze 1923–1933, p. 74, Hanau, 1933.
14. Keinath, G. "Hochmagnetische Legierungen aus Nickeleisen," *Archiv für technisches Messen*, **2**, T 173 (1932).
15. Pettit, I. C. "Magnetic Materials," *Bell Laboratories Record*, **12** (1934).
16. Yensen, T. D. "Magnetic Properties of the Fifty Per Cent Iron-Nickel Alloy," *Jour. Franklin Inst.*, **199**, 333 (1925).
17. Arnold, H. D. and Elmen, G. W. "Permalloy, an Alloy of Remarkable Magnetic Properties," *Jour. Franklin Inst.*, **195**, 621 (1923).
    Elmen, G. W. "Magnetic Alloys of Iron, Nickel, and Cobalt," *Jour. Franklin Inst.*, **207**, 583 (1929).

Elmen, G. W. "New Permalloys," *Bell Laboratories Record*, **10**, 2 (1931).
18. Dahl, O. "Zur Frage Unterkühlbarer Zustandsänderungen in Eisen-Nickel-legierungen," *Zeit. f. Metallkunde*, **24**, 107 (1932).
19. Haughton, J. L. and Payne, R. J. "Transformations in the Gold-Copper Alloys," *Jour. Inst. of Metals*, **46**, No. 2, 457 (1931).
20. Kussmann, A., Scharnow, B., u Steinhaus, W. "Uber das Permalloy-Problem," Heraeus Vacuum—Schmelze 1923–1933, p. 310, Hanau, 1933.
Dahl, O. and Pfaffenberger, J. "Beitrag zur Kenntnis der Eisen—Nickel legierungen," *Zeit. f. Metallkunde*, **25**, 241 (1933).
21. Dahl, O., Pfaffenberger, J. and Sprung, H. "Neurartige magnetische Werkstoffe für pupinspulen (Materialien höchster magnetischer Stabilität; Ispoerme)," *Elektrische Nachrichten Technik.*, **10**, 317 (1933).
22. Ellis, W. C. "A Study of the Physical Properties of Electrolytic Cobalt and its Alloys with Iron," *Renss. Poly. Inst. Eng. and Sci. Series* No. 16, Troy, N. Y., 1927.
Kussmann, A., Scharnow, B., u. Schulze, A. "Physikalische Eigenschaften und Struktur des Zweistoffsystems Eisen-Kobalt.," *Zeit. f. technische Physik*, **13**, 449 (1932).
23. White, J. H. and Wahl, C. V. U. S. Patent No. 1,862,559.
24. Elmen, G. W. "Magnetic Properties of Perminvar," *Jour. Franklin Inst.*, **206**, 317 (1928).
Auwers, O., u Kühlewein, H. "Beitrage zum Stereomagnetismus II—Über das Perminvar-problem," *Ann. der Physik*, [5] **17**, 121 (1933).
25. Kelsall, G. A. "Permeability Changes in Ferromagnetic Materials Heat Treated in Magnetic Fields," *Physics*, **5**, 169, 1934.
Bozorth, R. M., Dillinger, J. F. and Kelsall, G. A. "Magnetic Materials of High Permeability Attained by Heat Treatment in a Magnetic Field," *Physical Review*, [2], **45**, 742 (1934).
McKeehan, L. W. and Cioffi, P. P. "Magnetostriction in Permalloy," *Physical Review*, [2] **28**, 146 (1926).
McKeehan, L. W. "Ferromagnetism in Metallic Crystals," *Metals Technology*, p. 1, August 1934.
Bozorth, R. M. "Theory of the Heat Treatment of Magnetic Materials," *Phys. Review*, **46**, 232 (1934).
26. Heusler, F. "Über den Zusammenhang der magnetischen und mechanischen Eigenschaften der gewalzten Heusler-Bronz (Mangan-Aluminium-Kupfer)," *Zeit. f. Physik.*, **10**, 403 (1922).
Persson, E. "Über den Bau der Heuslerschen Legierungen," *Zeit. f. Physik.*, **57**, 115 (1929).
Potter, H. H. "Some Magnetic Alloys and Their Properties," *Phil. Mag.*, [7] **12**, 255 (1931).
Bradley, A. K. and Rogers, J. W. "The Crystal Structure of the Heusler Alloys," *Proc. Roy. Soc. A.*, **144**, 340 (1934).
Heusler, O. "Kristallstruktur und Ferromagnetismus der Mangen-Aluminium-Kupferlegierungen," *Annalen der Physik*, [5] **19**, 155 (1934).
27. Shackelton, W. J. and Barber, I. G. "Compressed Powdered Permalloy," *Trans. Amer. Inst. Elec. Eng.*, **47**, 429 (1928).
White, J. H. U. S. Patent No. 1,739,052.
Beath, C. P. and Heincke, H. M. E. U. S. Patent No. 1,669,649.
Andrews, J. W. U. S. Patent No. 1,703,287.
28. Sanford, R. L. "Some Principles Governing the Choice and Utilization of Permanent Magnet Steels," *Sci. Papers, Bu. Stds.* Paper No. 567, **22**, 557 (1927).
Mathews, J. A. "Retained Austenite—A Contribution to the Metallurgy of Magnetism," *Trans. Am. Soc. Steel Treat.*, **8**, 565 (1925).
29. Honda, K. and Saito, S. "On K. S. Magnet Steel," *Physical Review*, [2] **16**, 495 (1920).
30. Mishima, T. "Magnetic Properties of Iron-Nickel-Aluminum Alloys," *Ohm* (Coversheet), July, 1932. *Iron Age* (Abstract), p. 346, Sept. 1932.
French Patent No. 731,361.
English Patents Nos. 392,656; 392,657; 392,658; 392,659; 392,660 and 392,661.
Köster, W. "Das System Eisen-Nickel-Aluminium," *Archiv. für das Eisen-hüttenwesen*, **7**, 257 (1933).

Steinhaus, W. u Kussmann, A. "Neuere Entwicklung der Dauermagnetstähle," *Phys. Zeitschrift*, **35**, 377 (1934).
"Nickel-Aluminium Magnetstahl," *Archiv. für technisches Messen*, **3**, T56 (1934).
31. Seljesater, K. S. and Rogers, B. A. "Magnetic and Mechanical Hardness of Dispersion Hardened Iron Alloys," *Trans. Amer. Soc. Steel Treat.*, **19**, 553 (1932).
Köster, W. "Über die Beziehungen der magnetischen Eigenschaften, insbesondere der Koerzitivkraft zum Gefügeaufbau der Legierungen und die Entwicklung neuartiger Magnetlegierungen," *Zeit. f. Elektrochemie u. Angewandte Physikalische Chemie*, **38**, 549 (1932).
Köster, W. "Dauermagnetiwerkstoffe auf der Grundlage der Ausscheidungshärtung," *Stahl u. Eisen*, **53**, 849, 1933.
32. Kato, Y. and Tokei, I. "Permanent Oxide Magnet and Its Characteristics," *Jour. Inst. of Elec. Eng. of Japan*, **53**, 408 (1933).

Theory of Multi-Electrode Vacuum Tubes *

By H. A. PIDGEON

Physical principles underlying the characteristics and performance of multi-electrode vacuum tubes are presented in simple form in this paper. Presentation of the subject is based, as far as possible, on the well-known theory of the three-electrode tube. It is shown that the definitions of electrical tube parameters applicable to triodes are, with certain modifications in their interpretation, also applicable to tubes having more than three electrodes.

The characteristics of the screen-grid tetrode are discussed in detail since they are typical of those found in a number of multi-electrode structures. Except for the effects produced by the emission of secondary electrons from the plate and screen, it is shown that the characteristics of the screen-grid tetrode are in general accord with those to be expected from the application of simple theory. The presence of the electrostatic screen in such structures inherently results in high values of the plate resistance and amplification factor, but the transconductance remains normal and has about the same value as in comparable triodes.

One of the necessary modifications in the screen-grid tetrode to produce a satisfactory output power tube is some means of removing the fold in the plate current-plate voltage characteristics, which limits the permissible plate voltage swings. This is accomplished in the power pentode by the addition of a suppressor grid between the plate and screen grid. The efficiency of power pentodes, and of some other tubes having positive grids, is higher than that usually found in triodes. The reason for this is discussed and also certain peculiarities in the harmonic output of pentodes.

The arrangement of electrodes in the space-charge-grid pentode corresponds to that of the screen-grid tetrode with an additional grid inserted between the cathode and control grid. This space-charge grid, which is maintained at a positive potential of 10 to 20 volts with respect to the cathode, reduces the effects of space charge near the cathode surface. This results in extraordinarily high values of transconductance and, consequently, in high amplification. Practically, such tubes are limited to use as voltage amplifiers, since operation over the wide range necessary for large output power results in prohibitive distortion.

In the co-planar-grid tetrode the lateral wires of the positive grid are arranged in the same planes as those of the control grid. This results in comparatively low plate resistance while retaining the advantages of a positive grid. The plate efficiency is comparable with that in the power pentode but the available amplification is lower.

INTRODUCTION of the three-electrode vacuum tube into the field of communications and in other applications represented such a tremendous advance over the possibilities of any other known device that, despite some of its rather obvious limitations, it proved entirely adequate for the service required until comparatively recent years. However, with increasing demands made by service requirements for larger power output at higher efficiency, reduced distortion, higher

* Published in November 1934 issue of *Electrical Engineering*. Scheduled for presentation at Winter Convention of A. I. E. E., New York City, January 22–25, 1935.

gain, amplification at higher frequencies and greater frequency discrimination, it eventually became necessary to investigate the possibilities of making changes in vacuum tubes enabling them to meet these requirements more satisfactorily. Measures taken to meet this situation have included improvements in the three-electrode tube that reduce the effects of some of its limitations, and the development of vacuum tubes having more than three electrodes.

The purpose of this paper is to present, in simple form, the physical principles underlying the characteristics and performance of multi-electrode vacuum tubes. For present purposes, such tubes may be defined as those having more than the three electrodes of the conventional triode. The procedure will be to show that the definitions of electrical tube parameters applicable to triodes are, with certain modifications in their interpretation, also applicable to tubes having more than three electrodes; and, utilizing the theory of the triode, to analyze the characteristics of a few typical multi-electrode structures that illustrate the types of characteristics found in many such tubes. No attempt is made to present new material in the paper or to discuss in detail the many different types of multi-electrode tubes now in use. The author has attempted to present the subject from the viewpoint of those readers who have a satisfactory understanding of the physical principles, characteristics, and operation of the triode, but who do not have a similarly clear analysis available for the more complex structures.

Multi-electrode tubes may be divided conveniently into two classes. In the first class are those the purpose of which is to perform some function that cannot be performed readily by a triode, or which perform some function better by reason of the elimination or reduction of some limitation in triodes. The second class includes those structures in which additional electrodes are introduced to permit them to perform simultaneously more than one function, or to permit them to function in two or more ways, depending on the voltages applied to the various electrodes and on the manner of their operation. This paper will deal exclusively with typical structures of the first class.

Fundamental Definitions and Tube Equations

Regardless of the type of multi-electrode tube considered, the space current to any electrode may be expressed as some function of the voltages applied to the various electrodes. However, in the operation of any multi-electrode device, or any section of such a device performing a single function, one usually is concerned with variations in the voltages and corresponding currents of only two of the electrodes, the other electrodes being maintained at fixed potentials. One of these

two electrodes, which usually is maintained at a negative operating potential, is connected to the input circuit and acts as a control electrode. The second of these two electrodes, which is maintained at a positive potential with respect to the cathode, acts as an anode or collector of electrons and is connected in the output circuit. Just as in the case of the triode, then, a study of the characteristics of multi-electrode tubes is concerned with variations in the current collected by the anode with variations in the potential applied to the control grid.

This anode or plate current may be expressed as a function of the various electrode voltages by the following equation:

$$I_p = f(E_p, E_{g_1}, E_{g_2}, E_{g_3}, \text{etc.}), \tag{1}$$

in which E_p is the operating voltage applied to the output anode or plate, and E_{g_1}, E_{g_2}, E_{g_3}, etc., are the operating voltages applied to the various grids numbered outward from the cathode. The variation in the anode current, neglecting second and higher order terms, is given by

$$dI_p = \frac{\partial I_p}{\partial E_p} dE_p + \frac{\partial I_p}{\partial E_{g_1}} dE_{g_1} + \frac{\partial I_p}{\partial E_{g_2}} dE_{g_2} + \frac{\partial I_p}{\partial E_{g_3}} dE_{g_3} + \text{etc.} \tag{2}$$

The partial differential coefficients in equation 2 have the physical dimensions of conductances and, if these conductances be designated by the letter S with appropriate subscripts, the equation may be written

$$dI_p = S_{pp} \cdot dE_p + S_{p1} \cdot dE_{g_1} + S_{p2} \cdot dE_{g_2} + S_{p3} \cdot dE_{g_3} + \text{etc.} \tag{3}$$

The plate or output anode conductance of a multi-electrode tube is defined in the same manner as for the three-electrode tube. It is the rate of change of plate current with plate voltage, that is, it is the slope of the plate current-plate voltage characteristic at the selected operating point, the potentials of all the other electrodes remaining constant. Under this condition, from equations 2 and 3

$$\text{Plate conductance} = \frac{\partial I_p}{\partial E_p} = S_{pp} = S_p. \tag{4}$$

Obviously, the plate resistance also must be defined in the same manner as in the triode, that is

$$\text{Plate resistance} = R_p = \frac{1}{\dfrac{\partial I_p}{\partial E_p}} = \frac{1}{S_p} \cdot \tag{5}$$

In a similar manner, the transconductance from the control grid to

the output anode or plate of a multi-electrode tube is defined, as it is in the triode, by the rate of change of plate current with variation of the control-grid voltage; that is, it is the slope of the plate current-grid voltage characteristic at the given operating point, the potentials of all electrodes other than the control grid remaining constant.

In conventional screen-grid tetrodes and pentodes, the grid next to the cathode is the control grid. Consequently, for such structures, the transconductance is defined from equations 2 and 3 by

$$\text{Transconductance} = \frac{\partial I_p}{\partial E_{g_1}} = S_{p1}. \tag{6}$$

In space-charge-grid tetrodes and pentodes, the grid next to the cathode is maintained at a fixed positive potential and the second grid acts as the control grid. Consequently, in these and similar structures

$$\text{Transconductance} = \frac{\partial I_p}{\partial E_{g_2}} = S_{p2}. \tag{7}$$

Similarly, considering the control grid (assumed to be the first grid) and the output electrode of a multi-electrode tube, the amplification factor is defined, as it is in the triode, by the ratio of the transconductance to the plate conductance. It is expressed by

$$\text{Amplification factor} = \mu_{pg_1} = \frac{\dfrac{\partial I_p}{\partial E_{g_1}}}{\dfrac{\partial I_p}{\partial E_p}}. \tag{8}$$

Or, assuming that E_{g_1} and E_p are varied in such a manner that I_p remains constant, the amplification factor is expressed in the usual form by

$$\mu_{pg_1} = -\frac{dE_p}{dE_{g_1}}\Bigg]_{I_p = \text{constant}}. \tag{9}$$

Combining equations 5 and 6 with equation 8 gives

$$\text{Transconductance } S_{p1} = \frac{\mu_{pg_1}}{R_p} \tag{10}$$

just as in the case of the triode. Exactly similar equations apply if g_2 is used as the control grid.

Obviously, the currents to the other electrodes in a multi-electrode tube may be expressed by functions of the electrode voltages, similar to equations 1 and 2. The various differential coefficients of these equa-

tions define transconductances, electrode resistances, and amplification (or reflex) factors analogous to those just given. Since these quantities are not used in this paper, they will not be given further consideration here. The voltage applied to the control grid will be designated by E_g, regardless of the grid employed for the purpose; and the transconductance (or mutual conductance) and the amplification factor, applying to the control grid and plate, will be designated by S_m and μ, respectively.

If a load resistance, R, is inserted in the plate circuit of a multielectrode tube, and if the potentials of all of the elements other than the control grid and plate are maintained constant, equation 3 reduces to

$$dI_p = S_p \cdot dE_p + S_m \cdot dE_g = \frac{1}{R_p} \cdot dE_p + \frac{\mu}{R_p} \cdot dE_g. \tag{11}$$

In this case, the only independent variable is E_g, and E_p varies by reason of the changing potential drop across the external load resistance, R, due to variations in the plate current, I_p, produced by the varying grid potential. Consequently

$$dE_p = - dI_p \cdot R. \tag{12}$$

Substituting equation 12 in equation 11 and reducing,

$$dI_p = \frac{\mu}{R_p + R} \cdot dE_g. \tag{13}$$

For vacuum tubes having curvilinear characteristics, equation 13 applies rigorously, of course, only to infinitesimal variations in I_p and E_g. However, as in the case of the triode, the output from multielectrode tubes may be expressed by a power series in terms of finite voltage variations applied to the elements, the coefficients in the series being functions of the static characteristics. If these finite variations in I_p and E_g are designated by i_p and e_g, respectively, the output current is expressed to the first order by

$$i_p = \frac{\mu e_g}{R_p + R}, \tag{14}$$

which is identical with the equation expressing the output current from a triode.

Letting e_p represent the variable voltage across the load resistance, R, the voltage amplification is given by

$$A_v = \frac{e_p}{e_g} = \frac{i_p \cdot R}{e_g} = \frac{R}{R_p + R} \cdot \mu. \tag{15}$$

It may also be written in the following form which will be found useful later:

$$A_v = \frac{R_p \cdot R}{R_p + R} \cdot \frac{\mu}{R_p} = \frac{R_p \cdot R}{R_p + R} \cdot S_m. \tag{16}$$

It should be emphasized, perhaps, that the electrical parameters of multi-electrode tubes and the output current, as defined by the foregoing equations, are subject to the condition that the voltages applied to all of the electrodes other than the plate and control grid are maintained constant. The satisfactory operation of multi-electrode tubes in circuits also usually requires that this condition be fulfilled. It requires that the impedance to alternating current components in each of these circuit branches be very low. This is accomplished in practice by connecting these electrodes to ground, so far as alternating currents are concerned, through reasonably large capacitances.

From the foregoing analysis it is apparent that, with proper interpretation, the definitions of plate resistance, transconductance, and amplification factor applicable to triodes are also applicable to multi-electrode tubes; in addition, the same expressions for output current and voltage amplification are applicable. This follows from the fact that these quantities are expressed in terms of the differential coefficients of the static characteristics, that is, they depend only upon the slopes of these characteristics at the given operating voltages and not upon their form. However, as will appear later, the difference in the shape of the static characteristics of multi-electrode tubes from those of triodes is very important in determining great differences not only in the magnitude of the electrical parameters, but also in the character and amount of distortion resulting when the tubes are operated under conditions such that large portions of the characteristics are traversed.

In multi-electrode tubes, as well as in triodes, the total space current drawn from the cathode is determined by the extent to which the resultant field, due to the electrodes, overcomes the opposing field produced by space charge. While space charge extends throughout the interelectrode space, it is relatively so much more dense in regions of very low electron velocity that, as a first approximation, its effect usually may be neglected in other regions. Except in space-charge-grid tubes and a few other special tubes, the only important space-charge region is confined to a relatively thin sheath near the cathode surface. Consequently, in such structures the total space current is determined largely by the extent to which the resultant positive field due to the electrodes neutralizes the negative field near the cathode

surface produced by space charge. An appreciation of this fact is essential to a clear understanding of the characteristics of multi-electrode tubes.

Screen-Grid Tetrodes

Utilizing the simple theory of triodes, which has been shown to be applicable to multi-electrode tubes also, the characteristics of a multi-electrode tube will be analyzed next. For this purpose the screen-grid tube is chosen, since it admirably illustrates the type of characteristics found in several types of multi-electrode tubes.

The objective in this case is to reduce the direct capacitance between the plate and control grid through which energy is fed back from the plate to the input circuit. This is accomplished by inserting an

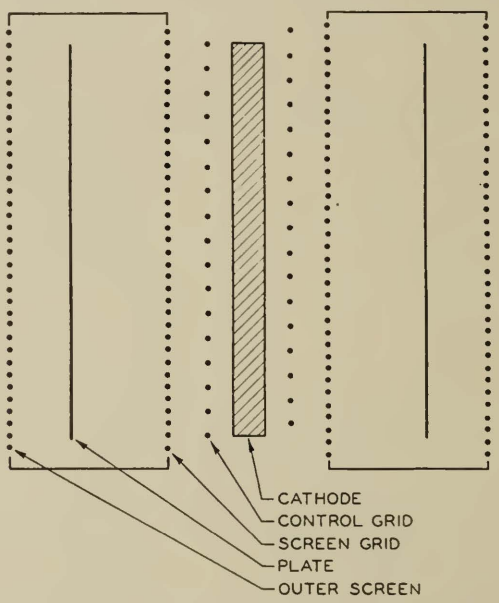


Fig. 1—Schematic diagram showing the arrangement of electrodes in a screen-grid tetrode.

electrostatic shield between the plate and control grid of what otherwise would be a three-electrode tube. The condition that such a screen must allow an electron stream to flow through it to the plate, is satisfied by making it of fine mesh or in the form of a finely wound grid structure. To be effective, it must be maintained at some constant potential with respect to ground.

The arrangement of the electrodes in such a tube is shown diagrammatically in Fig. 1. The screen structure outside the plate is added for

the purpose of completing the electrostatic isolation of the plate and its leads from the grid, thus reducing the capacitance between these two elements to the lowest practicable value. This outer screen is of no further concern, since it has no effect on the static characteristics of the tube. The usual characteristics of a typical screen-grid tetrode are shown in Figs. 2, 3, and 4. In this case, as is usual in screen-grid tubes, the shielding is very complete, reducing the direct capacitance between the plate and control grid to a few thousandths of what its value would be in the absence of the screen.

First, the characteristics of Fig. 2 will be considered. Since the direct capacitance between the plate and the control grid, g_1, is extremely small, the electric field in the immediate vicinity of the latter, produced by any potential on the plate, also must be extremely small. The cathode is electrically even more remote from the plate than the control grid, because it not only is shielded from the former by the screen grid, but also has some additional shielding from the control grid. Consequently, the field produced by the plate at points between the cathode and control grid is smaller even than the field produced at the control grid and, hence, is negligibly small. Since, as previously discussed, the total space current is determined almost wholly by the field very near the cathode surface, the plate in this case can have practically no effect on the total space current drawn from the cathode. This is shown by the curves giving the total space current, I_t, in Fig. 2. These curves are seen to be so flat as to be almost entirely independent of variations in the plate voltage.

The plate, then, in a screen-grid tube plays an essentially passive rôle which is to collect those electrons that succeed in passing through the screen. The remainder of the space current is collected by the screen, the sum of the plate and screen currents remaining nearly constant with changes in plate voltage.

There is nothing in the theory of the triode by which to determine the ratio in which space current divides between two or more positive electrodes in a multi-electrode tube. As a rough approximation, one might assume that when their potentials are nearly equal, the currents to the plate and screen would be proportional to the ratio of the area of the openings in the screen to the area subtended by its lateral wires. Also, it might be expected that this ratio would increase slightly with increasing voltage of the plate with respect to the screen, because of a tendency of the plate to pull more electrons through the screen. This effect should be less for very fine mesh screens than for coarse ones.

From this simple theory, one would expect the plate current-plate voltage characteristics to be very flat for plate potentials higher than

the screen potential. Consequently, the plate conductance, given by

$$S_p = \frac{\partial I_p}{\partial E_p} = \frac{1}{R_p}$$

is a very small quantity; and the plate resistance, given by

$$R_p = \frac{1}{\frac{\partial I_p}{\partial E_p}}$$

is a very large quantity compared with its value in triodes, and increases with the fineness of mesh of the screen grid. The plate-current curves of Fig. 2 are seen to be in general accord with this simple theory at the higher values of plate voltage, although they are not quite as flat as might be expected from the theory. This, and the rapid falling off in the vicinity of 100 volts, will be discussed later.

Taken alone, the extremely high resistance of a screen-grid tube would seem to be a very serious disadvantage. From equation 14, the tube may be considered as analogous to a generator the electromotive force of which is μe_g and the internal resistance of which is R_p, working into an external load resistance R—a generator with extremely high internal resistance. Why this is not fatal to the usefulness of the tube will be pointed out later.

While the plate current in a screen-grid tube is nearly independent of plate voltage for values of the latter higher than the screen potential, this obviously cannot hold at low values of plate voltage. At zero plate voltage, the plate current must be zero. At this point the screen collects the entire space current and $I_s = I_t$. As the plate potential increases from zero, the plate current would be expected to rise rapidly, with a corresponding drop in the screen current, as the plate collects more and more of the electrons passing through the screen. However, two factors tend to prevent an abrupt rise in the plate current to its nearly constant value when the plate becomes slightly positive. The first of these is space charge in the region closely adjacent to the plate produced by the electrons that pass through the screen, reach zero velocity in the region adjacent to the plate, and return to the screen. Some of them may perform several oscillations to and fro through the screen before being captured by it. This space-charge effect is largely masked by the more important second factor which is the deflection of the majority of the electrons from their normal paths by the intense electric fields about the lateral wires of the screen. This results in large differences in the components of velocity

of the electrons normal to the surface of the plate and, consequently, in the distance to which they approach the plate in their trajectories before being turned back to the screen. As a result, the plate must become positive by several volts with respect to the cathode before it captures substantially all of the electrons that pass through the screen.

From this simple theory, the plate-current and screen-current curves would be expected to have the form shown by the ideal curves of Fig. 2. Obviously, the screen-current curves must be complementary to the plate-current curves since the sum of the two currents is substantially constant.

The difference between these ideal curves and the actual characteristics, in the region extending from a few volts to potentials somewhat higher than the screen voltage, is attributed to the phenomenon of secondary electron emission. When electrons strike a metal surface with velocities equivalent to more than a few volts, other electrons, known as secondary electrons, are liberated from the surface. The number of electrons so liberated varies not only with the velocity of the primary bombarding electrons, but also with the character of the metal surface, the amount of adsorbed gases and other materials on the surface, and other factors. The number of such electrons leaving the surface may even exceed the number of primary electrons striking it, in which case the net current to the metal surface is negative. The velocity of the secondary electrons varies greatly. A very few have velocities approaching that of the primary electrons. The great majority, however, have low velocities equivalent to only a few volts.

In the screen-grid tube, an appreciable number of secondary electrons is liberated from the plate at potentials between 5 and 10 volts, and they increase in number with plate voltage. For plate potentials lower than the screen potential, in this case 75 volts, the secondary electrons from the plate are drawn to the screen, thus increasing the screen current by the amount the plate current is decreased. When the plate reaches a potential equal to that of the screen, secondary electrons no longer can escape from the plate to the screen, except those emitted with appreciable velocities; consequently, the plate current rises rapidly to its normal value.

At plate potentials higher than the potential of the screen, secondary electrons emitted from the latter are drawn to the plate. Consequently, in this region the plate current is slightly higher than it would be in the absence of secondary electron emission from the screen. The gradual rather than abrupt rise in the plate current curves at 75 volts is attributed primarily to the distribution of velocities with which the secondary electrons are emitted; to a lesser extent, it is dependent also

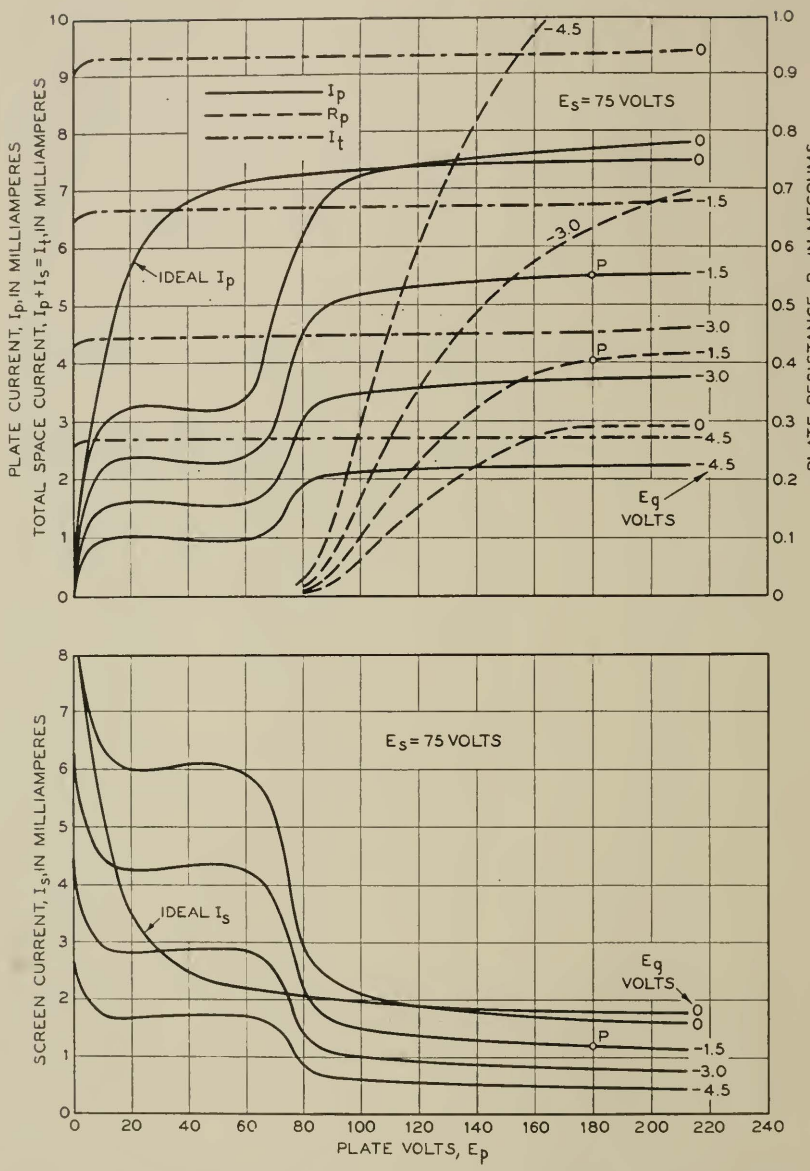


Fig. 2—Characteristics of a screen-grid tetrode. The curves show the plate current, I_p, screen current, I_s, total space current, $I_t = I_p + I_s$, and plate resistance, R_p, as functions of the plate voltage. Control-grid voltage, E_g, is indicated on the curves. Screen voltage, E_s, is maintained constant at 75 volts throughout.

on the combined effects of space charge, intensity and distribution of the field at the surface of the screen wires.

Normal operating conditions for the screen-grid tube the characteristics of which are shown in Fig. 2 are: $E_p = 180$ volts, $E_s = 75$ volts, and $E_g = -1.5$ volts. At this operating point P in Fig. 2, the plate current is 5.5 ma and the plate resistance is 400,000 ohms. The operating range is confined to the flat portion of the characteristics. If the tube is operated with plate-voltage swings sufficiently large that the instantaneous values of the plate potential extend into the region of rapidly falling plate current, serious distortion of the output results. This is a serious limitation in screen-grid tetrodes, because it requires that the operating plate potential be much higher than the screen potential. How this limitation may be removed by the introduction of an additional electrode in the tube will be shown later.

Curves showing the variation of plate resistance with plate voltage are shown in Fig. 2. The plate resistance decreases with plate voltage, falling off very rapidly as the plate voltage approaches the screen voltage. The ordinates of the plate-resistance curves give a measure of the flatness of the plate-current curves. That the latter are not as flat as the total-space-current curves, thus resulting in values of plate resistance approaching infinity, is attributed to two factors: secondary electron emission from the screen, and an increasing ratio of plate current to screen current with increasing plate voltage. The increasing percentage of the primary space current drawn to the plate with increasing plate voltage is an involved and undetermined function of several factors including: the ratio of the openings in the screen grid to the total conducting area subtended by it, the intensity and distribution of the field at the screen grid, the velocity and directional distribution of the electrons arriving at the screen grid, and space-charge effects in its vicinity.

Thus, there is the interesting situation in screen-grid tubes that, for any given set of operating voltages, the magnitudes of the plate resistance and amplification factor are determined largely by factors not directly determined by the geometry and design of the tube. This is quite different from triodes in which the plate resistance and amplification factor both are determined directly by geometrical dimensions and the arrangement of the electrodes.

Obviously, the number of electrical parameters in vacuum tubes increases rapidly with the number of electrodes. The curves of Fig. 2, which correspond to the usual plate current-plate voltage characteristics for a triode, were obtained with the screen maintained at a constant potential of 75 volts as a fixed parameter. To obtain a complete

charting of the characteristics would require several such families of curves taken with different screen voltages. One such additional family of characteristics, in which E_s is maintained at 90 volts, is shown in Fig. 3.

Since the plate in a screen-grid tube has practically no effect on the total space current, so far as consideration of the latter is concerned, one may regard the plate as being removed from the structure and consider only the remaining elements. The cathode, control grid, and screen then may be regarded as constituting an ordinary triode. By maintaining the screen at a positive potential, it is enabled to perform the function usually performed by the plate of a triode, *viz.*, that of supplying the positive field necessary to produce the flow of space current against the opposing resistance due to space charge. This function does not interfere with its screening action, so long as its potential is not allowed to vary. In fact, there is an actual gain in efficiency, as will be shown later, in having the electrode that provides the main driving field for the space current maintained at a fixed potential instead of varying over the operating cycle, as it must in the triode.

Usual design principles and equations applicable to triodes are also applicable to screen-grid tetrodes. By making the spacings between the electrodes small, particularly that between the cathode and control grid, a high transconductance can be obtained. Using the subscript t to designate total space current, the transconductance is given by $S_t = \frac{\partial I_t}{\partial E_g}$ and has the same value as that for a triode of the same dimensions. Let f represent the fraction of the total space current that passes through the openings in the screen and is collected by the plate, which is assumed to be at a higher potential than the screen. By proper design, f can be made large, say from 0.7 to 0.9, in the normal operating range. The transconductance from the control grid to the plate is given by

$$S_m = \frac{\partial I_p}{\partial E_g} = f \cdot \frac{\partial I_t}{\partial E_c},$$

which is 0.7 to 0.9 of the normal value of transconductance for the three-electrode tube.

The amplification factor, μ, for the tetrode, is given by

$$\mu = \frac{\frac{\partial I_p}{\partial E_g}}{\frac{\partial I_p}{\partial E_p}} = S_m \cdot R_p. \tag{17}$$

Since the transconductance, S_m, has a value not greatly different from the normal value for a triode, and since R_p is very large, μ, which is proportional to R_p, also must be very large. Referring to equation 14

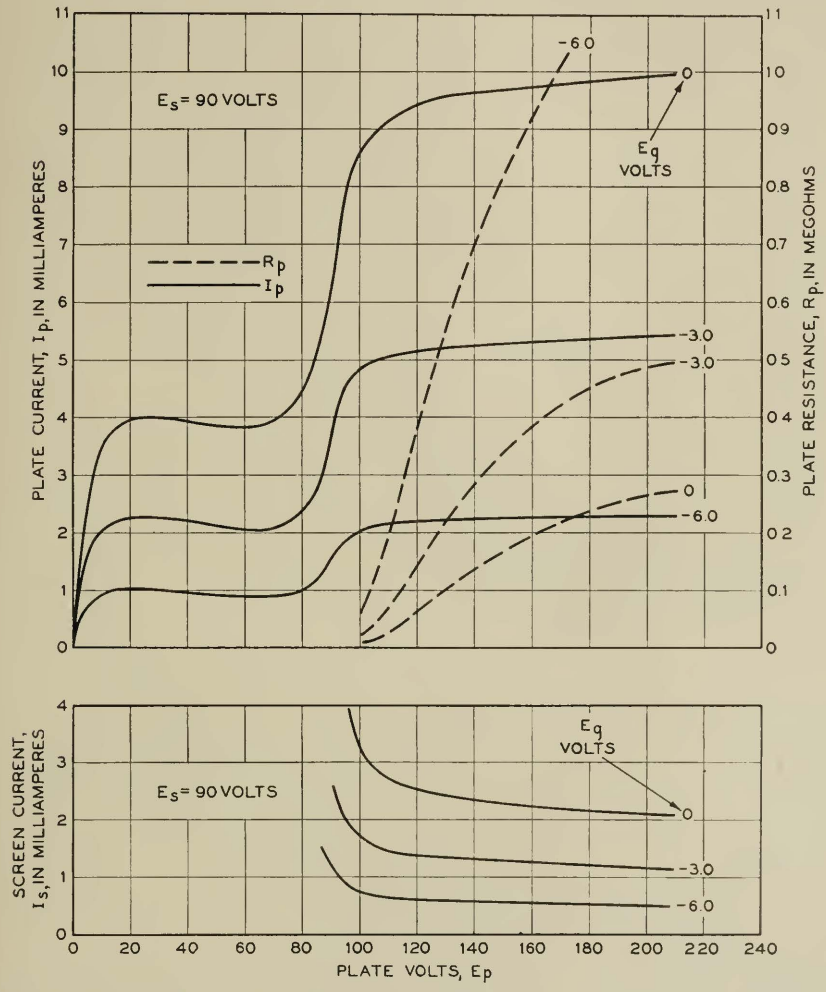


Fig. 3—Characteristics of a screen-grid tetrode similar to those shown in Fig. 2 except that E_s = 90 volts.

and its analogy with the generator equation, it may be seen that, despite the fact that the screen-grid tube considered as a generator has a very high internal resistance, it gives normal output because it is provided also with a very large electromotive force, μe_g.

From this analysis of the characteristics of the screen-grid tube,

based upon the simple theory of triodes, one would expect that families of characteristics corresponding very closely with those of ordinary triodes would be obtained if the total space current, or the plate current,

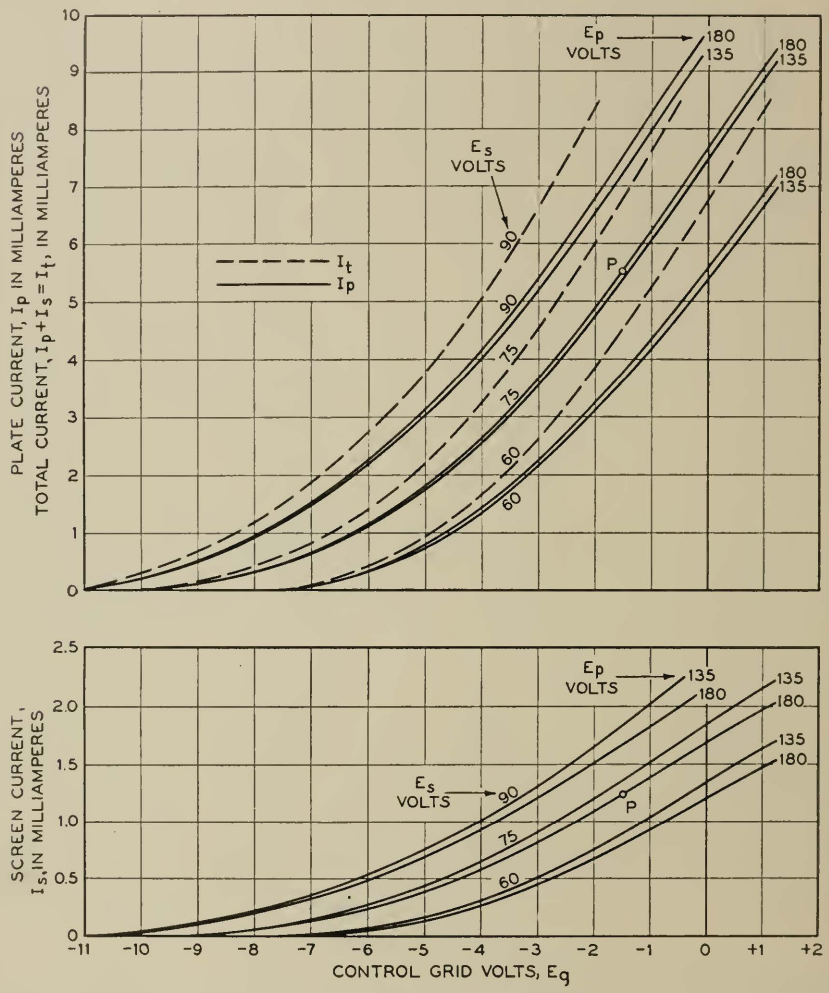


Fig. 4—Characteristics of a screen-grid tetrode. Plate current, I_p, screen current, I_s, and total space current, I_t, as functions of the control-grid voltage, with various values of plate voltage and screen voltage as parameters.

were plotted as functions of the control-grid voltage and screen-grid voltage, with the plate maintained at any fixed potential higher than the screen potential. It would be expected further that the family of curves showing total space current would remain practically invariant

with variations in plate voltage. The plate-current family of characteristics should show only small variations with plate voltage (so long as E_p is higher than E_s), depending on the magnitude of the change in the ratio of I_p to I_s.

Families of such characteristics are shown in Fig. 4. They are seen to correspond very closely indeed with similar characteristics for a triode. The I_t curves vary so little with plate voltage that the families of characteristics taken at $E_p = 135$ volts and $E_p = 180$ volts, coincide within the breadth of the curves. The plate current curves show a small variation with E_p, as was discussed previously.

In Fig. 5 the transconductance and amplification factor are shown as functions of the plate voltage for four different values of grid bias and with the screen potential maintained constant at 75 volts. At the normal operating point P, the transconductance is 1375 micromhos and the amplification factor is 550. The shape of these curves is typical of that for screen-grid tubes and pentodes. Throughout the normal operating range, the transconductance curves have about the same degree of flatness as the plate-current curves. This is to be expected from consideration of the plate-current curves of Fig. 4. Since these curves change only slightly with variations in plate potential, their slopes or transconductance values also change but slightly. The amplification-factor curves are very similar in form to the plate-resistance curves of Fig. 2. This follows at once from equation 17, for since S_m remains nearly constant with variations in E_p, μ must vary in the same manner as R_p.

In Fig. 6 the transconductance, plate resistance, and amplification factor are shown as functions of grid voltage for three different values of the screen voltage and with the plate voltage maintained constant at 180 volts. These curves are also typical of those found for several multi-electrode tubes. The transconductance curves agree in form with those for a triode, as would be expected from the plate-current curves of Fig. 4. The plate-resistance curves are similar in form to those for a triode, but rise more rapidly with increasing negative grid bias. The amplification-factor curves, however, are entirely different in form from those for a triode, since they rise with increasing negative grid bias, whereas in triodes the amplification factor decreases with increasing negative grid bias. The reason for this difference is that the plate resistance in many multi-electrode tubes, over the normal range of operation, increases more rapidly with decreasing plate current than in triodes. At sufficiently large negative values of grid bias and very low plate currents, the amplification-factor curves frequently reach maxima and then fall rapidly as they do in triodes.

In the operation of triodes as voltage amplifiers, it is usually possible to have the external load resistance large with respect to the plate resistance. Consideration of equation 15 shows that this results in a

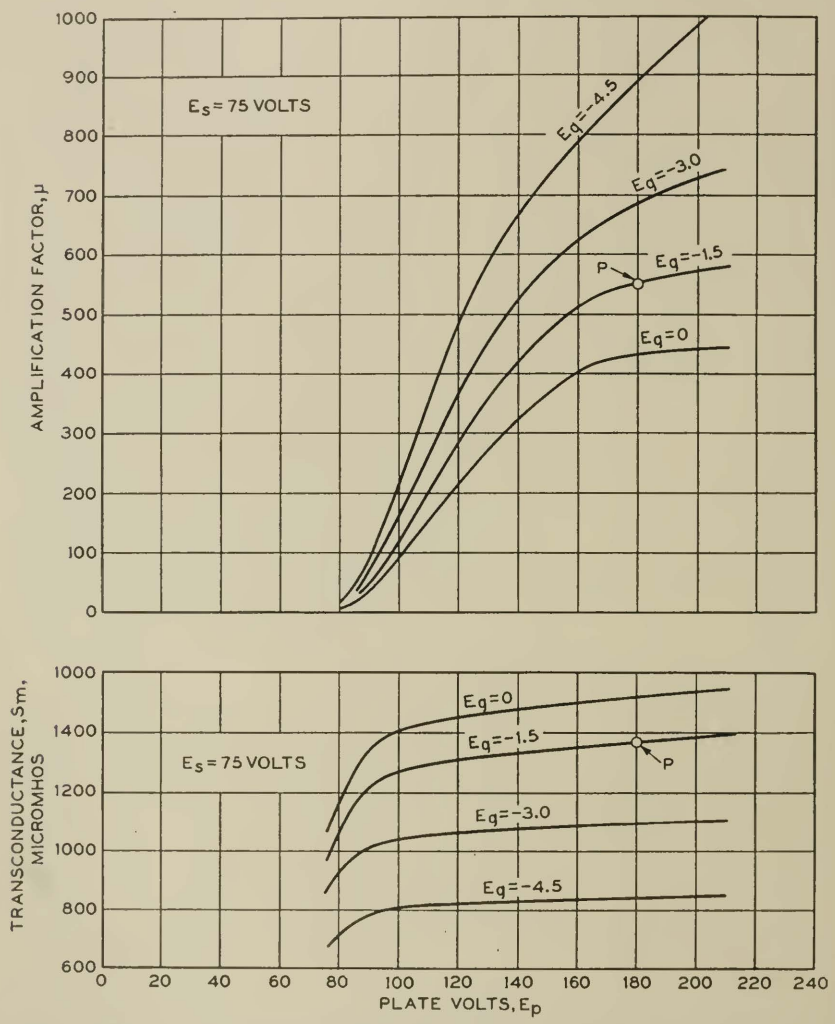


Fig. 5—Amplification factor and transconductance of a screen-grid tetrode as functions of the plate voltage.

voltage amplification ratio approaching the amplification factor. In the operation of screen-grid tubes and conventional pentodes as well, which have high plate resistances, it is usually necessary that the load resistances be much smaller than the plate resistances. This limitation

is imposed by circuit coupling requirements and sometimes by restrictions on the permissible harmonic content in the power output. This limitation results in an amplification ratio that is much smaller than

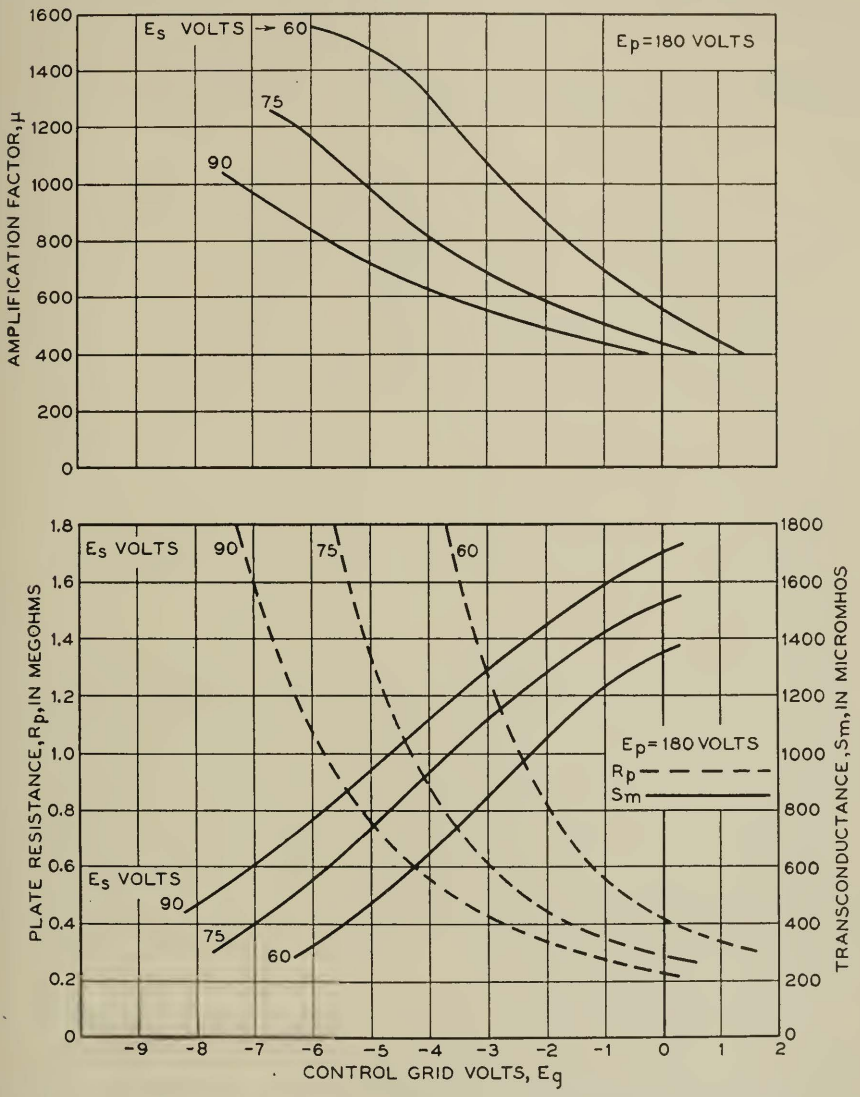


Fig. 6—Amplification factor, transconductance, and plate resistance of a screengrid-tetrode as functions of the control-grid voltage.

the amplification factor; but even so, a considerably higher amplification ratio usually can be obtained than with triodes. This is apparent

from consideration of equation 16 which may be written in the form,

$$A_v = S_m \cdot \frac{R}{1 + \frac{R}{R_p}} \cdot \tag{18}$$

Since in screen-grid tubes and pentodes, R/R_p is small, the voltage amplification is given approximately by

$$A_v = S_m \cdot R. \tag{19}$$

Now, in a comparable triode the transconductance will have a value comparable with S_m, and it may be assumed that a load resistance could be used comparable with R. However, the ratio R/R_p for the triode is not small and may be greater than unity. Hence, from equation 18, the amplification obtained with the triode is correspondingly smaller.

It has been shown that the presence of the electrostatic screen in the screen-grid tetrode results in a high plate resistance and a high amplification factor, but that the ratio of μ to R_p (transconductance) remains normal. Such tubes yield high amplification and, with suitable associated circuits, are relatively free from feed-back; but they are limited in their range of operation because of the fold in the plate current-plate voltage characteristics resulting from secondary emission from the plate.

Power Pentodes

In order to deliver a large power output, a vacuum tube must be capable of large variations in plate current and plate voltage from their normal operating values. Both of these conditions are fulfilled by the power pentode. The arrangement of the electrodes, shown schematically in Fig. 7, corresponds to that in a screen-grid tube except that an additional grid, g_3, is inserted between the plate and screen grid, g_2. As in the screen-grid tube, the first grid, g_1, has a negative voltage applied to it and acts as the control element. The second grid, g_2, is maintained at a fixed positive potential, E_s, and provides the main driving field for the space current.

As in the screen-grid tube, the total space current is determined almost wholly by the geometrical dimensions and spacings of the cathode, g_1 and g_2, and by the voltages applied to these electrodes. Consequently, in the design of this portion of the structure, the same considerations apply as in the design of an ordinary triode to deliver large power output. By making the inner grid comparatively coarse and by designing the second grid to operate at comparatively high potentials, a structure having a low amplification factor is obtained

which draws a large space current from the cathode at a control-grid bias sufficiently negative to permit relatively large swings of the control-grid voltage. In pentodes designed to operate at low frequencies, screening between the plate and control grid is unimportant;

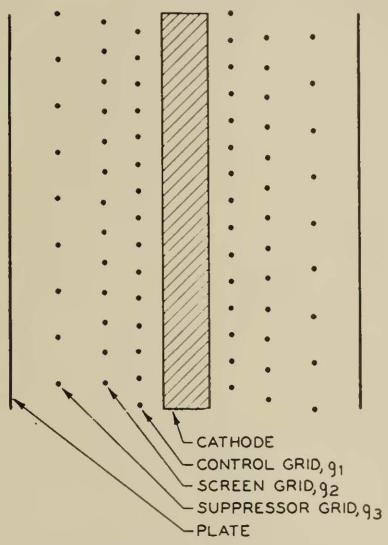


Fig. 7—Schematic diagram showing the arrangement of electrodes in a power pentode.

hence, the second grid can be comparatively coarse, thus permitting as large a portion as possible of the space current to pass through it to the plate.

To permit the largest possible swings in plate voltage, it is necessary to remove the "fold" in the plate-current characteristics, caused by secondary electrons emitted from the plate. This is accomplished by the insertion of a third grid, g_3, between the plate and second grid, g_2. This grid, known as a suppressor grid, must be maintained at a lower potential than the lowest instantaneous potential reached by the plate, and is usually maintained at the cathode potential by connecting it to the cathode inside the tube. The suppressor grid exerts a retarding force on the primary electrons flowing toward it from the cathode, but, because of its coarse structure, all but a small fraction succeed in passing through it and are accelerated again, finally reaching the plate with the same velocity they would have if the suppressor grid were absent. On the other hand, secondary electrons emitted either by the plate or screen grid find themselves in a retarding field, which they are unable to traverse because of their low velocity, and are constrained to return to the electrode from which they came.

Plate current-plate voltage characteristics and screen current-plate voltage characteristics for a power pentode of the indirectly heated cathode type, are shown in Fig. 8. The secondary emission "fold"

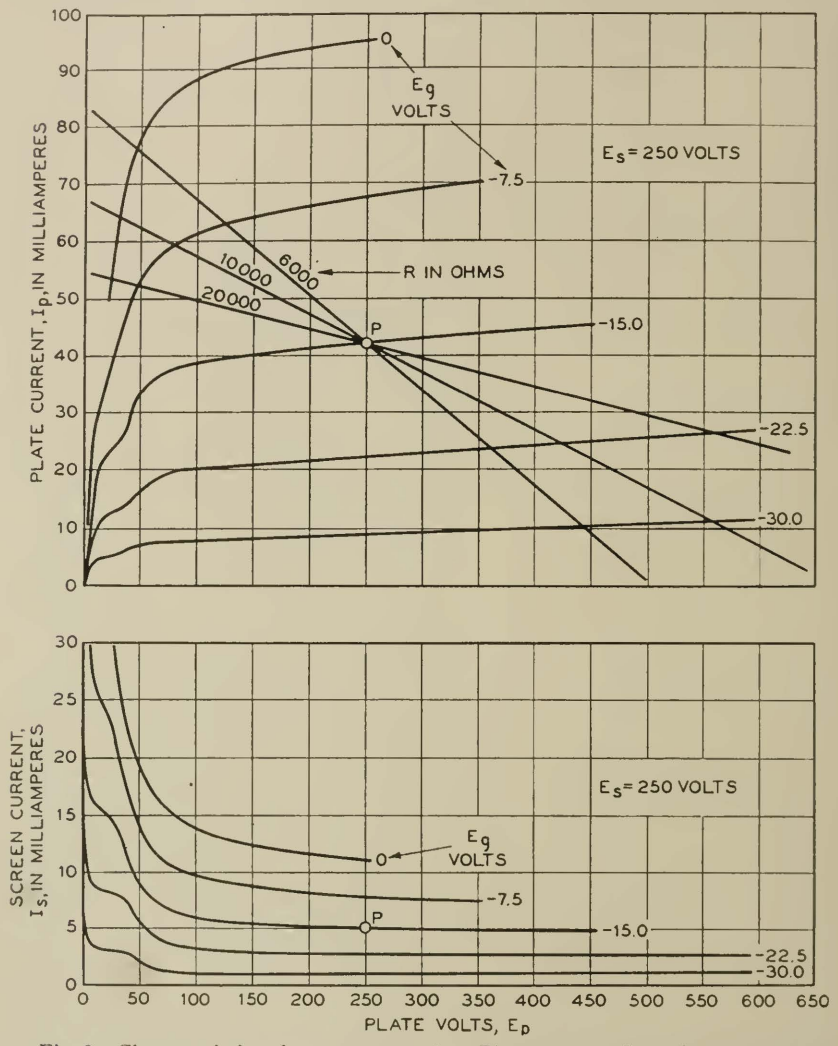


Fig. 8—Characteristics of a power pentode. Plate current, I_p, and screen current, I_s, as functions of the plate voltage with various values of the control-grid voltage, E_g, as parameters. Screen voltage, E_s = 250 volts throughout. Load lines are shown for resistance loads of 6000, 10,000, and 20,000 ohms.

in the characteristics is almost completely eliminated by the suppressor grid.

The effectiveness of the suppressor grid is shown by the charac-

teristics of Fig. 9. These curves were obtained from a tube of the same type as that for which the characteristics are shown in Fig. 8. One set of curves was obtained with the suppressor grid operating in the normal manner. The other curves were obtained with the suppressor grid tied to the screen grid and maintained at a positive

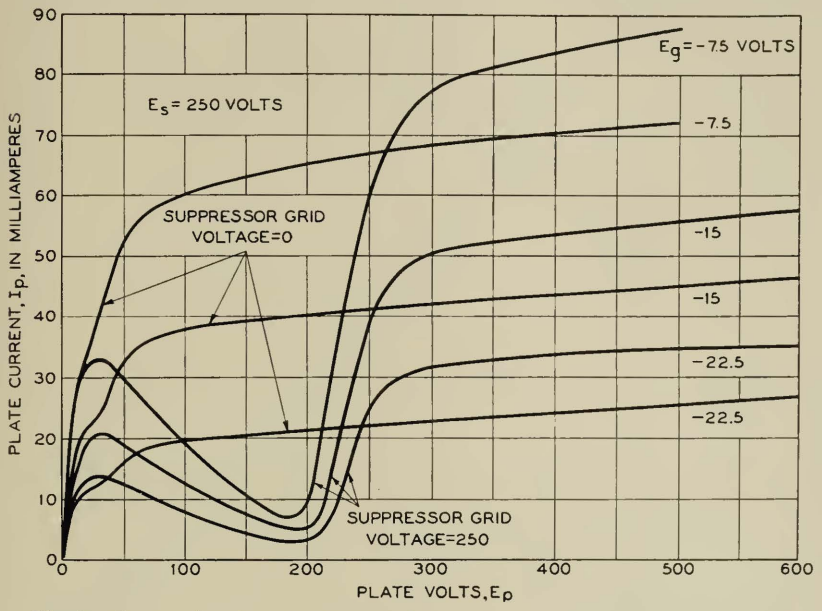


Fig. 9—Curves showing the effect of the suppressor grid in a power pentode.

potential of 250 volts. In this latter case, the number of secondary electrons escaping from the plate is practically the same as if the suppressor grid were removed from the tube. The presence of the suppressor grid not only permits the plate to swing to very much lower potentials than otherwise would be possible, but also permits the plate and screen to operate at the same potential, which is 250 volts in this case.

The characteristics of Fig. 8 closely approach the form that would be expected from simple theory. As is usual with power pentodes, the plate-current curves are not quite as flat as those for screen-grid tubes. This is because of the more open character of the grids which permits the plate to have a slightly greater effect on the magnitude of the space current. This is evidenced also by the tendency of the curves to turn up at the higher plate voltages.

The normal operating point for this tube is at point *P* in Fig. 8, at which the plate and screen potentials are both 250 volts and the control

grid potential is -15 volts. Under these conditions, the average characteristics are: $I_p = 42$ ma, $I_s = 5$ ma, $\mu = 156$, $R_p = 52{,}000$ ohms, and $S_m = 3000$ micromhos.

Curves showing the amplification factor, μ, the plate resistance, R_p, and the transconductance, S_m, for several different values of grid bias, are shown in Fig. 10. They correspond in general form to those

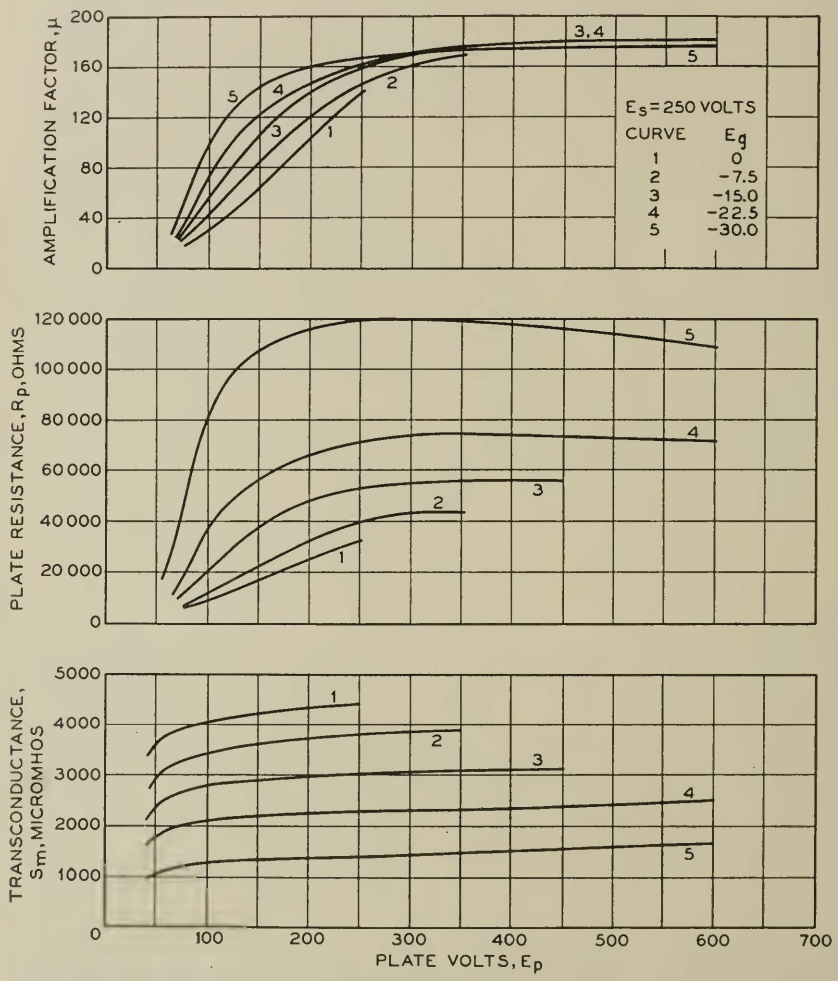


Fig. 10—Amplification factor, plate resistance, and transconductance of a power pentode as functions of the plate voltage, with various values of control-grid voltage as parameters. Screen voltage, E_s, is maintained constant at 250 volts.

shown in Figs. 3 and 5 for the screen-grid tube. The maxima in some of the plate-resistance curves result from the fact that the corre-

sponding plate-current curves turn up at the higher plate potentials and thus have points of inflection.

In Fig. 11, the plate current is shown as a function of control-grid

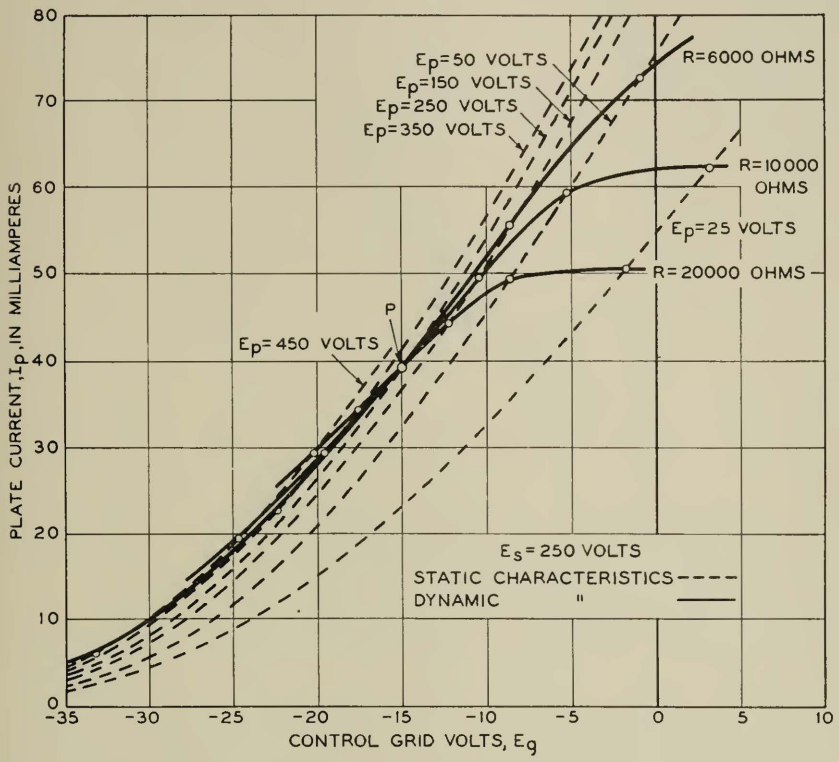


Fig. 11—Plate current-grid voltage characteristics of a power pentode, with dynamic characteristics for resistance loads of 6000, 10,000, and 20,000 ohms.

voltage for various values of the plate voltage, with the screen potential constant at 250 volts. The curves are very similar to corresponding ones for the screen-grid tube shown in Fig. 4. Here, curves are included at such low plate voltages that a large falling off in plate current occurs.

In Fig. 12, the output power in watts, and the second and third harmonics, expressed in decibels below the fundamental, are shown as functions of peak volts for a sinusoidal input applied to the grid. These data were obtained under the normal operating conditions previously given, and with the indicated load resistances. Since the curves are typical in form of those obtained in several types of multi-electrode tubes, it will be of interest to examine them in some detail.

Except at the higher inputs, the power output increases continuously with load resistance over the load range considered. This is to be expected, since the highest load resistance of 12,000 ohms is much

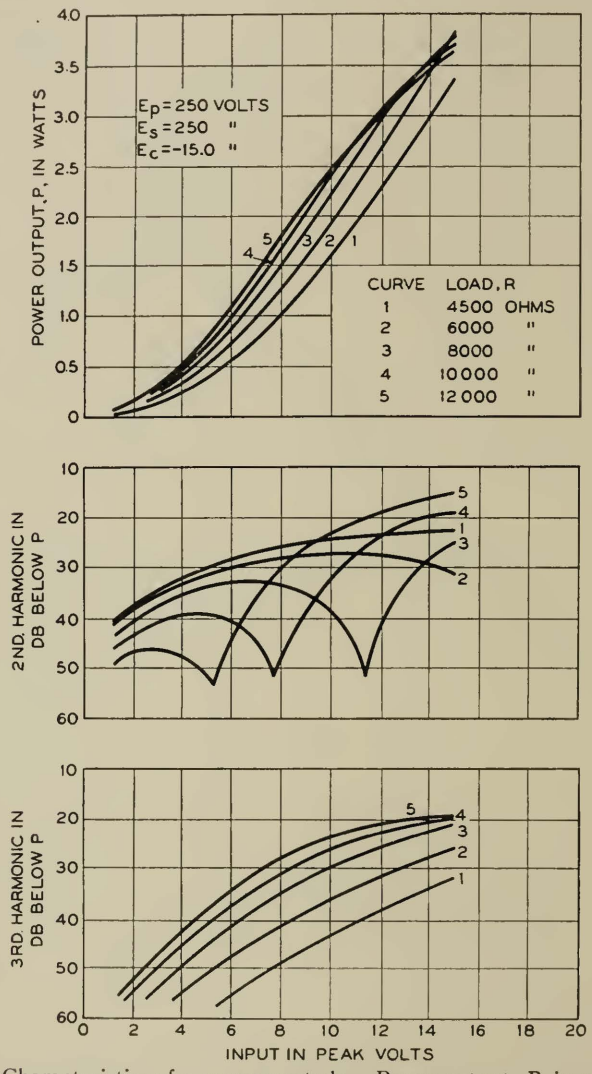


Fig. 12—Characteristics of a power pentode. Power output, P, in watts, second and third harmonics in decibels below the fundamental, as functions of the peak input when a sinusoidal voltage is applied to the control grid.

smaller than the plate resistance of the tube, which is 52,000 ohms. For small inputs, the maximum power output would be obtained, as it is in triodes, when the load resistance is equal to the plate resistance of

the tube. The decrease in power output with increasing load resistance, at the higher inputs, results from the progressive turning over of the dynamic characteristics as shown by the curves of Fig. 11.

The second harmonic decreases continuously with increasing load resistance at small inputs. At large inputs, it decreases at first, then increases with increasing resistance. At very low resistances, the second harmonic increases continuously with input. At higher load resistances, it rises to a broad maximum, then falls to a very sharp minimum, after which it again rises rapidly with increasing input. The explanation of these phenomena will be given later.

The third harmonic increases continuously both with increasing input and with increasing load resistance. At load resistances that are small compared with the plate resistance of the tube, it rises to a higher level than the second, over a certain range of input. Relatively high levels of third harmonic are characteristic of pentodes, screen-grid tubes, and also of some other types of multi-electrode tubes. For example, from the curves of Fig. 12, with a load resistance of 6000 ohms and an input of 15 peak volts, the power output is 3.8 watts with the volume level of the second and third harmonics 31 db and 26 db, respectively, below that of the fundamental. At an input of 10 peak volts, the volume level of the second harmonic rises to 27 db and of the third harmonic falls to 36 db below that of the fundamental.

These results are quite different from those obtained with triodes where the third harmonic is, in general, 10 db or more below the second. Furthermore, in triodes, both the second and third harmonics decrease continuously, as a rule, with increasing load resistance. There are exceptions to this, however, where the third harmonic curves show minimum points or cusps, similar to those shown by the second harmonic curves for pentodes.

The reason for the relatively large harmonic content in the output of pentodes is apparent from consideration of the load lines drawn through the operating point P in Fig. 8. At the lower values of plate voltage, the load lines cut across the rapidly descending portions of the plate-current characteristics. This effect is more marked and begins at more negative values of grid bias as the load resistance increases and the slope of the load lines becomes correspondingly less. The effect of this is to produce current variations through the external load resistance that are not proportional to the variations in grid voltage, thus resulting in distortion of the output.

The character of the distortion is made clearer by reference to the dynamic characteristics of Fig. 11, for resistance loads of 6000, 10,000, and 20,000 ohms. All the curves show a flattening out at the top which

increases progressively with load resistance. It readily is shown that such dynamic characteristics, having points of inflection at which the curvature changes sign, give rise to peculiarities in the harmonic output.

If such a characteristic be expressed by a power series in terms of grid-voltage variations from the operating point P, there usually is found a relatively large contribution by third and higher odd-power terms the coefficients of which are predominantly negative in sign. Since odd-power terms yield odd harmonics, this accounts for the relatively high levels of third harmonic at input voltages sufficiently large that the flat portion of the characteristic is traversed. However, positive and negative signs are about evenly divided among the coefficients of the even-power terms, which yield even harmonics. At some value of the input voltage, which varies with the load resistance, the contributions to the second harmonic by positive and negative terms are approximately equal, resulting in a very small value of this harmonic. This accounts for the cusps in the second harmonic curves of Fig. 12. For inputs less than the value at the cusp, the contribution of positive terms (largely the second-power term) prevails over that of negative terms, while at higher inputs the reverse is true. Consequently, there is a reversal in the phase of the second harmonic at the cusp.

If the point of inflection were at the operating point P in Fig. 11 and, if the dynamic characteristic were symmetrical about P, then only odd-power terms would appear in the equation of the curve. Consequently, in this special case, even harmonics would vanish from the output and only odd harmonics would remain.

Plate-circuit efficiency of pentodes is higher than that usually found in triodes. The underlying physical reasons for this difference are as follows: In the triode, the plate simultaneously performs two functions. First, it is the element in the output circuit whose fluctuating potential is impressed across the load resistance. Second, assuming that the grid potential is not positive at any time, the plate is the only positive electrode providing the necessary driving force for the space current. These two functions militate against each other to a certain extent, for, as is evident from consideration of the dynamic characteristic of either a triode or a pentode, the plate voltage reaches its minimum value at the instant when the plate current reaches its maximum value. This minimum voltage, which must be sufficiently large to draw the peak current through the tube, is a very substantial fraction of the operating plate voltage, particularly when the latter is comparatively low.

In pentodes and in some other multi-electrode tubes, the positive grid, maintained at a constant voltage, provides the necessary driving

force for the space current, thus relieving the plate from performing this function. Consequently, the plate is free to swing to lower voltages than otherwise would be possible, which results in a corresponding increase in efficiency.

In Table I, similar data are shown for typical pentodes and triodes.

TABLE I

COMPARATIVE DATA FOR TYPICAL TRIODES AND PENTODES

	Triodes			Pentodes		
	A	*B*	*C*	*D*	*E*	*F*
Plate voltage, E_p	200	250	325	180	250	250
Screen voltage, E_s				180	250	250
Grid voltage, E_g	−45	−50	−68	−18	−16.5	−15
Plate current, I_p, ma	45	34	60	14.5	34	42
Screen current, I_s, ma				2.8	6.5	5
Transconductance, S_m, micromhos	2,810	2,175	5,200	1,050	2,200	3,000
Amplification factor, μ	2.9	3.5	3.8	105	220	156
Plate resistance, R_p, ohms	1,030	1,610	730	100,000	100,000	52,000
Load resistance, R, ohms	2,060	3,900	2,750	12,000	7,000	6,000
Input, peak volts	45	50	68	18	16.5	15
Power output, watts	1.9	1.9	7.8	1.2	3.0	3.8
Second harmonic, %	7.4	5.7	5.0	5.0	2.3	2.8
Third harmonic, %	0.7	1.2	1.1	7.5	6.6	5.0
Total harmonic, effective %	7.45	5.9	5.1	9.0	7.0	5.7
Plate efficiency, %	21	22	40	46	35	36
Plate-grid capacitance, $\mu\mu$f	12	7.2	20	1.2	0.8	0.2
Plate-ground capacitance, $\mu\mu$f	3.2	3.0	5.5	7.8	8.5	15
Grid-ground capacitance, $\mu\mu$f	6.8	4.5	9	6.4	8.4	9

While the triodes and pentodes are not directly comparable with each other, the data are indicative of the differences between the two types of tubes. The chief differences between pentodes and triodes may be summarized as follows: (1) Pentodes yield higher gain and require correspondingly lower input voltages to drive them. (2) Pentodes have much higher amplification factors and plate resistances than triodes. The latter constitutes a handicap in coupling the tube to its circuit, particularly if transformer coupling is employed. (3) Pentodes yield high power output, generally at higher plate efficiency than triodes. (4) The harmonic content in the output of pentodes is high, the third harmonic being particularly high compared with its level in triodes. This requires that pentodes work into load impedances that are very low compared with the plate resistance. Ratios of R to R_p of 1/3 to 1/10 are common. (5) The plate-grid capacitance is much lower and the plate-ground capacitance is somewhat higher in pentodes than in comparable triodes.

Space-Charge-Grid Pentodes

One limitation in the three-electrode tube and in the multi-electrode tubes considered thus far in this paper, is the resistance offered by space charge to the flow of space current. Tubes having so-called space-charge grids overcome this limitation to some extent by having a positive grid close to the cathode, which partially neutralizes the negative field very near the cathode surface due to space charge. A comparatively large current is drawn from the cathode by the space-charge grid. A portion of this current (usually about half of it) is collected by this grid, while the remaining portion passes through it and is acted on by the remaining elements of the tube.

The arrangement of the electrodes in a space-charge-grid pentode is shown in Fig. 13. The space-charge grid, g_1, is maintained at a

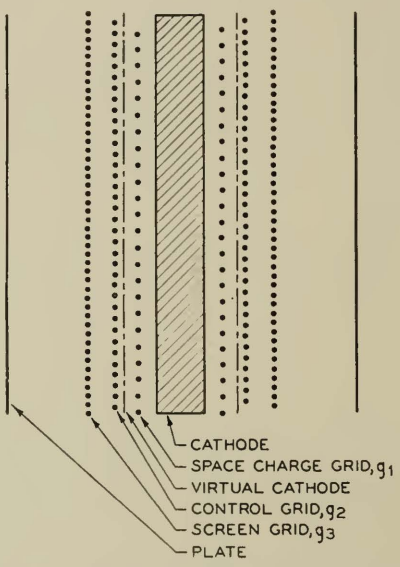


Fig. 13—Schematic diagram showing the arrangement of electrodes in a space-charge-grid pentode.

relatively low positive potential with respect to the cathode, usually in the range from 10 to 20 volts. The second grid, g_2, is the control grid and is maintained at a negative potential with respect to the cathode.

Ideally, at some cylindrical surface between these two grids (assuming the structure to be cylindrical) the electrons are retarded to nearly zero velocity forming a second space-charge region which may be regarded as a virtual cathode. Since this space-charge sheath is larger in area than the original cathode and is very close to the control grid, it

results in very large values of transconductance. Practically, the ideal condition is not fully realized, largely because velocity components other than radial are imparted to the electrons in passing through the space-charge grid. Consequently, these electrons reach any given cylindrical surface outside the space-charge grid with rather widely varying radial components of velocity. This, as will be seen later, places a rather serious limitation on the performance of such tubes.

The arrangement and functioning of the other electrodes in Fig. 13, outward from the virtual cathode, correspond with that of the screen-grid tetrode. The screen grid, g_3, is maintained at a fixed positive potential necessary to accelerate the electrons from the region of the virtual cathode. The plate must be maintained at a potential higher than that of the screen for the same reason as in the screen-grid tetrode. It will be shown that the characteristics of this pentode correspond roughly with those of the screen-grid tube previously discussed.

If g_3 were omitted, the structure outside the virtual cathode would correspond to that of a triode and the characteristics in the resulting tetrode would correspond roughly to those of a triode. In Fig. 14, characteristics are shown for a pentode of this type, the cathode and general dimensions of which are the same as those of the power pentode, the characteristics of which were shown previously. The plate-current and screen-current curves are seen to correspond very closely with those previously shown for a screen-grid tube. The characteristics exhibit the same "folds" due to secondary electrons, although this portion of the characteristics is not shown. The net or space-charge-grid current, I_n, increases as the plate current decreases with increasing negative control-grid voltage. This is to be expected since, as the control grid becomes more negative and reduces the current passing through it to the plate, the excess current returns to the net rather than to the cathode as in the screen-grid tube.

If values of the amplification factor, plate resistance, and transconductance are plotted as functions of the plate voltage, families of curves are obtained similar in all respects to those for a screen-grid tube as shown in Figs. 3 and 5. These characteristics are not shown for this tube.

In Fig. 15, plate-current, screen-current, and net-current characteristics are shown as functions of the control-grid voltage with different values of the screen voltage as parameters. As in the case of the screen-grid tube, only a slight displacement of the characteristics results from variation of the plate voltage. It is of interest to note the high values of net current, particularly as the plate current drops toward zero with increasing negative grid bias. For example, if the operating point is

chosen at P in Fig. 15, with a screen potential of 90 volts and a grid potential of -2 volts, the plate current, screen current, and net current are 24, 0.5, and 23 ma, respectively, making a total space

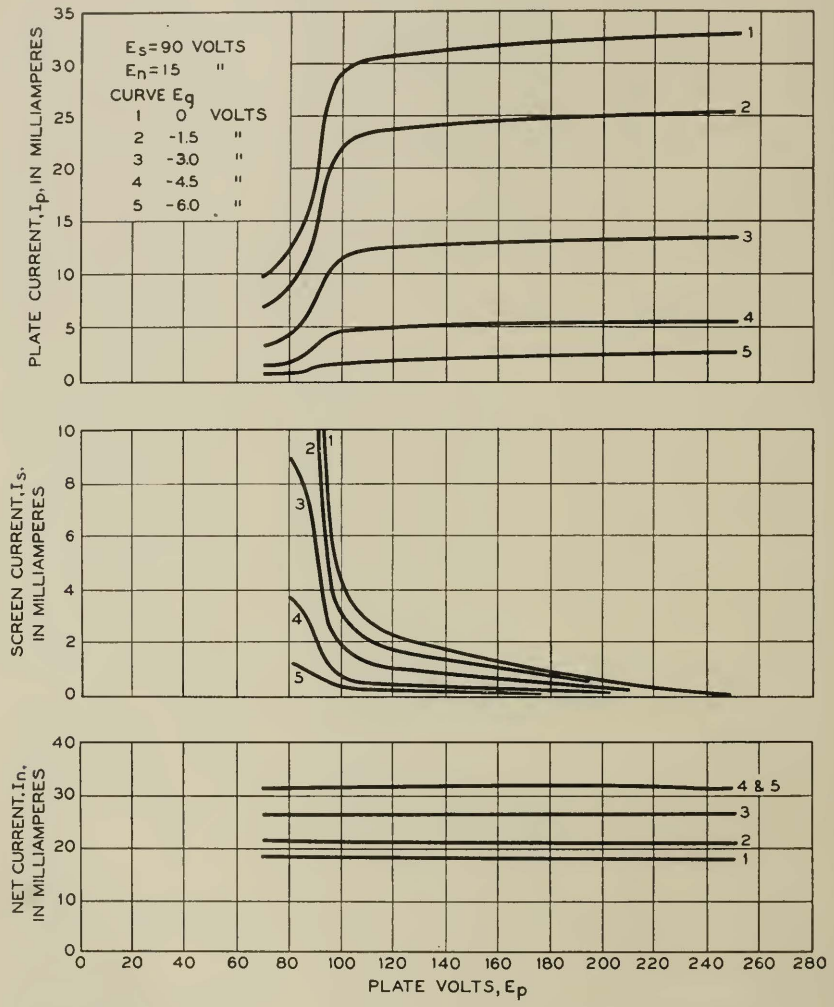


Fig. 14—Characteristics of a space-charge-grid pentode. Plate current, I_p, screen current, I_s, and net (space-charge-grid) current, I_n, as functions of the plate voltage with the indicated values of control-grid voltage as parameters. Screen-grid voltage, $E_s = 90$ volts. Net (space-charge-grid) voltage, $E_n = 15$ volts.

current of 47.5 ma drawn from the cathode, nearly half of which is collected by the space-charge grid.

Another point of interest in Fig. 15 is the flattening out of the plate-current characteristics at the higher values of plate current, in a manner

very similar to that exhibited by triodes of low cathode emission. In fact, this phenomenon is caused by the partial exhaustion of the space charge in the region of the virtual cathode. This is a fundamental characteristic of space-charge-grid tubes, caused largely by the

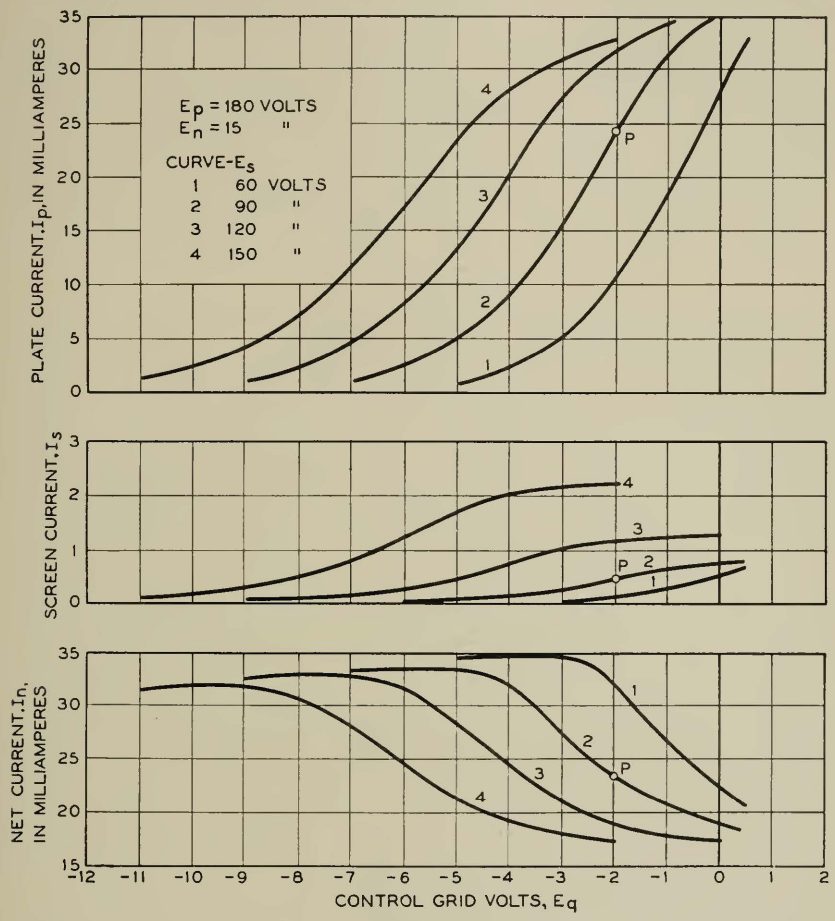


Fig. 15—Characteristics of a space-charge-grid pentode. Plate current, screen current, and net current as functions of the control-grid voltage, E_g, with the indicated values of screen voltage as parameters. Plate voltage, E_p = 180 volts. Net voltage, E_n = 15 volts.

imperfect character of the virtual cathode, which results from electrons entering the space-charge region with widely varying normal components of velocity. It constitutes a serious limitation on the practicable range of operation of such tubes, since it prevents the plate potential from swinging over a sufficiently large range to obtain a large power output without resulting in prohibitive distortion.

In Fig. 16, the amplification factor, plate resistance, and transconductance are shown as functions of control-grid voltage, with the same parameters as were used in obtaining the curves of Fig. 15. Of particular interest in these curves, are the unusually high values of

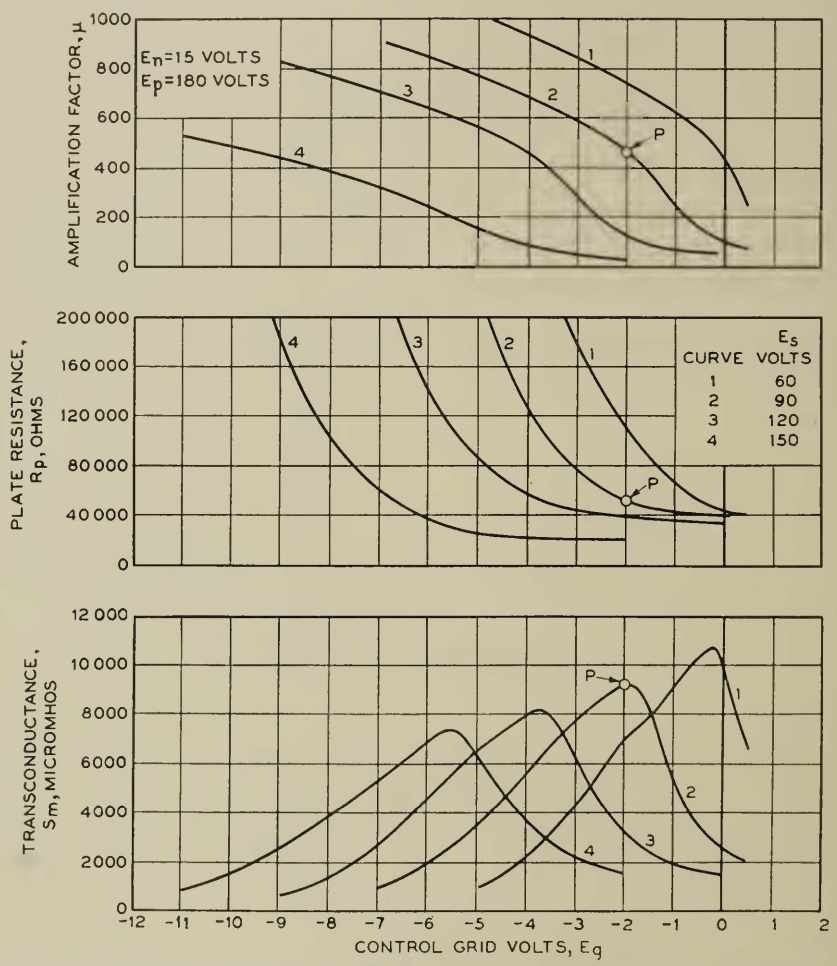


Fig. 16—Amplification factor, plate resistance, and transconductance of a space-charge-grid pentode as functions of the control-grid voltage.

transconductance for a tube of this size. For example, at the selected operating point P it is 9200 micromhos. Also of interest is the peaked character of the transconductance curves. Since transconductance is defined by the slope of the plate current-grid voltage curves, which for this tube have points of inflection in them, maximum points must occur

in the transconductance curves. The rapid falling off in transconductance on either side of the maxima, with change in grid voltage, also indicates that a large amount of distortion would result if the tubes were operated with large swings in grid potential.

Space-charge-grid pentodes are capable, then, of yielding very high amplification because of their extraordinarily high transconductance; but, practically, they are limited to use as voltage amplifiers for fairly small inputs, since operation over the wide range necessary for large power output results in prohibitive distortion. These statements apply also to space-charge-grid tetrodes. These tubes have the advantage over ordinary triodes, however, of yielding high values of transconductance at comparatively low voltages applied to the plate.

Co-Planar-Grid Tubes *

Thus far only multi-electrode tubes have been considered in which the electrodes are arranged in concentric order one about another. In the case of the power pentode it has been shown that this arrangement results in characteristics yielding comparatively high output power at high amplification and efficiency but with relatively large percentages of harmonics. It has also been pointed out that the high plate-circuit efficiency is largely due to the presence of a positive grid held at a fixed potential, which permits larger plate voltage swings than would otherwise be possible.

It is evident also from the underlying physical principles that in such a structure, with three concentric grids between the plate and cathode, even though the lateral spacings are fairly wide, enough shielding is interposed between the plate and cathode to make the plate resistance high. The question then arises: Is there any possible way of obtaining the advantages of a positive grid in tubes designed to operate at comparatively low plate voltages without the accompanying high plate resistance of the pentode structure?

This objective is accomplished reasonably well by a four-electrode structure in which the positive grid is placed in the same plane as the control grid. The arrangement of the electrodes in such a structure, which will be referred to as a co-planar-grid tube, is shown schematically in Fig. 17. The lateral wires of the two grids are asymmetrically arranged, the positive and negative lateral wires being arranged in pairs close together. The object of this arrangement may be described roughly as follows: By such an arrangement, when the control grid is very negative, the field about each negative wire very largely neutral-

* This section, "Co-Planar Grid Tubes," is an addition to the material published in *Electrical Engineering*.

izes the field due to its positive companion so that the latter becomes ineffective in drawing electrons away from the cathode. As the negative grid approaches zero potential it uncovers the positive grid electrically so that the latter becomes highly effective in drawing a large space current from the cathode at the moment when the peak current is drawn to the plate. This would not be so effectively accomplished if the lateral wires were symmetrically arranged.

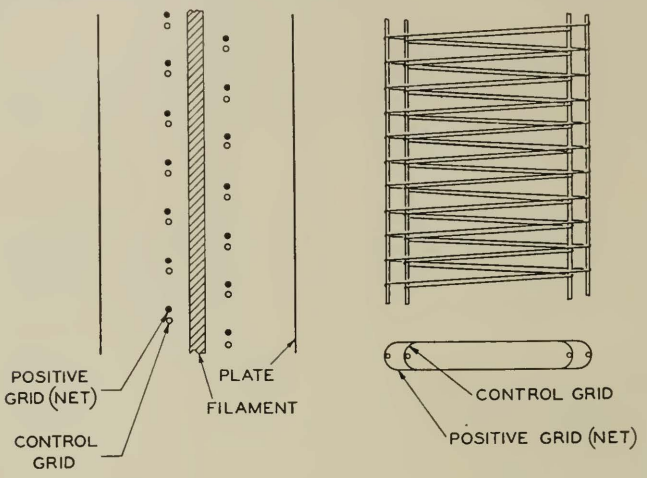


Fig. 17—Schematic diagram showing the arrangement of electrodes in a co-planar-grid tube.

In Fig. 18, characteristic curves are shown for a co-planar-grid tube designed to operate at a plate potential of 130 volts and at a plate current of 35 milliamperes. Each of the three groups of curves shows the plate current as a function of the control-grid voltage for plate voltages of 100, 130 and 160 volts, and with the positive-grid voltage maintained constant at the indicated value. It is evident from these curves that the primary effect of increasing the potential of the positive grid, or net, is to translate the group of plate-current curves to the left. By so doing it is obvious that larger control-grid swings are possible, at any given operating plate current, without the potential of the control-grid becoming positive at any time. This results in a comparatively large power output, to obtain which large grid swings are essential.

The lower curve in Fig. 18 shows that a nearly linear relation exists between the positive and negative increments in voltage that must be applied to the positive and negative grids, respectively, to maintain a constant operating current, in this case 35 milliamperes.

In Fig. 19, a more extended family of characteristics is shown for a positive-net potential of 65 volts. Across these static characteristics a dynamic load characteristic is drawn for a resistance load of 3400 ohms, which matches the plate resistance of the tube at the operating point

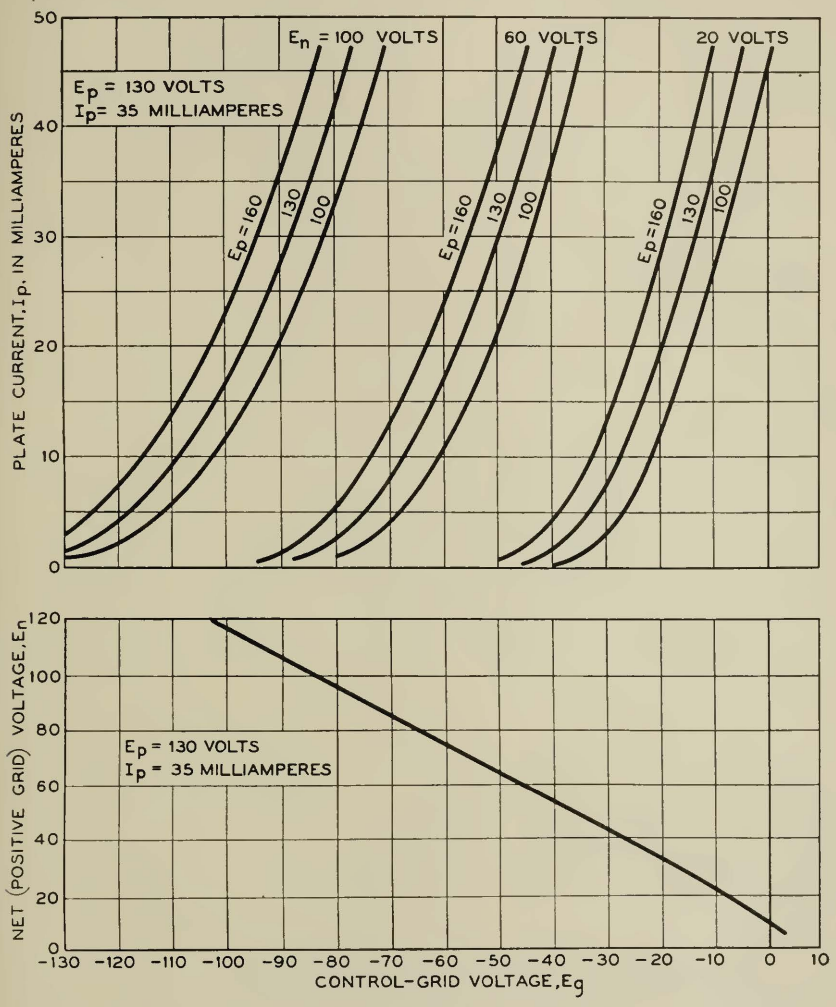


Fig. 18—Plate current-control grid voltage characteristics of a co-planar-grid tube with various values of plate voltage and net (positive-grid) voltage as parameters.

P. At this point E_p = 130 volts, E_n = + 65 volts, E_g = − 51 volts, I_p = 35 milliamperes, and μ = 5.3. The turning over of the dynamic characteristic at the top is due to the increasing fraction of the total

space current collected by the positive net. The magnitude of the net current in this region is shown by the family of curves in the lower part of Fig. 19.

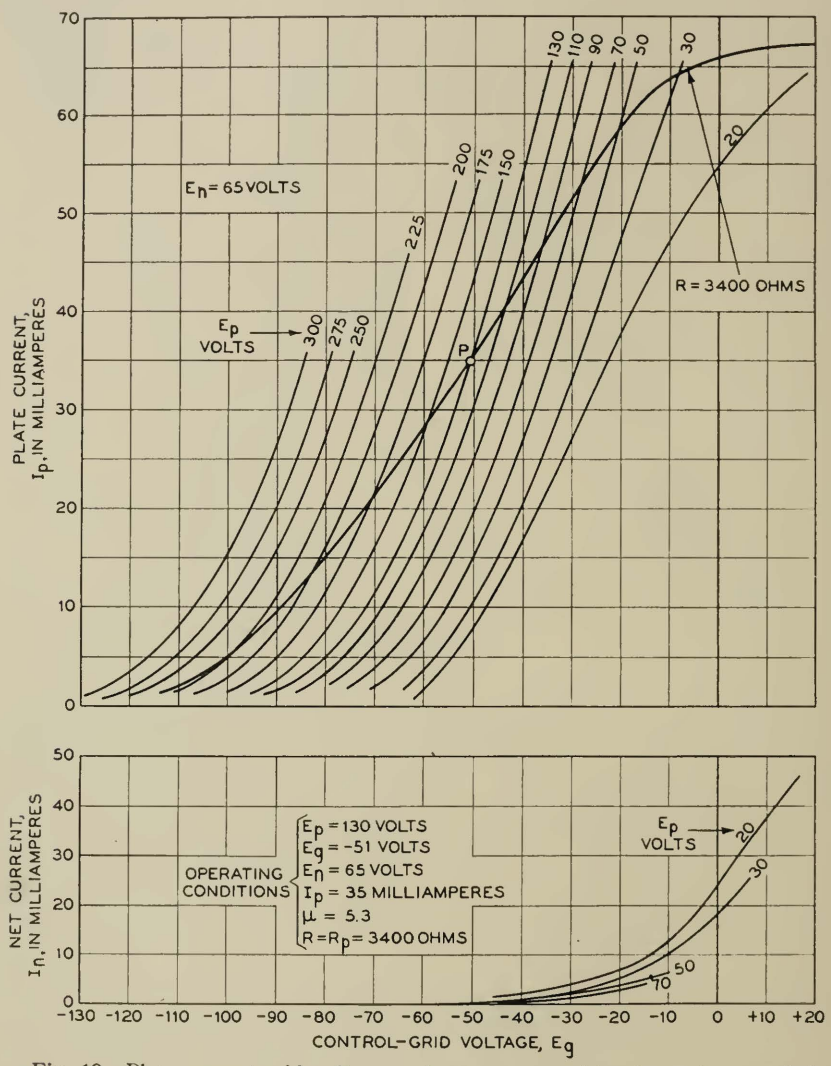


Fig. 19—Plate current-grid voltage and net current-grid voltage characteristics of a co-planar-grid tube. The net (positive-grid) potential is maintained constant at 65 volts. A dynamic characteristic is shown for a resistance load of 3400 ohms.

In Fig. 20, curves are shown giving the power output, and second and third harmonics expressed in decibels below the fundamental, as functions of the peak input when a sinusoidal voltage is applied to the

control grid. The curves were obtained with the indicated values of load resistance, and with the indicated bias voltages applied to the grids. In each case the grid voltages were adjusted to give a plate current of 35 milliamperes. The curves are very similar in form to

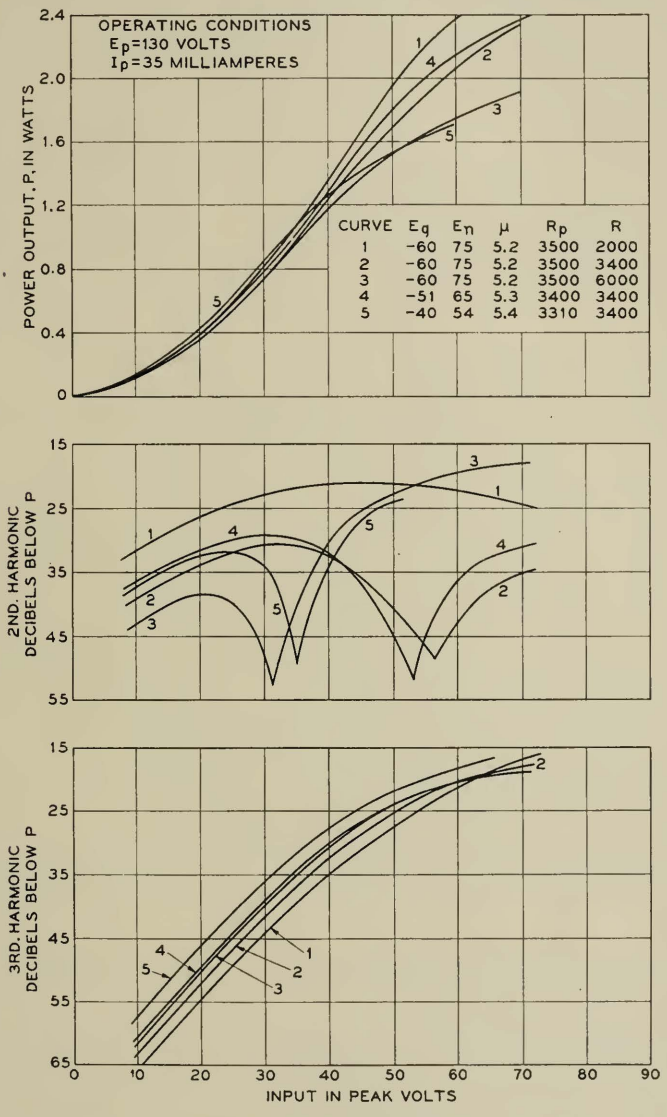


Fig. 20—Co-planar grid tube.

Power output, P, in watts, second and third harmonics in db below the fundamental, as functions of peak input when a sinusoidal voltage is applied to the control grid.

those for the power pentode shown in Fig. 12. For the co-planar-grid tube, as for the pentode, the third harmonic is comparatively high and increases with the load resistance, while the second harmonic curves exhibit minima. This is to be expected from the form of the dynamic characteristic which is very similar to that of similar curves for the pentode previously discussed. For example, consider the curves marked 2 in Fig. 20, obtained with $E_p = 130$ volts, $E_n = 75$ volts and

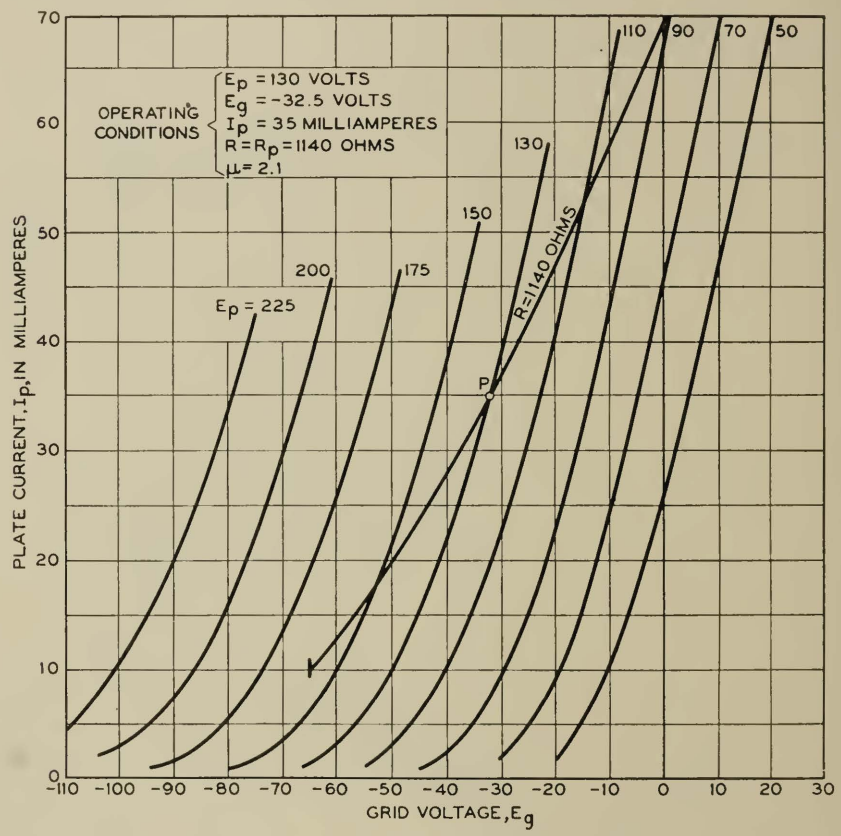


Fig. 21—Plate current-grid voltage characteristics of a triode with a dynamic characteristic shown for a resistance load, $R = R_p = 1140$ ohms.

$E_g = -60$ volts. The output power under these conditions, with an input of 50 peak volts, is 1.68 watts with the second and third harmonics 40 db and 25 db, respectively, below the fundamental. At an input of 30 peak volts, the level of the second harmonic rises to 29 db and that of the third harmonic falls to 41 db below that of the fundamental.

It is of interest to compare the characteristics of this co-planar-grid tube with those of a three-electrode tube designed also to operate at a plate potential of 130 volts and at a plate current of 35 milliamperes. The plate current-grid voltage characteristics of such a triode are shown in Fig. 21. The dynamic characteristic is for a load resistance of 1140 ohms which matches the plate resistance of the tube at the operating point P. At zero grid bias on the dynamic characteristic, the plate potential is 92 volts, a drop of only 38 volts from the operating

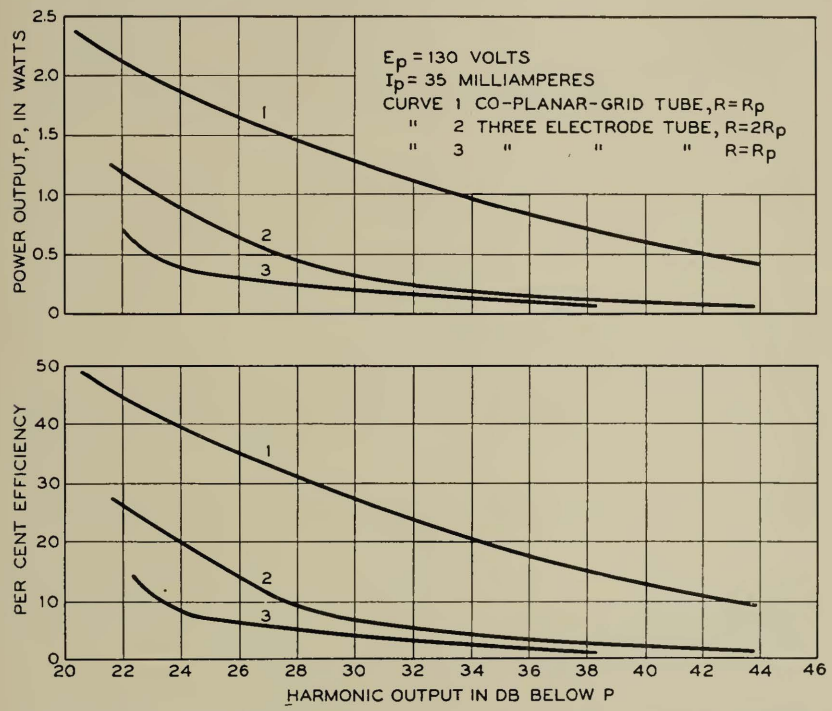


Fig. 22—Curves showing the output power and plate efficiency of a triode and a co-planar-grid tube with varying effective values of second and third harmonics expressed in db below the fundamental.

potential of 130 volts. In the co-planar-grid tube the dynamic characteristic cuts the axis of zero grid bias at a plate potential of 27 volts, a drop of 103 volts from the operating plate potential. The larger plate-voltage variation in the co-planar-grid tube results in higher efficiency than in the triode. The reasons for this difference are identical with those previously discussed in the case of the power pentode.

In Fig. 22, curves are shown comparing the power output and plate-

circuit efficiency of the triode and co-planar-grid tube at varying harmonic levels expressed in db below the fundamental. Curves are shown for the triode with the load resistance, R, equal to the plate resistance, R_p, and also with $R = 2R_p$. In both tubes the plate voltage is maintained at 130 volts, and the plate current at 35 milliamperes.

With the harmonics 26 db below the fundamental (5 per cent) and with $R = 2R_p$, the output from the triode is 0.65 watt. At the same harmonic level the output from the co-planar-grid tube is 1.65 watts with a resistance load matching its plate resistance. The efficiencies are 14 per cent and 35 per cent respectively.

The most significant differences between the co-planar-grid tube and the power pentode are as follows: (1) The former has a much lower plate resistance which simplifies coupling the tube to its output circuit and is particularly advantageous in applications where it is desirable to operate the tube into an impedance matching its plate resistance. (2) The pentode yields higher amplification and, consequently, requires a correspondingly smaller input voltage to drive it. (3) The plate-circuit efficiency is not greatly different except at low plate voltages where the co-planar-grid tube is somewhat more efficient.

It should be borne in mind in considering the efficiency of output power tubes used at audio frequencies, that the usual criterion of available power is the harmonic content. As has been shown, fundamental physical factors tend to make the harmonics, particularly the odd harmonics, relatively high in pentodes and co-planar-grid tubes. On the other hand, the recent trend in the design of triodes has been toward tubes of very low plate resistance, which operate into comparatively large load resistances. This tends to decrease the harmonics on the one hand and increase the efficiency on the other. The net result has been that triodes of more recent design, particularly those operating at 250 volts or higher, are not far below the pentode or co-planar-grid tube in efficiency.

Limits to Amplification *

By J. B. JOHNSON and F. B. LLEWELLYN

The amplification obtainable in a vacuum tube amplifier is limited by the noise in the circuit. Of the various sources of noise the most fundamental and inevitable is thermal agitation of electricity. Other sources are the influence of ions and of shot effect and flicker effect on the current in vacuum tubes, poor contacts, mechanical vibration, and hum from a.-c. cathode heating. These noises and their effects in limiting amplification are discussed in this paper. Although the natural noise level of an amplifier is exceedingly low, modern amplifiers have reached such a stage of perfection that their noise levels often are practically at the natural limit.

NOISE in amplifiers is now a familiar term. Any one who has had his favorite radio hour ruined by static knows the effect of an incoherent background of noise. Although static was one of the first noises observed in radio amplifiers, its origin is really outside the circuits. At one time it seemed that other sources of noise of purely local origin, such as poor batteries, loose contacts, gassy tubes and induction from power lines, might be eliminated entirely so that the circuits would be capable of amplifying any signal, no matter how small. It was found, however, that the noise level cannot be lowered indefinitely; that there are limits below which, in the nature of things, noise cannot be reduced.

Of the sources of noise, the most fundamental and inevitable is thermal agitation of electricity. In a perfect amplifier all other noises would be reduced to a level below that of thermal agitation. Next in order comes the influence of ions and of shot effect and flicker effect on the current in the vacuum tubes. Under control to a greater extent, but nevertheless of a malignant nature, are the effects of poor contacts, mechanical vibration, and hum from a.-c. cathode heating. In dealing with these disturbances, the circuit and vacuum tube of the first stage of the amplifier are the most important, for here the signal being amplified is at its lowest level. When the signal is so faint that it is masked by the noise remaining as the natural limit of the circuit, then the only possible remedy is to raise the signal level.

The natural noise level is exceedingly low, yet modern amplifiers have reached such a stage of perfection that their noise levels often are practically at the natural limit. This is true not only of special amplifiers built for experimental purposes, but of many amplifiers

* Published in November 1934 issue of *Electrical Engineering*. Scheduled for presentation at Winter Convention of *A. I. E. E.*, New York City, January 22–25, 1935

used in commercial circuits. The natural limits to amplification which will be discussed in this review are therefore of very practical interest.

Thermal Agitation [1, 2, 3, 4, 5, 6]

The free charge of any conductor is in random motion in equilibrium with the thermal motion of the molecules of the conductor and this flow of charge creates a random voltage across the terminals of the conductor. This voltage usually is observed in a system composed of an amplifier with an input circuit and an output device. Its mean-square value across the output device is given by the expression

$$V_{to}^2 = 4kT \int_0^\infty RG^2 df, \tag{1}$$

where the symbols have the following meanings:

k is the Boltzmann gas constant and is equal to 1.37×10^{-23} watt-second per degree,
T represents absolute temperature, degrees Kelvin,
R represents the resistive component in ohms of the input impedance as measured across the input terminals of the amplifier,
G represents the voltage gain of the amplifier, and is equal to the ratio of voltage across the output device to voltage across the input terminals of the amplifier,
f represents frequency in cycles per second,
R and G are in general functions of frequency.

In the simple case where the amplifier has a constant gain over a frequency range F and no gain outside of this range, and where R is also constant over the same frequency range and is at the normal temperature of 300 degrees, the mean-square noise voltage across the input terminals of the amplifier is

$$V_t^2 = 1.64 \times 10^{-20} RF. \tag{2}$$

This is the voltage that would be produced by a generator supplying to the resistance R the power

$$W = \frac{V_t^2}{R} = 1.64 \times 10^{-20} F. \tag{3}$$

The power W, sometimes expressed as 1.64×10^{-20} watts per cycle, is independent of R and may be regarded as the apparent input power of the thermal agitation. It depends only on the frequency range of the amplifier, since the temperature cannot be varied conveniently or

[1] For all numbered references see bibliography at end of paper.

very effectively and it sets a lower limit to the possibility of amplifying electrical impulses of any kind. Any signal much smaller than the thermal noise would be masked hopelessly. The only factor under control in the noise equation is the frequency range F, which should be no greater than is needed for the transmission of the signal.

An example will illustrate the magnitude involved in this limit to amplification. When the signal is speech requiring a frequency band of 6,000 cycles, then the apparent power generated at the input of the amplifier by thermal agitation is 0.985×10^{-16} watts, which is about 138 db below the common reference level of 0.006 watts. (The level of 10^{-16} instead of 0.006 watts is being considered as a reference point for the decibel scale in communication circuits. This is approximately the level of thermal noise in a 6,000-cycle channel.) If the input resistance were one megohm the corresponding r.m.s. noise voltage would be 9.94 μv.

A signal represents a certain amount of available power, and when this is so small that it is near the thermal noise level it must be used efficiently to produce voltage at the grid of the amplifier tube.[7, 8] Let the signal be supplied by a generator of voltage E and internal resistance R_1 which delivers power to a load resistance R_2, the combination forming the input circuit of the amplifier as shown in Fig. 1. The mean-square signal voltage on the grid of the amplifier tube is

$$V_{sg}^2 = \frac{E^2}{R_1^2}\left(\frac{R_1R_2}{R_1 + R_2}\right)^2. \tag{4}$$

However, the resistance required by equation 2 for the noise is the

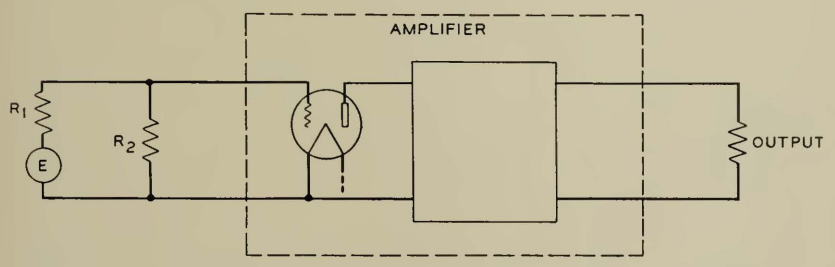


Fig. 1—Schematic diagram of a vacuum tube amplifier showing equivalent input circuit.

combination of R_1 and R_2 in parallel, so that the mean-square noise voltage on the grid of the amplifier tube is, from equations 2 and 3

$$V_n^2 = WR = W\left(\frac{R_1R_2}{R_1 + R_2}\right). \tag{5}$$

Hence the signal-to-noise ratio is

$$\frac{V_{s0}^2}{V_n^2} = \frac{E^2}{WR_1}\left(\frac{R_2}{R_1 + R_2}\right). \tag{6}$$

In general, the internal resistance R_1 of the signal generator is fixed, so that R_2 is the only available variable. In the usual case of matched impedances where R_1 and R_2 are equal, the signal-to-noise ratio is 3 db poorer than in the ideal case where R_2 is made very large compared with R_1. This is one of the few examples where a mismatch of impedances is advantageous. The use of an ideal step-up transformer between R_1 and R_2 in Fig. 1 will be of no avail, so far as the thermal noise is concerned, because its effect in equation 6 will be only to replace R_2 by R_2/N^2 where N is the turns ratio of the transformer.

In some systems the impedance at the input of the amplifier is unavoidably small. It may be so small that the voltage of the thermal agitation of the input circuit, even when amplified by the first tube, is lower than the noise voltage originating in the output circuit of that tube. Ideally, the noise in the plate circuit also should be caused by thermal agitation only, and the equations for it have been derived.[9, 10] In practice, however, the noise in the plate circuit is found to be considerably greater, for reasons that will be discussed presently.

Shot Effect and Flicker Effect Without Space Charge

Early in the study of noise arising in vacuum tubes it was shown [1, 11] that under certain conditions a noise is produced that depends on the fact that the electric current is a flow of discrete particles, the electrons, which are emitted from the cathode in a random manner. The random electron emission produces a statistical fluctuation in the current that flows through the tube and coupling impedance. This fluctuation, called the shot effect (German Schroteffekt, in analogy with the random scattering of shot from a shot gun), appears as noise in the output of the amplifier. When the current in the tube is limited by the rate of emission of electrons rather than by the space charge, so that the resistance of the tube is nearly infinite, then the shot effect produces a mean-square voltage across the output impedance of the amplifier given by [12, 9]

$$V_{s0}^2 = 2\epsilon i \int_0^\infty Z^2 G^2 df, \tag{7}$$

in which

ϵ represents the charge on the electron and is equal to 1.59×10^{-19} coulomb,

i represents space current in amperes,
Z represents the magnitude in ohms of the coupling impedance,
G represents the voltage gain of the amplifier from Z to the output,
f represents frequency in cycles per second,

For an amplifier having a flat frequency response curve over frequency range F the expression becomes, for the effective shot noise across the impedance Z

$$V_s^2 = 31.8 \times 10^{-20} i Z^2 F. \tag{8}$$

The expression holds quite accurately for tubes in which the cathode is made of either clean or thoriated tungsten and for high-vacuum photo-electric tubes, and it has been used in determining the charge on the electron.[13]

When an oxide coated cathode is used, fluctuations of a larger magnitude [14] are superimposed on the true shot effect. These fluctuations are inappreciable above about 10 k.c., but increase rapidly in magnitude toward the lower frequencies. They also increase with the current at a faster rate than the shot effect fluctuations. This disturbance has been ascribed to a state of flux and change in the activating material on the surface of the cathode,[14, 15] and the phenomenon has been called the "flicker effect" (from the analogy of a flickering candle).

There are two practical circuits in which the pure shot effect may set the ultimate noise level. One of these is the circuit in which the grid of an amplifying tube is left "floating" at its equilibrium potential as usually is done in the first stage of amplifiers used for ion counters and other instruments for measuring very small charges.[18, 10] The grid then emits a few electrons and receives positive ions and electrons from the surrounding space. These currents are very small, but are not subject to space charge limitation so far as the grid is concerned. Because the grid impedance is very high, the shot voltage developed by the small grid current may exceed the thermal voltage of the grid impedance. The second circuit is that in which a photo-electric cell works into the amplifier.[19, 20, 21] Vacuum cells generate shot noise of very nearly the theoretical value given by equation (7), while gas filled cells give even greater noise.

The total noise generated in the output of the vacuum photo-electric cell is the sum of the shot noise and thermal noise across the coupling resistance R, as given by equations (8) and (2). The mean square of the signal voltage, however, is $(\Delta i R)^2/2$ where Δi is the amplitude of the current variation. The ratio of signal to noise is then

$$V_{sg}^2/(V_t^2 + V_s^2) = 1.59 \times 10^{18} \frac{(\Delta i)^2}{F} \frac{R}{iR + 0.0516}. \tag{9}$$

This equation shows the expected fact that for a given value of Δi it is better to keep the direct photo-electric current small (high modulation). It also brings out the curious result that when the direct voltage drop in the coupling resistance is much more than 1/20 volt the noise is largely shot noise and the signal-to-noise ratio is independent of the coupling resistance, while if this voltage is much less than 1/20 volt the thermal noise predominates and the signal-to-noise ratio is proportional to the coupling resistance.

Shot Effect and Flicker Effect With Space Charge

When, as in an amplifier tube, the current in the tube is limited partly or wholly by space charge rather than by the cathode temperature, then the conditions are changed [13, 14, 16] in two respects. First, while the electrons still are emitted from the cathode at random times, they must arrive at the plate in a more orderly manner. Simple statistical laws no longer apply, the flow of current is smoother and the fluctuations are greatly reduced. Second, the impedance of the tube is no longer infinite, but has a finite value. The equation for the shot effect (equation (7)) now must be modified,[9, 16, 17] by substituting for the current i the quantity $j(\partial i/\partial j)^2$, where j is the total current emitted by the filament and $\partial i/\partial j$ is the rate of change of space current with emission current for the particular conditions used in the observation of the fluctuating voltage. Furthermore, in place of the coupling impedance Z the effective impedance Z_e of this in parallel with the tube resistance r_p must be used. The equation now reads

$$V^2 = 2\epsilon j \left(\frac{\partial i}{\partial j}\right)^2 \int_0^\infty Z_e^2 G^2 df. \tag{10}$$

In the absence of space charge j and i are identical, $\partial i/\partial j$ is unity, and Z_e becomes Z, so that the equation then represents the pure shot effect. With increasing space charge the value of $\partial i/\partial j$ approaches zero and Z_e becomes smaller so that the shot effect becomes very small. Similarly the flicker effect, being connected with the process of emission and not with the subsequent history of the electrons, also is made ineffective by the space charge. In fact, in well designed tubes the fluctuation noise of both shot effect and flicker effect in the space current appears to be reduced to such an extent as to be negligible.

Ions in the Space Charge

The effect of ions in the grid current already has been discussed. Ions also may cause fluctuations in the plate current of the tube.

The space charge which limits the current between cathode and anode consists of electrons in rapid progress toward the anode. A massive ion placed in this region travels much more slowly and contributes to the space charge for a much longer time than does an electron. While its own charge contributes little to the current, one ion may cause the current to change by the amount of hundreds of electrons during its flight through the space charge region, and the action of many ions would be additive.

Probably most of the ions existing in a tube are positive. Some of them are molecules of residual gas that have lost an electron by collision with an electron of the space current. Residual gas has been found to increase the noise of tubes, especially at the higher pressures. Observations at very low pressures are not conclusive, and it is not certain whether in any modern tubes the noise level is determined by the presence of residual gas.[9, 14, 19, 20, 21, 22, 23, 24]

Positive ions may be emitted also by the cathode. These never can attain a high velocity because they remain in a region of low field intensity. They may be trapped for a time in the region of the potential minimum near the cathode before they finally pass to the grid or possibly become neutralized by an electron. In modern tubes with low temperature filaments the effect of these ions is reduced greatly, yet still may account for a large part of the difference between the observed tube noise and the theoretical thermal noise of tubes.[14, 25, 26, 27, 28]

Noise in Commercial Tubes

Noise generated in an amplifier should consist largely of the thermal noise of the input circuit, to which is added the noise produced in the plate circuit of the first tube. It is convenient to consider that the tube noise comes, not from the plate circuit of the tube, but from a fictitious resistance R_G in series with the resistance R_e of the input circuit.[2] The effective thermal noise of the input circuit then is given by the expression

$$V^2 = 4kTF(R_e + R_G) = 1.64 \times 10^{-20}F(R_e + R_G). \qquad (11)$$

The tubes therefore may be rated conveniently in terms of R_G. The transformation to volts or to watts can be accomplished readily by equations (2) or (3). If, with a given tube and circuit, R_G approaches or exceeds R_e in value, the tube is responsible for an appreciable part of the total noise. The choice of another tube in which the ratio of R_e to R_G is more favorable then may be considered.

For the calculation of tube noise several formulas have been proposed, either entirely empirical [2] or with some basis on theory.[9, 24, 29] These formulas generally fail in the prediction of noise in tubes for the reason that the greater part of the noise in practical tubes is caused by things that have not been included in theory and that are still in a state of flux so far as manufacturing is concerned. It is best, therefore, to rely only on actual measurements of the noise in specific types of tubes. With modern tubes, the noise level of a given type of tube can be represented reasonably well by measurements made on a small number of samples.

Published data on noise of tubes are rather meager. The best series of measurements is that of Pearson,[10] which covers four Western Electric tubes at different frequency bands. These tubes are known commercially as types 102*G*, 262*A*, 264*B*, and 259*B*. The General Electric tube type *PJ*-11, designed specially for work at low frequencies, was studied by Metcalf and Dickinson.[24] They also give data, for the low frequency region, on the tubes known commercially as types 222, 240, 201, and 112. Johnson and Neitzert [30] have given data for the *PJ*-11 and the type '38 tube. Certain British tubes were studied by Moullin and Ellis,[29] and of these the type *AC*/2*HL* tube was found to have the lowest noise level. Brintzinger and Viehmann [31] studied a few German tubes. Of these the type *RE*-084 appears to have the lowest noise rating, but the data cannot be reduced to absolute measure.

In many of these studies the tubes were operated at voltages different from those usually employed. For these, the original papers should be consulted. In general for the best triodes R_G has a value of a few thousand ohms, while for screen tubes it has a value of a few tens of thousands. At the lowest voice frequencies the values may be somewhat greater.

Other Sources of Noise

While the more fundamental sources of noise have been discussed, it may be well to add some remarks on a few types of disturbance that often can be eliminated.

Noise From A-C Cathode Heating [32–39]

The indirectly heated cathode may be operated on alternating current when the tube is employed in radio frequency circuits. In audio amplifiers with gains in excess of 50 db, additional precautions must be taken to reduce the effects of the electric and magnetic fields of the heater and of coupling impedance between the heater and the other electrodes. Even under the best conditions, however, the hum

level is of the order of 20 db above the tube noise measured with d-c heating.

Noise From Vibration [32, 33, 39, 40]

Mechanical vibration changes the relative positions of the tube elements and hence causes disturbing noise. This is especially objectionable at audio frequency, although a radio frequency carrier may become modulated sufficiently to produce noise.

The remedy, used in the so-called "low microphonic" tube designs, is to stiffen the construction of the tube elements and to apply damping to their vibration, as well as to cushion the tube by a suitable mounting and to shield it from sound waves. The indirectly heated cathode is superior to the filamentary cathode in regard to noise from vibration.

Noise From Poor Insulation [32, 33, 39]

Noise arises from resistance changes at contacts and across thin films of conducting material deposited on insulating supports in the vacuum tube. Leaky capacitors may produce a similar noise.

Noise From Faulty Resistances

Many resistors in which the resistance element is a thin film are sources of noise. If no current flows in them only thermal noise is generated, but when direct current passes through them more noise is produced. The noise voltage is roughly proportional to the direct current. These resistors must be chosen carefully for circuit branches where current flows.

Signal-to-Noise Ratio [7, 29, 41, 42]

In so far as noise is concerned, the merit of a transmission system is dependent not only on the amount of noise present, but also on the strength of the signal, so that a determination of the ratio of the signal level to the noise level is necessary. Fortunately, this ratio has a reference value for any given transmission system determined uniquely by the ratio of the signal to thermal noise in the input circuit. The ratio of the signal level to noise level is here the greatest the ratio ever can attain, because noise that originates at subsequent points in the amplifier contributes to the noise level without increasing the signal.

This fact provides a basis for the rating of amplifiers, the thermal noise of the input circuit being used as a comparison signal. For example, the noise output of an amplifying system may be 0.3 mw which falls to 0.2 mw when the input circuit is short-circuited. The thermal noise from the input circuit is then the difference between

these two readings, namely, 0.1 mw, and the signal-to-noise ratio of the actual system is 3 times, or 4.8 db worse than its best possible value with a given signal. These data may be expressed in terms of an equivalent input resistance which has the advantage that the amplifying properties of the tube have been taken into account. This leaves for the engineer only the problem of selecting a tube having an input capacity of such a value that the construction of a relatively high impedance input circuit is possible.

So far, the discussion has been based upon the properties of amplifiers only, no mention being made of the effects of modulators, detectors, frequency converters, and other nonlinear devices on the signal-to-noise ratio. A detailed discussion of the noise in such devices is beyond the scope of this paper, but the relations in the most commonly used ones may be indicated and their general properties outlined.

First, consider a system composed of a radio frequency amplifier followed by a detector and a pair of headphones. A certain amount of noise will be heard in the phones if the gain of the amplifier is great enough. This noise is caused by the various components of the radio frequency noise beating together in the detector to form audio frequency components. Next, suppose that an unmodulated carrier is introduced into the amplifier. It will be observed that the audio noise in the phones increases. The increase in audio noise is produced by the radio frequency carrier beating with the radio frequency noise components and this increase is proportional to the strength of the carrier.

If a small percentage of modulation is added to the carrier, the audio signal-to-noise ratio in the phones will be determined by the properties of the amplifier in the same way as though the system were a straight amplifier without any detector. Comparison of the actual system with the ideal may be made by introducing the carrier into one of the amplifier stages subsequent to the input, and then measuring the audio noise with the input circuit in its normal condition and again with the input circuit short-circuited. The ratio of these two energy values gives the ratio of the equivalent input resistance of the actual system to the equivalent input resistance of the noisy amplifier alone. The ratio of the ideal signal-to-noise ratio to the actual one may be found by dividing the difference between the two audio energy readings by the reading taken with the input circuit in its normal condition.

If the percentage of modulation of the carrier is large, the system will be noisier because there will be appreciable audio noise components caused by beats between the side bands and the radio frequency noise components. Again, if the carrier level is not large compared with

the noise level in the amplifier, the system will be noisier because the beats between the noise components are appreciable compared with the beats between carrier and the noise.

The same considerations apply to the first detector in a double-detection receiving system. If, as is usual, the beating oscillator voltage is large compared with the noise components, then the frequency band of the noise will be shifted in position in the same manner as the signal, and the signal-to-noise ratio of the system will be unchanged by the frequency conversion.

The signal-to-noise properties of any system are considered satisfactory when the total output noise differs only slightly from that produced by thermal agitation in the input circuit alone, and this difference may be measured by eliminating the input thermal noise (as by the short-circuit method) and noting the change produced in the output noise.

BIBLIOGRAPHY

1. W. Schottky, "Ueber spontane Stromswankungen in vershiedenen Elektrizitätsleitern" ("On Spontaneous Current Fluctuations in Various Electric Conductors"), *Ann. d. Phys.*, v. 57, 1918, p. 541–67.
2. J. B. Johnson, "Thermal Agitation of Electricity in Conductors," *Phys. Rev.*, v. 32, 1928, p. 97–109.
3. H. Nyquist, "Thermal Agitation of Electric Charge in Conductors," *Phys. Rev.*, v. 32, 1928, p. 110-13.
4. E. K. Sandemann and L. H. Bedford, "The E.M.F. of Thermal Agitation," *Philosophical Mag.*, v. 7, 1929, p. 774–82.
5. N. H. Williams and D. A. Wilbur, "Thermal Agitation of Electrons in a Metallic Conductor," *Science*, v. 76, 1932, p. 519–20.
6. E. B. Moullin and H. D. N. Ellis, "A Measurement of Boltzmann's Constant by Means of the Fluctuations of Electron Pressure in a Conductor," *Cambridge Philosophical Soc. Proc.*, v. 28, 1932, p. 386–402.
7. F. B. Llewellyn, "A Rapid Method of Estimating the Signal-to-Noise Ratio of a High Gain Receiver," *I. R. E. Proc.*, v. 19, 1931, p. 416–20.
8. H. S. Black, "Stabilized Feed-Back Amplifiers," *Elec. Engg.*, v. 53, 1934, p. 114–20.
9. F. B. Llewellyn, "A Study of Noise in Vacuum Tubes and Attached Circuits," *I. R. E. Proc.*, v. 18, 1930, p. 243–65.
10. G. L. Pearson, "Fluctuation Noise in Vacuum Tubes," *Physics*, v. 5, 1934, p. 233–43 and *Bell Syst. Tech. Jour.*, v. 13, 1934, p. 634–53.
11. W. Schottky, "Zur Berechnung und Beurteilung des Schroteffektes" ("On the Calculation and Critical Examination of the Shot Effect"), *Ann. d. Phys.*, v. 68, 1922, p. 157–76.
12. T. C. Fry, "The Theory of the Schroteffekt," *Franklin Inst. Jour.*, v. 199, 1925, p. 203–20.
13. A. W. Hull and N. H. Williams, "Determination of Elementary Charge E From Measurements of Shot Effect," *Phys. Rev.*, v. 25, 1925, p. 147–53.
14. J. B. Johnson, "The Schottky Effect in Low Frequency Circuits," *Phys. Rev.*, v. 26, 1925, p. 71–85.
15. W. Shottky, "Small Shot Effect and Flicker Effect," *Phys. Rev.*, v. 28, 1926, p. 75–103.
16. E. W. Thatcher and N. H. Williams, "Shot Effect in Space Charge Limited Currents," *Phys. Rev.*, v. 40, 1932, p. 474–96.
17. E. W. Thatcher, "On the Reduction of Shot Effect Fluctuations," *Phys. Rev.*, v. 40, 1932, p. 114–15.
18. R. L. Hafstad, "The Application of the FP-54 Pliotron to Atomic Disintegration Studies," *Phys. Rev.*, v. 44, 1933, p. 201–13.

19. B. A. Kingsbury, "The Shot Effect in Photoelectric Currents," *Phys. Rev.*, v. 38, 1931, p. 1458–76.
20. F. V. Orbán, "Schroteffekt und Wärmegräusch im Photozellenverstärker" ("Shot Effect and Thermal Noise in Photoelectric Cell Amplifiers"), *Zeits. f. Techn. Phys.*, v. 13, 1932, p. 420–4 and v. 14, 1933, p. 137–43.
21. E. Steinke, "Natürliche Schwankung schwächster Photoströme" ("Natural Fluctuations of Very Small Photoelectric Currents"), *Zeits. f. Phys.*, v. 38, 1926, p. 378–403.
22. S. Ballantine, "Fluctuation Noise Due to Collision Ionization in Electronic Amplifier Tubes," *Phys.*, v. 4, 1933, p. 294–306.
23. N. P. Case, "Receiver Design for Minimum Fluctuation Noise," *I. R. E. Proc.*, v. 19, 1931, p. 963–70.
24. G. F. Metcalf and T. M. Dickinson, "A New Low Noise Vacuum Tube," *Phys.*, v. 3, 1932, p. 11–17.
25. J. S. Donal, "Abnormal Shot Effect of Ions of Tungstous and Tungstic Oxide," *Phys. Rev.*, v. 36, 1930, p. 1172–89.
26. H. N. Kozanowski and N. H. Williams, "Shot Effect of the Emission From Oxide Cathodes," *Phys. Rev.*, v. 36, 1930, p. 1314–29.
27. L. P. Smith, "Effect of Positive Ion Shot Effect on Space Charge Limited Currents," *Phys. Rev.*, v. 35, 1930, p. 1430.
28. E. W. Thatcher and N. H. Williams, "Shot Effect in Space Charge Limited Currents," *Phys. Rev.*, v. 39, 1932, p. 474–96.
29. E. B. Moullin and H. D. M. Ellis, "The Spontaneous Background Noise in Amplifiers Due to Thermal Agitation and Shot Effects," *Jour. I. E. E.*, v. 74, 1934, p. 323–356.
30. E. A. Johnson and C. Neitzert, "The Measurement of Small Alternating Voltages at Audio-frequencies," *Rev. Sci. Inst.*, v. 5, 1934, p. 196–200.
31. W. Brentzinger and H. Viehmann, "Das Rauschen von Empfängern" ("Noise in Radio Receivers"), *Hochfr. und Elektroauk.*, v. 39, 1932, p. 199–207.
32. M. J. Kelly, "Vacuum Tube and Photoelectric Tube Developments for Sound Picture Systems," *Jour. Soc. Motion Picture Engrs.*, v. 18, 1932, p. 761–81.
33. J. O. McNally, "Analysis and Reduction of Output Disturbances Resulting From the Alternating-Current Operation of the Heaters of Indirectly Heated Cathode Triodes," *I. R. E. Proc.*, v. 20, Aug. 1932, p. 1263–83.
34. H. M. Freeman, "A Practical Alternating-Current Radio Receiving Tube," *Elec. Jour.*, v. 19, Dec. 1922, p. 501–5.
35. F. G. McCullough, "Thermionic Tubes," *I. R. E. Proc.*, v. 10, Dec. 1922, p. 468–85.
36. A. W. Hull, "A Combined Kenotron Rectifier and Pliotron Receiver Capable of Operation by Alternating Current Power," *I. R. E. Proc.*, v. 11, April 1923, p. 89–96.
37. B. F. Meissner, "AC as a Filament Supply Source," *Radio Broadcast*, v. 10, Feb. 1927, p. 393–6, March 1927, p. 495–7.
38. W. J. Kimmell, "The Cause and Prevention of Hum in Receiving Tubes Employing AC Direct on the Filament," *I. R. E. Proc.*, v. 16, Aug. 1928, p. 1089–1106.
39. D. B. Penick, "The Measurement and Reduction of Microphonic Noise in Vacuum Tubes," *Bell Sys. Tech. Jour.*, v. 13, Oct. 1934, p. 614–33.
40. A. C. Rockwood and W. R. Ferris, "Microphonic Improvement in Vacuum Tubes," *I. R. E. Proc.*, v. 17, Sept. 1929, p 1621–32.
41. S. Ballantine, "Fluctuation Noise in Radio Receivers," *I. R. E. Proc.*, v. 18, Aug. 1930, p. 1377–87.
42. E. N. Dingley, "A Common Source of Error in Measurements of Receiver Selectivity," *I. R. E. Proc.*, v. 22, May 1934, p. 546.

Vacuum Tubes as High-Frequency Oscillators *

By M. J. KELLY and A. L. SAMUEL

Vacuum tubes as oscillators and amplifiers at frequencies greater than 100 megacycles (3 m.) are considered in this paper. The type of construction used in a large number of different tubes, and the characteristics of the tubes, are presented. Circuits for operating the tubes also are considered and the theory of operation and the factors limiting ultra-high frequencies are discussed. Principal attention is given to the tubes as oscillators, with brief consideration of the problem of amplification.

THE three types of oscillation generators which at present are the most efficient in the range from 100 megacycles to 3000 megacycles per second will be discussed in the following survey. These are: the negative grid tube which at lower frequencies is the conventional regenerative oscillator, the positive grid or Barkhausen oscillator, and the "magnetron" oscillator. The amplification problem will be briefly discussed. Because of the present unsettled state of the theory, only the most elementary and generally accepted part will be included. Much theoretical and experimental work remains to be done before knowledge of the mechanism of oscillation and amplification in this frequency range will be satisfactory. As is often the case, the empirical knowledge of some of these mechanisms has outdistanced the theoretical interpretation.

The Negative Grid Oscillator

The conventional thermionic vacuum triode, whether it be a large water-cooled power tube or a small receiving tube, may be used as a generator of oscillations varying in frequency from a few cycles per second to some 20 or 30 megacycles with substantially undiminished efficiency and output. In this range the frequency at which a tube is to be employed is a factor of almost negligible importance in the determination of its characteristics and its form. Beyond this range, however, frequency plays an increasingly important part, and as one approaches 300 megacycles, it becomes the most important factor in the determination of tube design.

When an attempt is made to operate a standard triode at increasingly high frequencies it is found that the output and efficiency begin to decrease. The frequency at which this is first observed will depend

* Published in November 1934 issue of *Electrical Engineering*. Presented at Winter Convention of *A. I. E. E.*, New York City, January 22–25, 1935.

upon the design of the tube but it will usually be in the 10- to 60-megacycle range. By successive modifications of the circuit arrangement and size this decrease in power output and efficiency can be minimized. With optimum circuit arrangements, however, this decrease continues until finally a frequency is reached beyond which oscillations can no longer be produced.

In Fig. 1 are shown typical data for a standard 75-watt tube, the Western Electric type 242*A*, operated with reduced potentials over

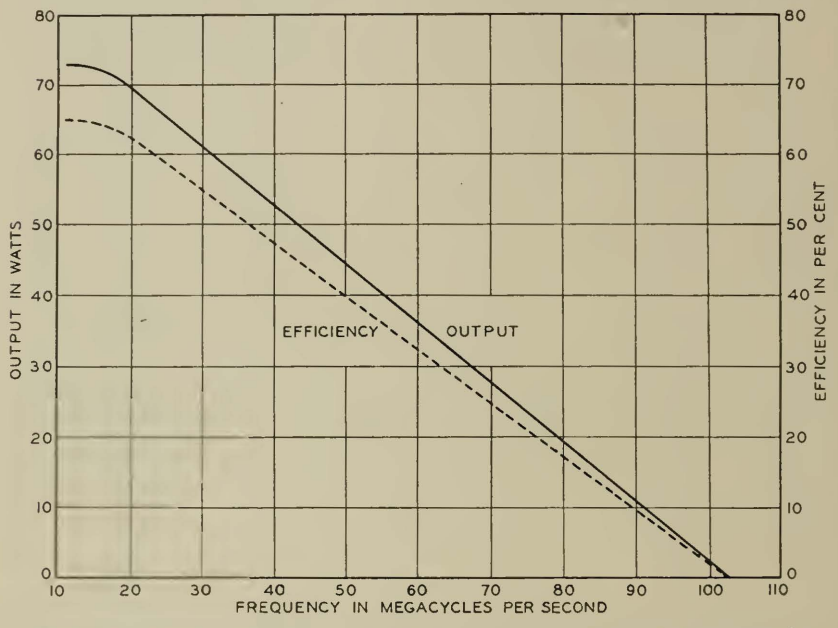


Fig. 1—Power output and anode efficiency as a function of frequency for a standard triode. These curves are typical of all tubes as they approach their upper limiting frequency.

the frequency range from the point where oscillation frequency noticeably affects performance to the point where oscillations can no longer be produced. The plate potential was held constant at 750 volts throughout the entire frequency range. The oscillation circuit was modified at each point in order to obtain maximum output and efficiency, keeping the anode dissipation within the maximum rating of 100 watts and the anode current within the maximum rating of 0.150 amp. It can be seen from the curves that the output and efficiency are independent of the frequency until about 20 megacycles is reached, when they begin to decrease. At 100 megacycles the output power is only 2.5 watts and the efficiency only 2 per cent. The tube will not oscillate at 105 megacycles.

Effect of Energy Losses on Performance

A tube operating in the range where frequency affects performance must withstand energy losses, and the resulting heating within its structure, which occur to only a negligible degree at the lower frequencies. Some of these losses are due to dielectric hysteresis in the insulating materials of the tube, particularly in the portions of the glass supporting stem or bulb which lies between the tube leads. The glass is sometimes so softened by the heat thus developed that it is punctured by the outside air pressure. Losses also occur in the auxiliary metallic parts of the tube structure due to the increased eddy currents that occur at high frequencies. Losses in the tube electrodes and their lead-in wires are also greatly increased due to skin effect which increases their resistance, and due to the increased charging current required by the inter-electrode capacitances. The increased lead temperature, depending upon its amount, will cause a more or less rapid deterioration of the lead-to-glass seals which may ultimately destroy the vacuum. In order to protect the tube from damage because of these new types of energy dissipation, the operating potentials and currents must be reduced to values less than those established for low-frequency operation. Some manufacturers are now giving special ratings on such of their standard tubes as may be used at ultra-high frequencies. These ratings should be adhered to when operating in this range.

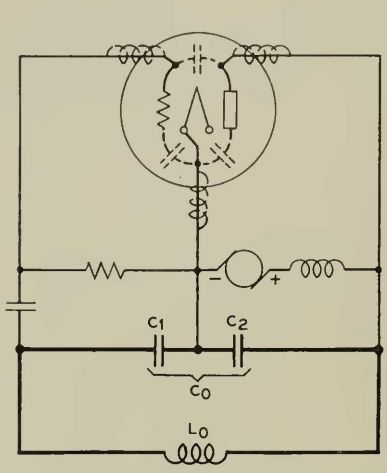
Effect of Circuit on Performance

The decrease in output power and plate efficiency which sets in with the increase in frequency, while due in part to the losses described above, and to the rapid increase in radiation losses, is also due to two additional effects of fundamental importance. The first to become evident, with increasing frequencies, is circuital in nature. This can be explained by reference to the conventional oscillator shown in Fig. 2. The frequency of such an oscillator is given by

$$f = \frac{1}{2\pi\sqrt{LC}},$$

where L and C are the effective inductance and capacitance of the oscillating circuit. In the lower frequency range, the LC product which determines the frequency is substantially equal to L_0C_0, that is, to the product of the inductance and capacitance of the external circuit. The inductance of the tube leads and the capacitance between its electrodes (indicated by the dotted lines in the figure) play a negligible rôle. In order to tune the circuit to higher and higher frequencies,

the capacitance C_0 is first reduced and finally eliminated, leaving the inter-electrode capacitance as the only capacitance in the oscillating circuit. The tube leads then form a part of the main oscillating circuit, in which large circulating currents must exist for stable operation. For a further increase in frequency the external inductance L_0



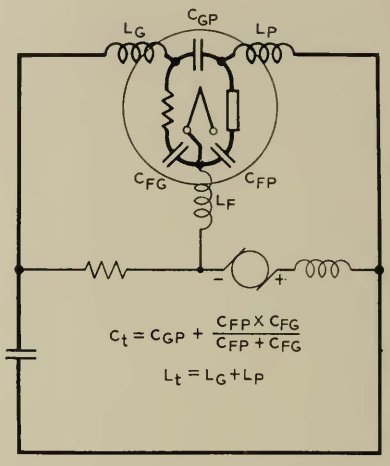


Fig. 2 (left)—A standard Colpitts oscillator circuit. The heavy lines indicate the main oscillating circuit. The dotted portions represent the inter-electrode capacitances and lead inductances which play a minor rôle at low frequencies.

Fig. 3 (right)—The limiting circuit with the external tuning capacitance eliminated and the external inductance reduced to a short circuiting bar between the grid and plate leads. The main oscillating circuit, indicated by the heavy lines, is seen to include the inter-electrode capacitances and lead inductances.

must be reduced and, in the limit, it becomes the shortest possible connection between the grid and plate terminals. The oscillating frequency for this limiting circuit, shown in Fig. 3, is determined by the product of the lead inductance L_t and the inter-electrode capacitance C_t; that is,

$$f_0 = \frac{1}{2\pi\sqrt{L_t C_t}},$$

where L_t is the sum of the grid and plate lead inductance and C_t is the total grid-plate capacitance. Even before this frequency limit is reached the output power and plate efficiency are seriously reduced by the lack of full control over the relative amplitude and phase of the alternating grid and plate potentials. Whereas the ratio of these amplitudes is controlled in the circuit shown in Fig. 2 by the condensers C_1 and C_2, it is determined in the limiting case primarily by

the fixed ratio of the grid-filament to plate-filament inter-electrode capacitance. Most tubes made especially for ultra-high-frequency use are constructed so as to minimize these circuit limitations by a reduction in the inter-electrode capacitance and lead inductance and by adjusting the capacitance ratio.

Effect of Transit Time on Performance

The second fundamental effect has to do with the time required for the electrons to travel from the cathode to the anode within the tube structure. This time, the so-called transit time, is very small in present-day commercial types of power tubes, usually much less than one micro-second. Obviously at low frequencies it can be neglected and, in fact, for many tubes it still plays a minor rôle either in determining the output and efficiency in the high-frequency range or in establishing the limiting frequency for oscillations. When the frequency range of oscillation of a tube is extended by an adequate decrease in energy losses and by improvements in electrical design, transit time becomes a dominating factor in the reduction of output power and efficiency and in establishing the limiting frequency of oscillation.

This comes about in two ways. In the first place, the relative phase of the alternating grid and plate potentials for best operation must be altered to compensate for the time required for the electrons to travel from the region in which the grid has its greatest effect upon their motion to the region in which their motion has the greatest effect upon the plate current. The available control over these phases is usually insufficient to permit a realization of the optimum adjustment. In terms of the measured characteristics of the tube, the transconductance has become complex. But even with the optimum phase adjustment the efficiency is reduced by losses which occur because of the variations in grid and plate potentials during the transit time. Electrons arriving at the plate will in general have velocities greater than the velocity corresponding to the potential of the anode at the instant of their arrival. The excess energy corresponding to the greater velocity is obtained from the oscillating circuit and is dissipated at the plate in the form of heat. Again in terms of the measured characteristics, the input conductance has been increased above its low-frequency value.

The mechanism which enables electrons to take energy from the oscillatory circuit in their passage across the tube is evident from a consideration of a somewhat simplified case as shown in Fig. 4. Assume that the anode is held at a constant positive potential of 100

volts, and that the grid is held at 50 volts positive just long enough to allow an electron to come from the cathode to the grid plane (very near one of the wires), where its velocity will correspond to a fall of 50 volts. The potential of the grid is then suddenly changed to 50 volts negative. The electron will then fall through an additional potential difference of 150 volts, arriving at the anode with a velocity corre-

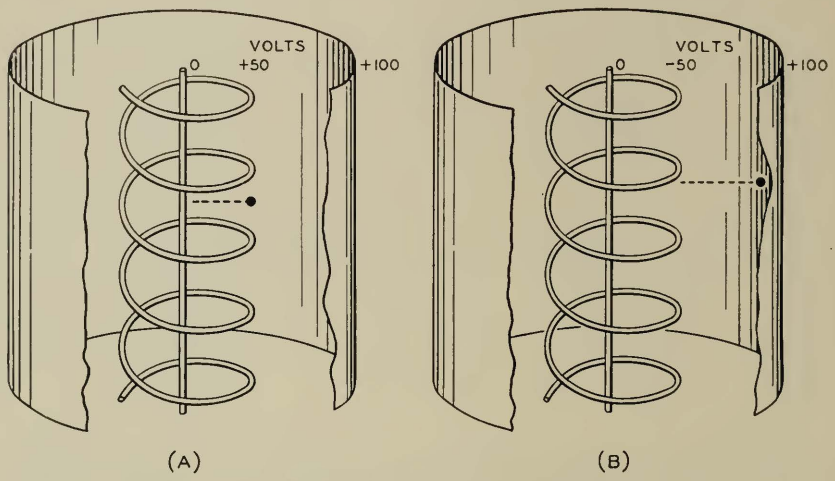


Fig. 4—Illustrating the mechanism which enables electrons to take energy from the oscillating circuit.

sponding to 200 volts, producing just twice as much heat as it would have done had the grid potential not been changed during the transit time. This added energy must come from the source which produced the change in the grid potential. In the actual case the change in grid potential is not abrupt but a similar loss occurs. This limits the useful frequency range of a tube to values for which the oscillation period is long compared to the electron transit time.

Special Designs Required for Different Ranges of High Frequencies

Most standard power tubes reach their upper frequency limit of oscillation somewhere in the 10- to 100-megacycle frequency range. For frequencies above this, specially designed tubes are required. The frequency range in which a given design is near the optimum is limited. Therefore, there is a succession of tubes, each rated for a band of frequencies. Characteristics such as a high mutual conductance and a sharp cut-off which make a tube a good oscillator at low frequencies, while still of importance at ultra-high frequencies, are apt to be secondary to the special frequency requirements

Although some progress has been made in the modification of conventially designed water-cooled tubes for use above 100 megacycles, more attention has so far been given to the development of radiation-cooled tubes for this frequency range.

A departure from conventional design with increasing frequency is illustrated by a radiation-cooled tube described by McArthur and Spitzer [1] in which the ratio of the plate diameter to the plate length

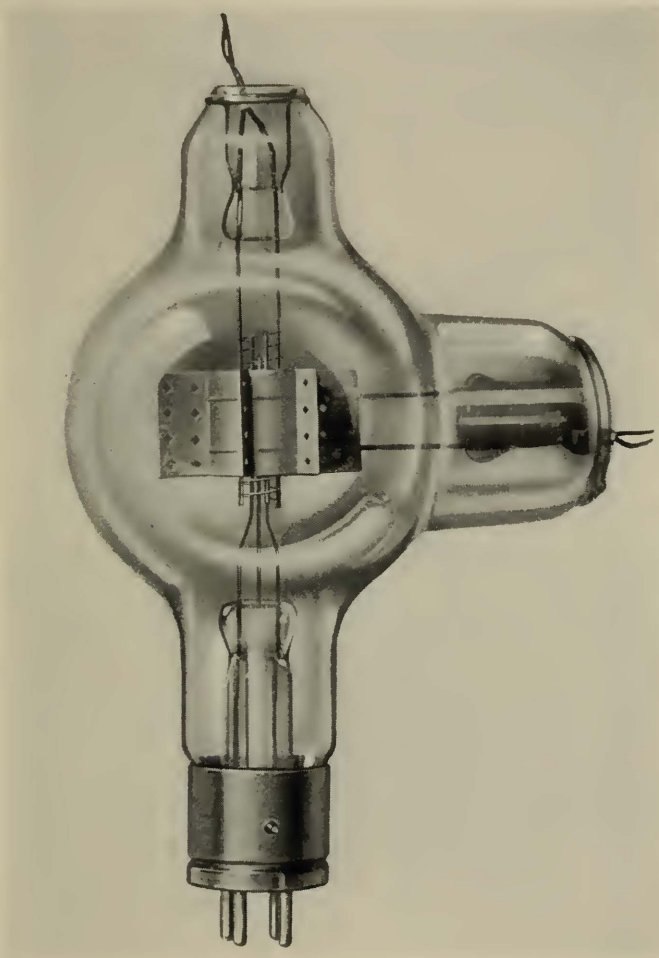


Fig. 5—A radiation cooled tube for use in the frequency range from 60 to 180 megacycles per second. Note the large ratio of the plate diameter to plate length and the special arrangement of the leads.

[1] For all numbered references see list at end of paper.

is much larger than for the conventional tube. Radiating fins are employed to compensate for the decrease in heat radiating ability of the plate which would otherwise occur because of its short length. In Fig. 5 is shown a photograph of this tube. It will be noted that the tube electrodes are supported directly from their leads. The complete absence of auxiliary supporting members either of metal or of insulating material and the large size of leads reduce radiation, eddy current, and conduction current losses. That portion of the inter-electrode capacitances due to the supporting structure is also made small by this method of support.

The inter-electrode capacitances are given below, together with the corresponding values for the type 242*A* tube, which has the same plate dissipation rating but is designed for use at lower frequencies:

	High Frequency Tube	242*A*
Plate to grid	3 $\mu\mu$f	13.0 $\mu\mu$f
Grid to filament	2 $\mu\mu$f	6.5 $\mu\mu$f
Plate to filament	1 $\mu\mu$f	4.0 $\mu\mu$f

The decrease in capacitance by a factor of approximately 4 makes possible a much greater improvement in performance in the 60- to 150-megacycle-per-second frequency range than the corresponding degradation in performance due to the lower mutual conductance and the increased electron transit time resulting from the increased spacing. The material increase in plate impedance makes it necessary to employ an anode potential approximately twice as great with the high-frequency tube.

Output and efficiency curves are shown in Fig. 6. (For the sake of uniformity, curves taken from published papers have, in most cases, been redrawn.) The particular shape of the output curve is due to the manner in which the applied anode potential was reduced with increasing frequency to minimize the danger of tube failure from the increased energy losses which occur at high frequencies. The limiting frequency as set by the inter-electrode capacitances and lead inductances is given by the authors as 230 megacycles. An extension of the efficiency curve to higher values indicates that the tube will probably fail to oscillate before this limit is reached. From this it can be inferred that the decrease in efficiency in the range from 150 to 200 megacycles is due largely to the effect of the relatively large transit time, since the authors' method of arriving at the output by taking the difference between the measured input and the measured plate losses includes circuit and lead losses as a part of the output. A comparison of the data of Fig. 1 and Fig. 6 shows that at 100 megacycles the output of the type 242*A* tube is 2 watts and the corre-

sponding output of the high-frequency tube is approximately 86 watts with substantially the same plate loss. This strikingly illustrates the improvement obtained by taking into account the factors which become important in the high-frequency range.

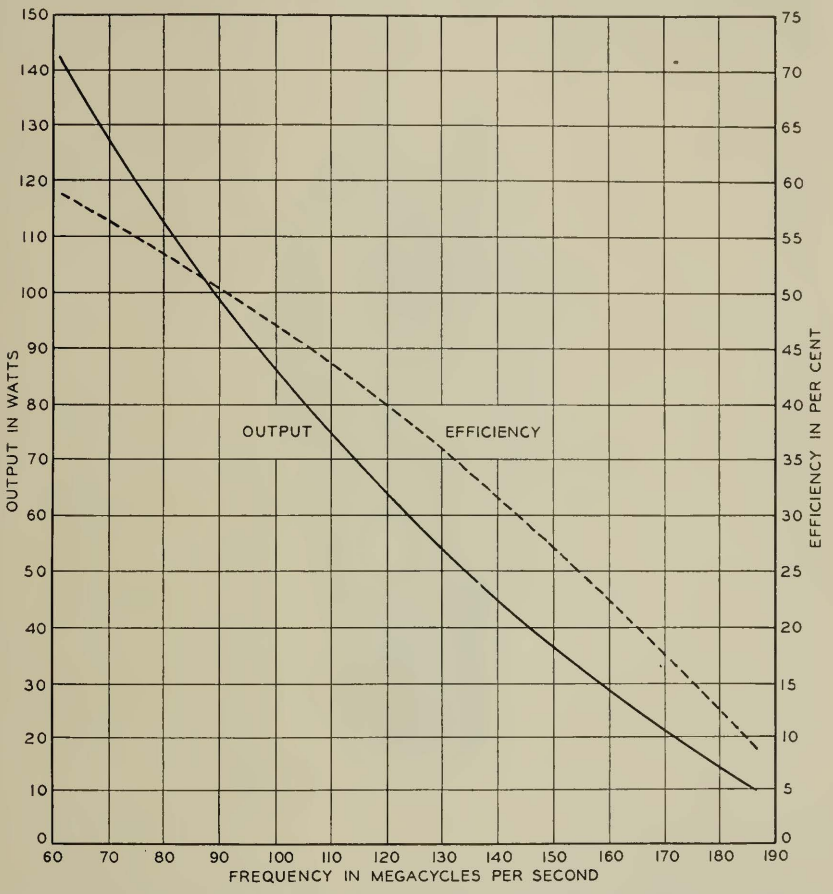


Fig. 6—Output and efficiency as a function of frequency for the tube shown in Fig. 5. A comparison of these curves with those shown in Fig. 1 illustrates the improvement obtained by taking account of those factors which become important at high frequencies.

Transit Time Becomes More Important

In extending the frequency range to 300 megacycles the importance of electron transit time becomes relatively greater. It must be kept as low as possible even at the expense of relatively higher inter-electrode capacitances. In the tube just described for the frequency range around 100 megacycles the reverse procedure is followed; in

order to make the inter-electrode capacitances as small as possible, transit times are increased. Fay and Samuel [2] in a recent paper presented before the International Scientific Radio Union (U.R.S.I.) describe a tube designed for use at 300 megacycles which well illustrates this point. The tube is shown in Fig. 7. It differs from the one

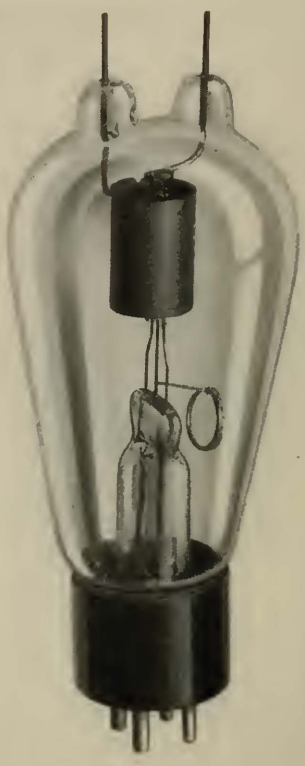


Fig. 7—A tube for use at frequencies up to 350 megacycles.

previously discussed in the close spacings between elements, particularly between the grid and filament. The lead length is further decreased and lead diameter made considerably larger in order to decrease lead inductance and resistance. The inter-electrode capacitances are:

Plate to grid	2.5 $\mu\mu$f
Grid to filament	2.0 $\mu\mu$f
Plate to filament	0.07 $\mu\mu$f

While these capacitances are substantially the same as for the tube shown in Fig. 5, the limiting frequency, as set by circuit resonance, is

somewhat beyond 400 megacycles as contrasted with 230 megacycles for the other tube. This is due, primarily, to the material decrease in lead inductance. The decreased losses resulting from the minimized transit time more than compensate for the increased circuit loss resulting from the required higher charging currents to the inter-electrode capacitances. Its mutual conductance of 2200 micro-

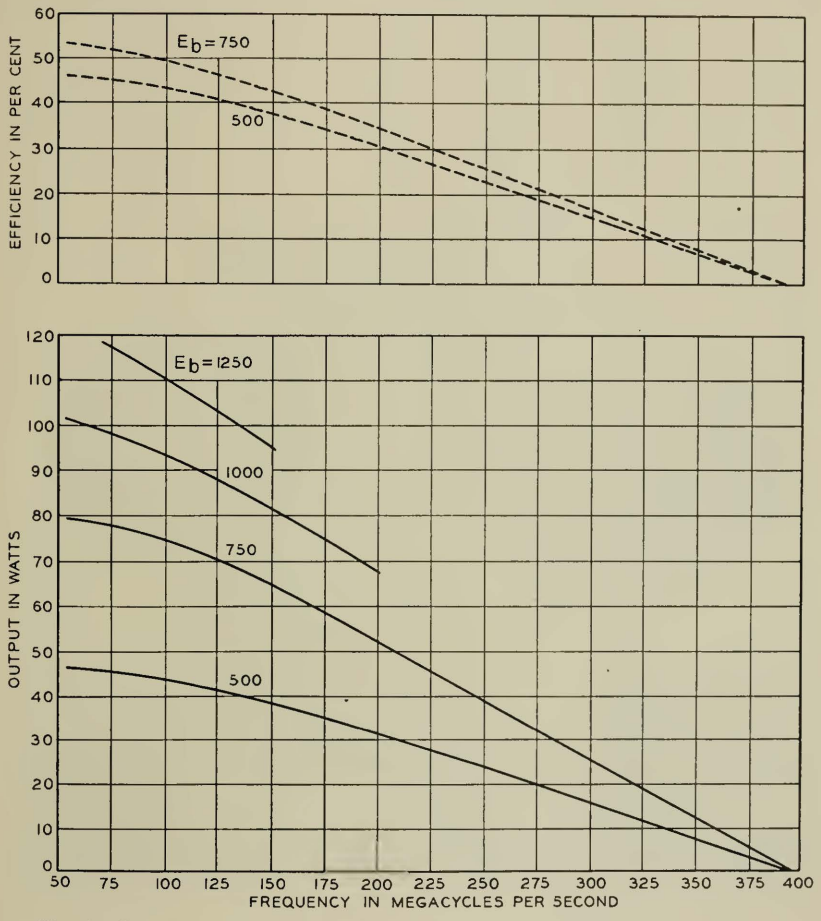


Fig. 8—Output and efficiency curves for 2 tubes of the type shown in Fig. 7.

ohms and the sharp cut-off shown by the static characteristics indicate that those electrical characteristics which are important for efficient oscillators in the low-frequency range have not been sacrificed in meeting the requirements of 300-megacycle operation.

Output and efficiency curves for this tube are shown in Fig. 8.

These data are for two tubes operated in the push-pull circuit shown in Fig. 9. The output shown represents only useful power, since it is the photometrically measured power consumed in a lamp load. It will be noted that, whereas the maximum output at 100 megacycles is 55

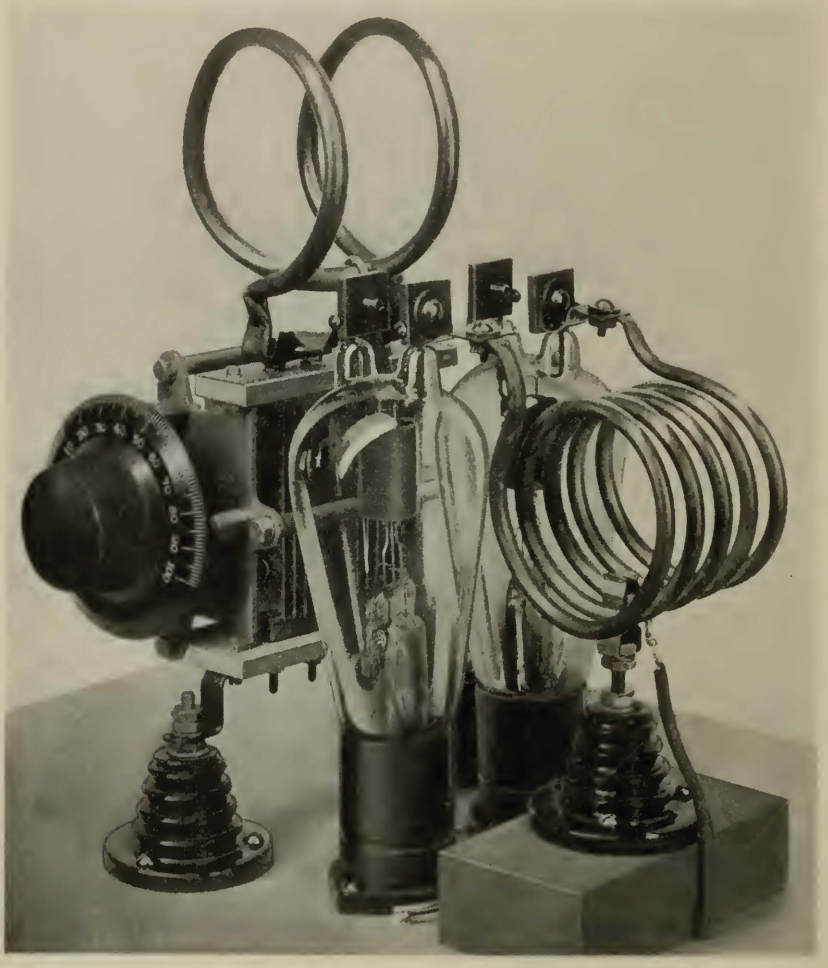


Fig. 9—A typical push-pull circuit for use at ultra-high frequencies. A circuit of this type was used to obtain the data shown in Figs. 8 and 11.

watts per tube and the efficiency 50 per cent, corresponding roughly to the 86 watts output at 47 per cent efficiency for the tube of Fig. 5, the output at 200 megacycles is 34 watts at 33 per cent efficiency as compared to less than 10 watts at an efficiency of only a few per cent

for the other tube. At 300 megacycles the higher frequency tube gives an output of 13 watts, while the tube of Fig. 5 no longer oscillates. This difference in behavior is due primarily to the decreases in transit time and in circuit losses, and to the more nearly optimum ratio existing between the inter-electrode capacitances. It is due to a considerably less extent to the increase in the frequency limit set by circuit resonance.

A Still Further Departure in Construction

The tube illustrated in Fig. 10 represents a still further departure from conventional construction with a corresponding increase in the

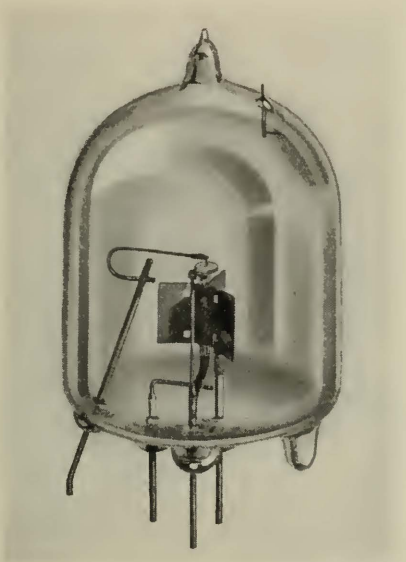


Fig. 10—This tube will oscillate at frequencies up to 740 megacycles. Note the absence of the usual press and the extremely small size of the elements.

frequency limit. Fay and Samuel report an output of 6 watts per tube at 500 megacycles and a frequency limit of 740 megacycles. Unusual features of the design are: The complete elimination of the usual press, the close spacing of the leads and the special construction of the tube elements, particularly the grid, made necessary by their small size. The grid is in the form of a number of straight wires (parallel and equidistant from the axial filament) supported by cooling collars at each end. The plate, in spite of its small size, can dissipate 40 watts with safety.

The inter-electrode capacitances of this tube are:

Plate to grid	1.8 μμf
Grid to filament	1.0 μμf
Plate to filament	0.75 μμf

The output and efficiency curves, shown in Fig. 11, were obtained with two tubes in a push-pull circuit. The efficiency of the tube is

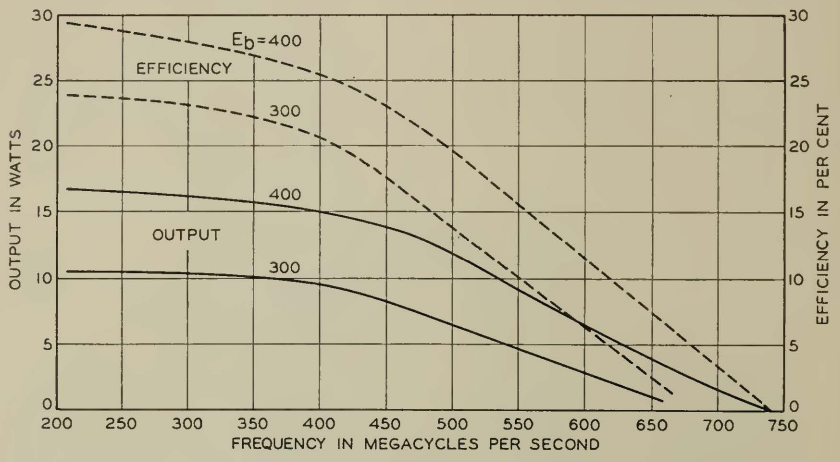


Fig. 11—Output and efficiency curves for 2 tubes of the type shown in Fig. 10.

about 28 per cent at 300 megacycles, contrasting with only 18 per cent for the previously discussed tube, while the output is only 8 watts as compared to 13 watts. The higher efficiency suggests that above 300 megacycles two of the smaller tubes are preferable to one of the large tubes. The fact that the limiting frequency varies with the applied anode potential indicates that the transit time effect is largely responsible for the decreased efficiency. This suggests modulation difficulties if the tube is used near the upper limit of the frequency range. The outputs and efficiencies in the 400- to 600-megacycle range, although low, represent a substantial increase over the usual values obtained at these frequencies.

The relatively low efficiency at 200 megacycles is due to insufficient filament emission. In order to maintain space charge conditions near the filament, at high frequencies, the electron emission must be large enough to supply not only the actual electron current to the plate but also the charging currents to the grid-filament and plate-filament capacitances. In a tube of this small size an increase in the filament emission is possible only by an impractical and unwarranted increase in the filament heating current.

The logical extension of these principles to increasingly high frequencies requires the use of closer and closer inter-electrode spacings. Severe mechanical difficulties are encountered. Curiously enough the limiting factor in the power dissipating ability of the tube turns out to be the grid temperature rather than the temperature of the plate as might be expected. This comes about because of the required close grid-filament spacing, and makes necessary the adoption of some method of cooling the grid. One of the writers has constructed a series of tubes in which the grid is a tungsten helix, each turn of which is attached to a common cooling fin projecting through a slot in the plate. This construction simplifies the mechanical problems involved and provides ample grid cooling. Two of these tubes are shown in Fig. 12.

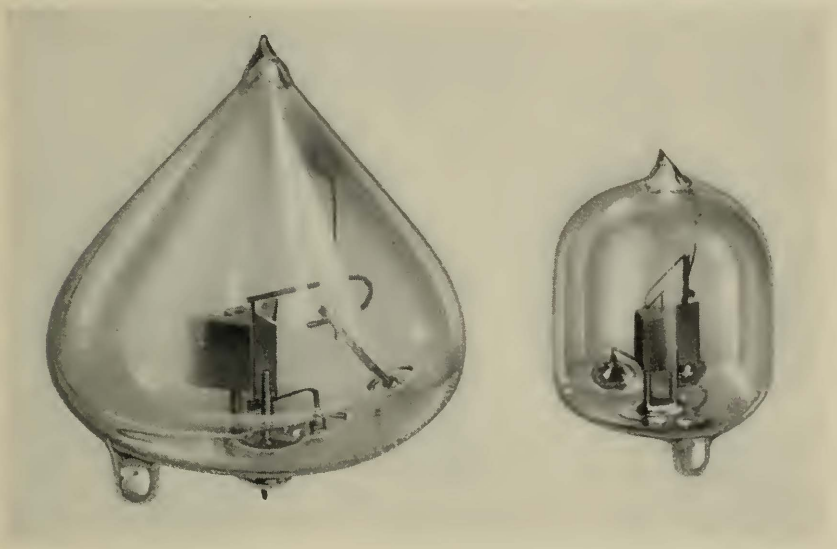


Fig. 12—These tubes represent a further extension in design according to the principles under discussion. The smaller one will oscillate at 1200 megacycles per second.

The larger one will deliver 10 watts at 670 megacycles with an efficiency of 20 per cent and the smaller one will deliver one watt at 1200 megacycles with an efficiency of 10 per cent. These tubes are in no sense commercial, the results representing the limit that has been obtaned by specially constructed tubes under controlled laboratory conditions at voltages and currents above those for which the tube would have a long life. With further advances it is reasonable to expect that outputs of this sort will be commercially realizable and that the frequency

range of the negative grid oscillator can be extended beyond 1200 megacycles.

Tubes for Receiving Purposes

For receiving purposes where large outputs are not needed, the ultra-high-frequency requirements may be met by shrinking all the tube dimensions in proportion to the desired wave-length. Tubes constructed by Thompson and Rose [3] based upon this principle are shown in Fig. 13 compared in size with the conventional receiving type tube.

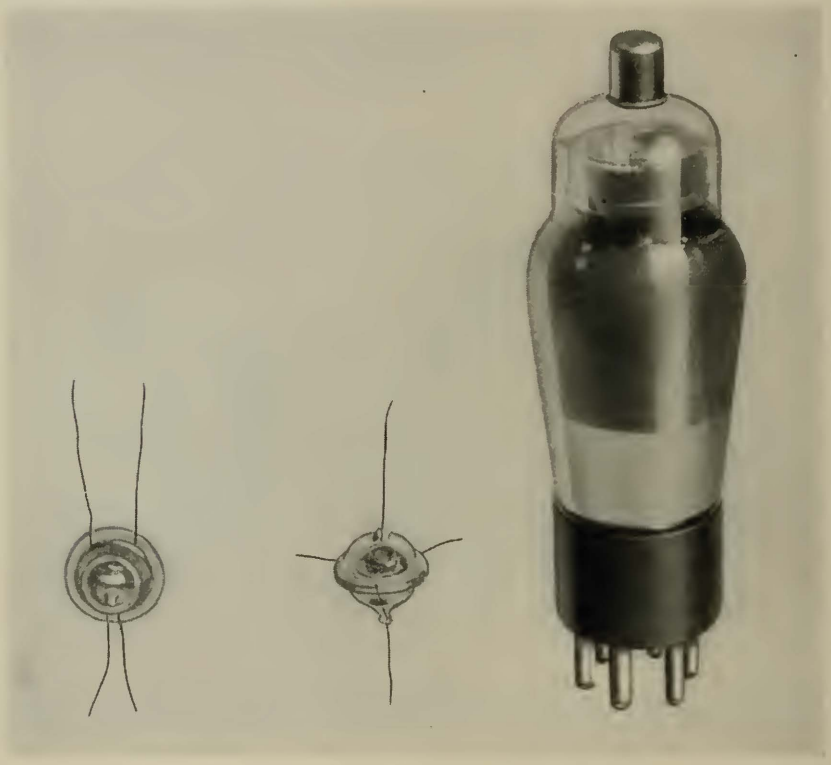


Fig. 13—Receiving tubes of extremely small dimensions. A conventional receiving type tube is shown at the right for comparison.

These tubes make use of parallel plane electrodes, the cathode being oxide coated and indirectly heated, and the grid being in the form of a mesh. A cross-section view of the triode is shown in Fig. 14. This tube will oscillate at frequencies up to 1000 megacycles in miniature replicas of the customary circuits used at longer wave-lengths. A photograph of a complete oscillator is shown in Fig. 15 and the circuit diagram in Fig. 16.

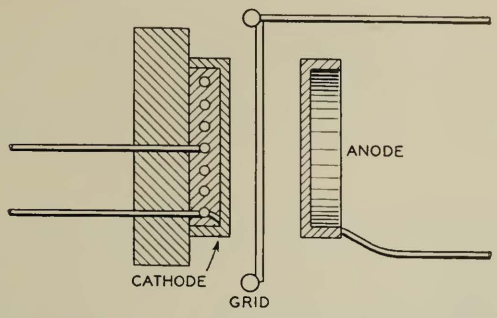


Fig. 14—A sectional view of the triode developed by Thompson and Rose.

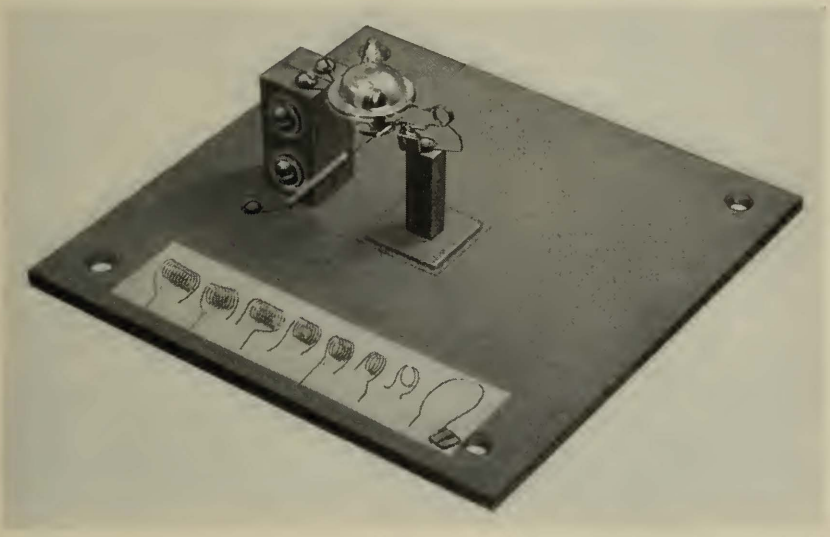


Fig. 15—A complete oscillator using a miniature receiving tube.

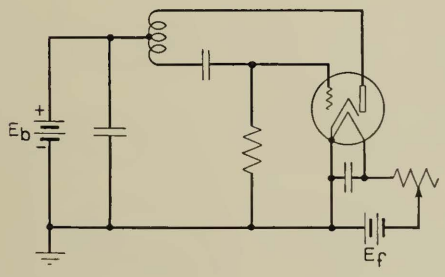


Fig. 16—The circuit diagram of the oscillator shown in Fig. 15.

Factors Limiting Ultra-High Frequencies

The limiting factor in the continued extension of the negative grid oscillator to higher and higher frequencies appears to be the dependence of physical size and output power on the operating frequency range. This dependence is well illustrated by the comparison in Fig. 17 of

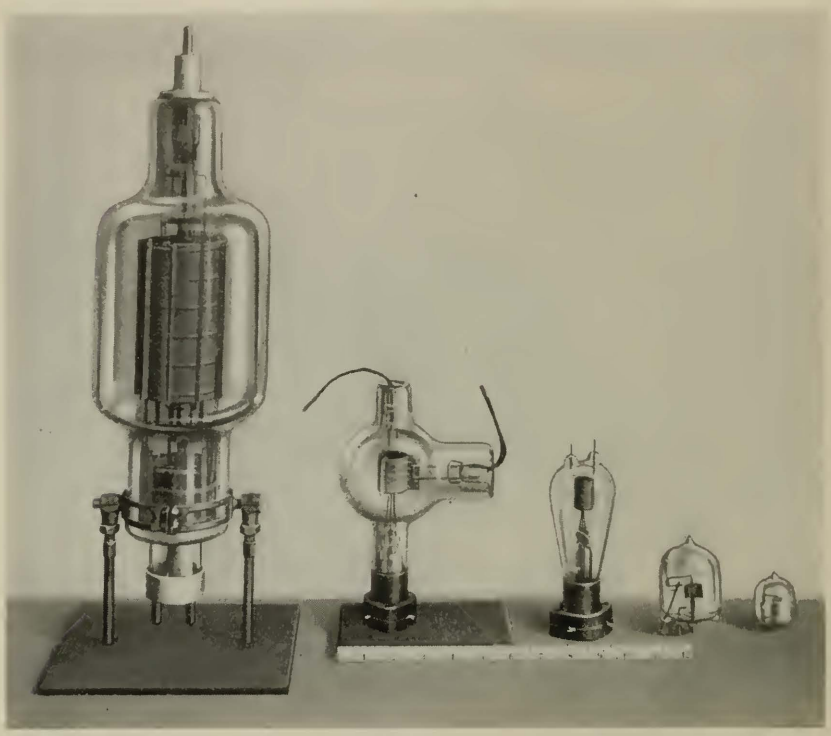


Fig. 17—Some of the tubes, previously discussed, compared in size with a lower frequency tube on the left.

some of the tubes so far discussed with a standard one-kilowatt tube for which the frequency limit is approximately 75 megacycles. Dimensional considerations indicate that the linear dimensions of a series of tubes of optimum design must be decreased in proportion to the operating wave-length. Since the heat dissipating ability depends upon the surface area of the plate, the output (assuming the same efficiency) will decrease as the square of the wave-length. That this is approximately true for the tubes shown in Figs. 5, 7, and 10 may be seen by reference to Fig. 18 where the outputs as a function of frequency are plotted on a logarithmic scale. The sloping lines are for

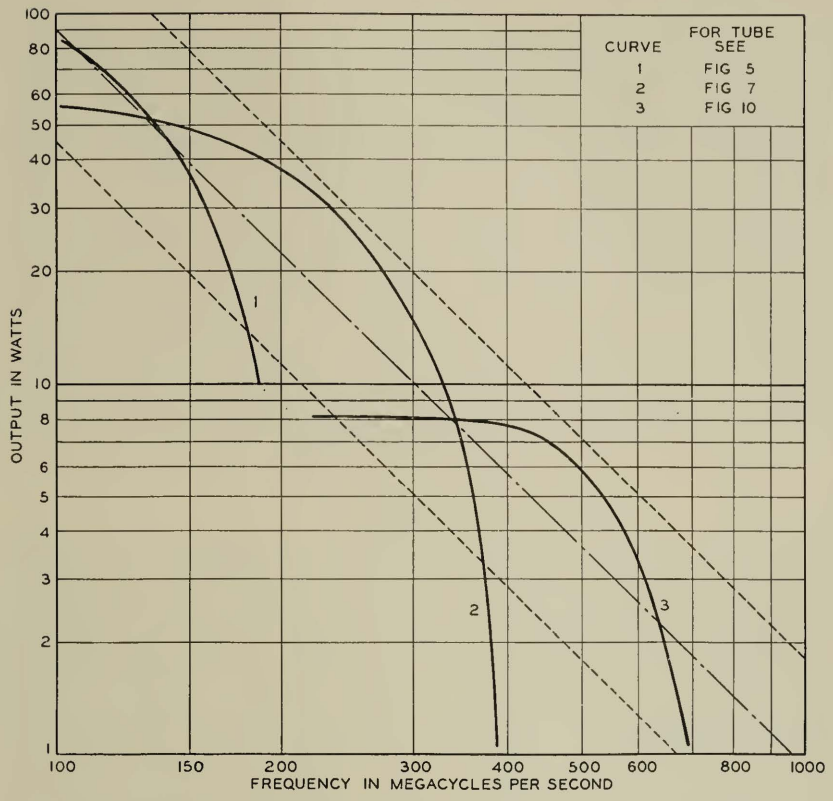


Fig. 18—A comparison plot of the outputs obtained with the tubes of Figs. 5, 7, and 10. The sloping lines are fixed values of the ratio of output to the square of wave-length.

different values of the ratio of output to the square of the wave-length. With radiation-cooled tubes patterned after those illustrated, outputs of only a few tenths of a watt at 3000 megacycles can be expected. If larger outputs are to be obtained, innovations in tube design must be made.

Positive Grid (Barkhausen) Oscillator

As first reported by Barkhausen and Kurz [7] in 1920, oscillations at frequencies greater than 300 megacycles can be produced by most high-vacuum triodes having symmetrical cylindrical structures when the grid is operated at a fairly high positive potential and the plate is held at or near the cathode potential. When so used they are variously known as oscillators of the Barkhausen and Gill-Morell [8] types after the earliest experimenters, or oscillators of the positive grid or re-

tarding field type to designate the arrangement of the electrode potentials. The relative ease with which such oscillations can be obtained at frequencies above 300 megacycles by the use of conventional tubes, and the widespread interest in this frequency range for communication purposes, have led to the appearance of a large number of papers on the experimental and theoretical aspects of such operation.

With the positive grid oscillator there are found to exist preferred frequencies of operation fixed by the electrode spacings and the applied electrode potentials. For the lowest preferred frequency mode of oscillation the relationship is such that the period of one complete oscillation is approximately equal to the total transit time of an electron which fails to strike the grid on its first transit, is retarded and finally turned back by the plate potential, and again missing the grid, returns to the cathode. Under these conditions the relationship,

$$\frac{E_g}{n^2} = \text{constant}$$

is found to hold approximately, where E_g is the applied grid potential, n is the frequency, and the constant is a function of the tube geometry. Other high-frequency modes of oscillation can be obtained. One of these is particularly easy to excite if the grid of the tube is in the form of a simple helix. The important rôle played by the electron transit time in determining the frequency of the positive grid oscillator contrasts sharply with the minor rôle it plays in determining the frequency of the negative grid oscillator.

For maximum output it is necessary to adjust the tuning of the external circuit to correspond to the preferred frequency fixed by the applied electrode potentials. The relative dependence of the frequency upon the circuit tuning and on the applied electrode potentials varies greatly with the design of the tube. In any case the improper adjustment of either parameter results in a marked decrease in output. In general it appears that the better the tube design and the higher the operating efficiency the greater will be the dependence of frequency upon circuit tuning and the less will be its dependence upon the applied electrode potentials.

The most efficient operation of the positive grid oscillator is obtained when the space current is limited by the cathode emission, as contrasted with the most efficient operation of the negative grid oscillator when the current is limited by space charge. Not only must the space current be emission limited but it must have a fairly critical value. This makes it necessary to adjust the cathode temperature critically. Since the cathode emission characteristics are apt

to change with time, frequent readjustments of the cathode temperature are usually required.

No completely satisfactory and generally accepted theory of the positive grid oscillator has as yet been given. Many theoretical papers dealing with the mechanism of oscillation have been published. Some of these papers resort to pictorial explanations which, from their very nature, must leave out certain basic factors. Readers interested in a résumé of the various theories are referred to the excellent review by Megaw [13] and to the original papers. It is now recognized that any accurate theory must be based upon a general consideration of all the forces acting upon the electrons in their flight between the electrodes. This may take the form of either a particular solution of the classical electromagnetic equations for the conditions within the tube or an analysis of the energy contributions due to individual electrons in their passage across the inter-electrode space.

Construction of a Positive Grid Tube

A representative positive grid tube of current design described by Fay and Samuel [2] before the International Scientific Radio Union is shown in Fig. 19. This tube differs from the conventional negative grid tube primarily in the construction of the grid and in the arrangement of the leads. While designed primarily for use in the frequency range from 500 to 550 megacycles, it illustrates the general problems encountered in the construction of the positive grid oscillator of this type for any frequency range.

The grid consists of a number of parallel wires supported by cooling collars at each end, the so-called squirrel cage construction. It will withstand 150 watts heat dissipation safely, and provides a minimum of circuit inductance and resistance. The grid diameter is fixed by the frequency for which the tube is designed and by the desired operating potential, such that the relationship

$$d_g = \frac{K_1\sqrt{E_g}}{n} \tag{1}$$

is approximately satisfied, where d_g is the diameter of the grid, K_1 is a constant, n the frequency, and E_g the applied grid potential.

An indefinite increase in output at a fixed frequency by the simultaneous increase in the grid diameter and in the applied grid potential is not possible because of the limited permissible grid dissipation per unit area. The optimum grid current is found to follow roughly a 3/2

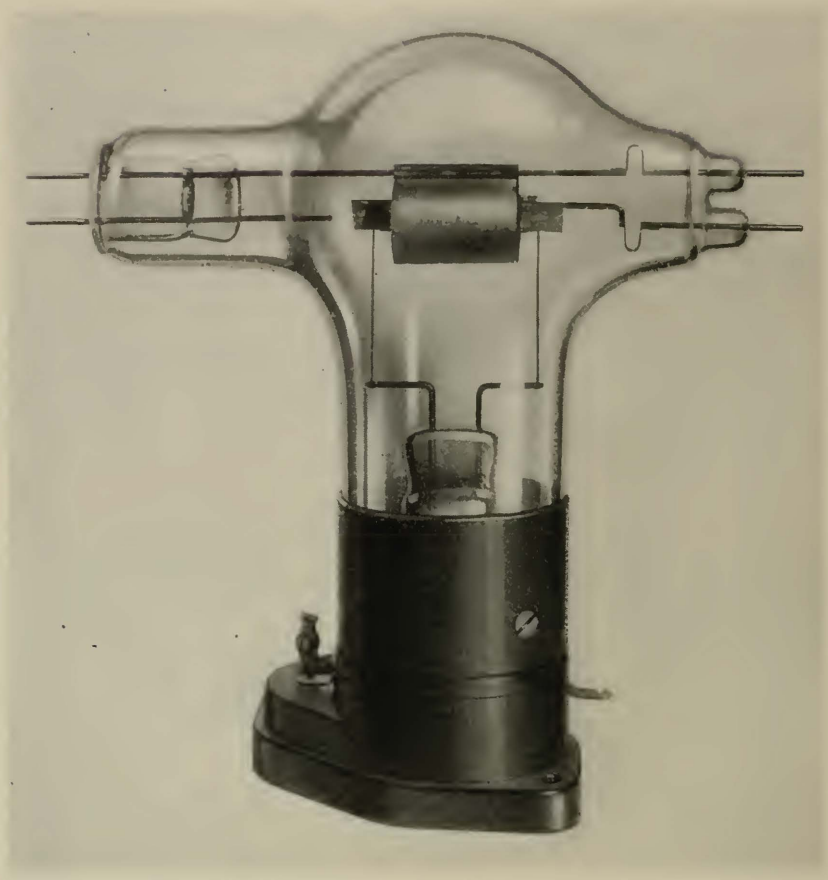


Fig. 19—A positive grid oscillator designed for the frequency range from 500 to 550 megacycles.

power law, that is

$$I_g = K_2 \frac{E_g^{3/2}}{d_g}, \tag{2}$$

so that the grid power will increase as the fourth power of the grid diameter while the grid area and hence the heat dissipating ability only increases as the first power of the diameter. Because of this an upper limit in output exists, fixed by the maximum permissible heat dissipation per unit area for the grid structure. The optimum grid diameter will vary directly with the wave-length for which the tube is designed, and if the ratio of the grid length to its diameter is main-

tained constant, the maximum available power output (assuming the same efficiency) will vary as the square of the desired wave-length.

Circuit of a Positive Grid Oscillator

In Fig. 20 is shown a diagram of a positive grid tube of the straight-wire-grid type and its associated circuit. Tuned circuits, in this case in the form of so-called Lecher systems, are connected between the grid and plate leads, extending approximately a half wave-length (30 cm) beyond the lead seals. Because of the existence of preferred

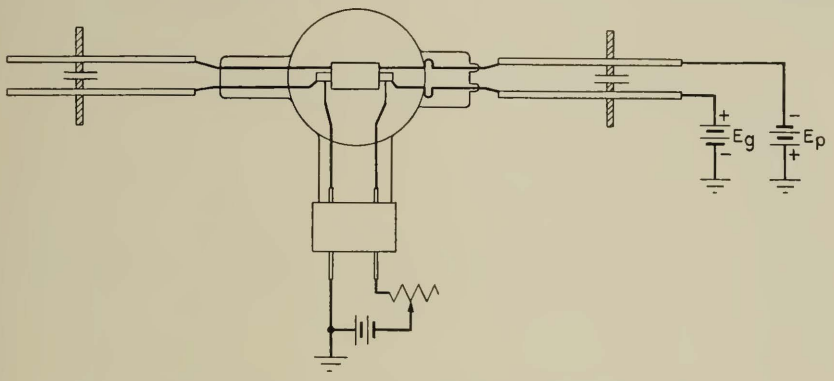


Fig. 20—A typical positive grid oscillator circuit.

frequencies of operation fixed by the potentials applied to the tube electrodes, distributed-constant circuits, if used, may be operated at frequencies corresponding to harmonic modes of oscillation. In this case the length of the leads within the tube envelope has been adjusted so that the glass seals come at or near potential nodal points for the Lecher systems of which the leads form a part. This minimizes dielectric losses in the glass. The effective paralleling of the two sets of leads greatly reduces the resistance losses, while the balanced arrangement decreases radiation losses. Strict attention to these details is required because of the already low efficiency of the mechanism of generation.

Characteristics of Positive Grid Oscillators

The dependence of output and anode efficiency on frequency is shown in Fig. 21. These data were taken by adjusting the circuit tuning, filament current, and the grid and plate potentials to their optimum values for each frequency. The curve showing the grid voltage will be observed to follow equation (1) above, at least roughly, and a similar correspondence will be observed between the curve for

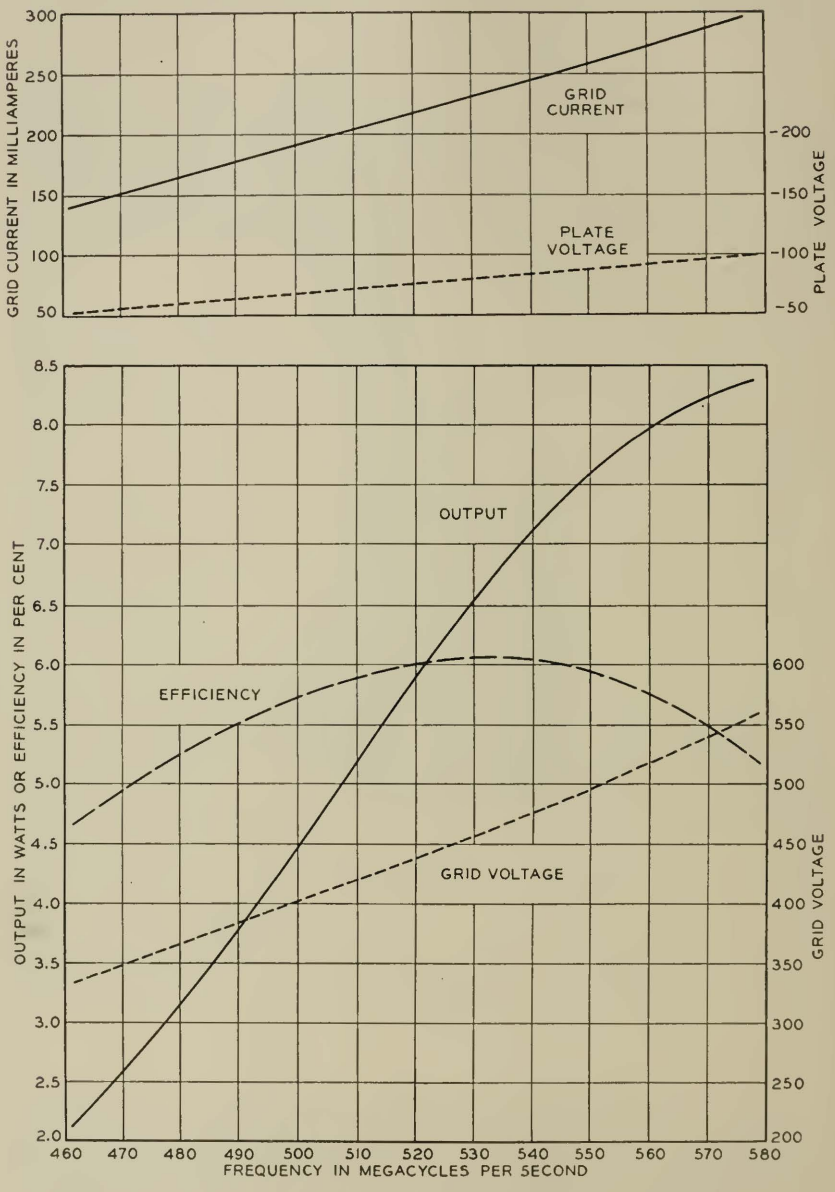


Fig. 21—Output and efficiency curves for the tube shown in Fig. 19.

the grid current and equation (2). Some variation in the required negative plate potential is observed. A maximum efficiency will be noted at a frequency of approximately 530 megacycles. The output, however, continues to increase with increasing frequency, the limit in output as well as in frequency being set by the safe grid dissipation. The outputs over the 500- to 600-megacycle range vary from 4.5 to 8 watts, comparing with outputs from 6 to 3 watts for the negative grid tube. The low efficiencies of 5 to 6 per cent are to be compared with the somewhat higher efficiencies of 19 to 11 per cent for the negative grid type tube.

The influence of the grid voltage on the frequency and on the output and efficiency is shown by the curves in Fig. 22. These data are for a

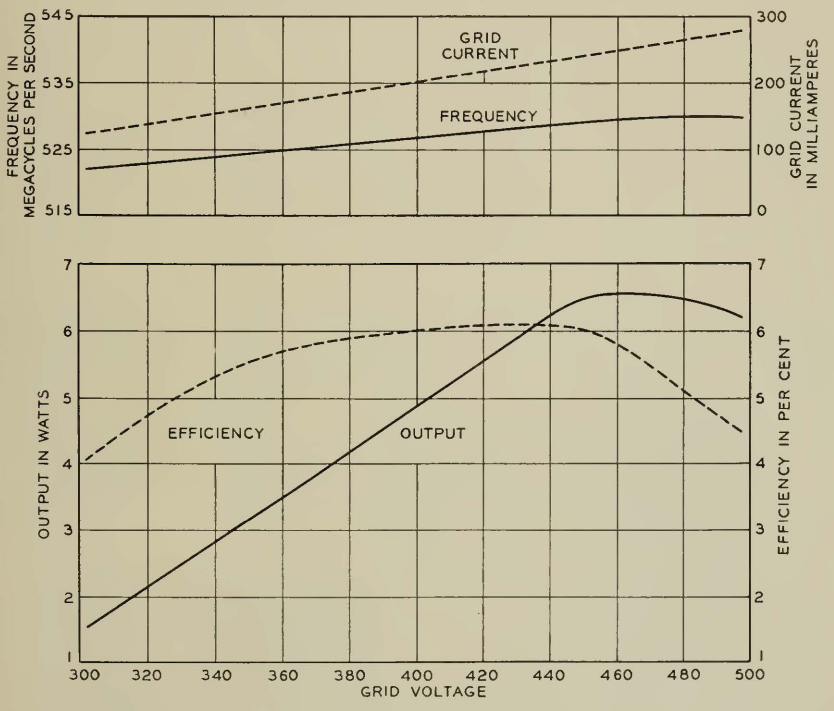


Fig. 22—Result of variation of grid voltage, tuning fixed, on tube of Fig. 19.

fixed circuit tuning, the grid voltage being adjusted to the values indicated. This corresponds to the condition that might obtain if a grid voltage modulation scheme were to be used. The lack of linearity of the output curve and the large shift in frequency indicate that amplitude modulation by this method would be unsatisfactory. The

frequency shift, large as it is, is much less than the shift observed in tubes of poorer design and correspondingly lower efficiency.

The dependence of the output and efficiency as well as frequency upon the grid current is shown in Fig. 23. These data were taken with

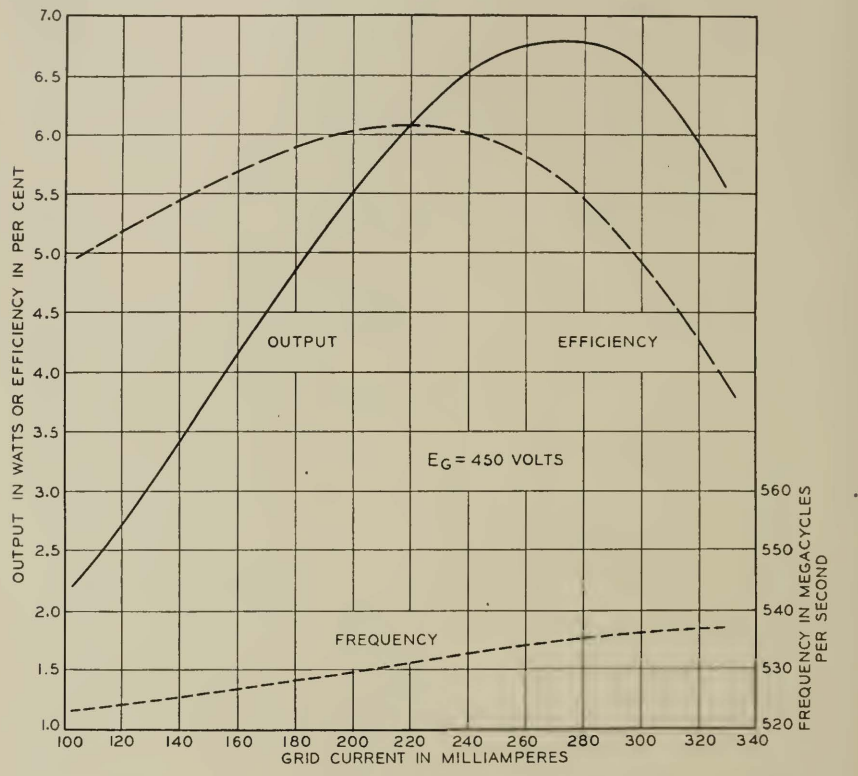


Fig. 23—Result of variation of grid current, other adjustments fixed, on tube of Fig. 19. f = 532 megacycles.

a grid potential of 450 volts and a fixed circuit adjustment corresponding roughly to a frequency of 532 megacycles. The current to the grid was varied by adjusting the temperature of the filament. The maxima observed in both the output and the efficiency correspond to conditions for which the grid current is limited primarily by the available emission rather than by space charge. As conditions corresponding to complete space charge are approached the output and efficiency fall off rapidly. The limit on the permissible grid dissipation prevents the extension of these curves to the condition of complete space charge. Because of this dependence of output on grid current, the adjustment of filament temperature is extremely critical.

At lower frequencies the efficiency of operation of a correctly designed positive grid tube is substantially the same as that exhibited by this tube. The negative grid oscillator on the other hand, as has been shown, increases both its output and efficiency rapidly with decreasing frequency. The positive grid oscillator is, therefore, at an increasing disadvantage at lower frequencies. With the present state of development, the negative grid oscillator will give larger outputs with higher anode efficiencies at all frequencies less than about 300 megacycles.

For frequencies much higher than 600 megacycles, it is found that the power input requirements for efficient operation of tubes having grids of the straight wire type are in excess of that which can be tolerated in the grid structures. Operation at very much less than optimum input results in considerable decrease in output as indicated in Fig. 25.

Spiral Grid Barkhausen Tubes

If the grid of a Barkhausen oscillator is in the form of a simple helix, oscillations at frequencies greater than those predicted by the relationship of equation (1) are readily obtained. When so constructed they are called spiral grid Barkhausen tubes. Some experimental models are shown in Fig. 24. The tubes used in the Lympne to St.

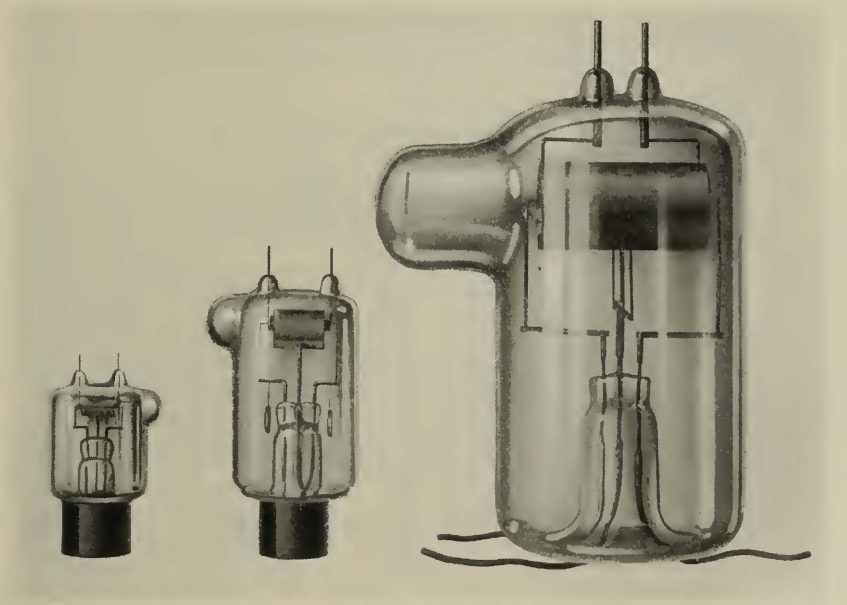


Fig. 24—Three optimum positive grid oscillators of the spiral grid type. Smallest tube designed for 2500 megacycles, largest for 500 megacycles.

Inglevert "micro-ray link" are of this general type.[14] Such tubes have been used to produce oscillations up to 3000 megacycles.

Because of the fact that the severe limitation on the optimum grid diameter is modified by the presence of a tuned circuit within the tube formed by the helical grid helix, tubes of this type are particularly useful at frequencies above 600 megacycles. The former restriction on grid diameter is replaced by the requirement that the expanded length of the grid spiral be approximately 1.24 times the wave-length at which the maximum output is required, and that there be a correct proportioning of the other dimensions of the tube. The dependence of wave-length on the grid wire length is illustrated by the experimental data in Table I covering a frequency range from 460 megacycles

TABLE I

DEPENDENCE OF WAVE-LENGTH ON GRID WIRE LENGTH FOR BARKHAUSEN OSCILLATOR WITH HELICAL GRID

Grid Wire Length in Cm.	Optimum Wave-length in Cm.	Ratio
16.3	13.5	1.21
18.6	18.6	1.00
19.7	14.5	1.36
20.4	18.0	1.13
21.4	17.5	1.22
21.3	25.0	0.85
22.3	20.2	1.10
25.2	18.6	1.35
30.6	25.0	1.22
32.0	23.6	1.36
32.0	25.5	1.25
33.4	25.0	1.33
42.6	29.0	1.47
42.6	29.5	1.44
42.6	30.2	1.41
42.6	30.7	1.38
53.2	43.5	1.22
80.0	65.0	1.23
Average		1.24

to 2220 megacycles. Graphic evidence of the independence of shape is shown by the largest and the smallest tube shown in Fig. 24 for which the dimensional ratios are nearly the same. The largest tube delivers several watts at 500 megacycles, while the smallest one delivers only a few tenths of a watt at 2500 megacycles. The efficiency in both cases is about one per cent. These dimensional considerations lead to the conclusion that there exists a maximum output at any given wave-length for a tube of a given design and that this output is proportional to the square of the optimum wave-length. From this it appears that the only advantage offered by the spiral grid tube over the other type is the simplification in mechanical design which

permits the construction of rigid grid structures capable of high energy dissipation for the higher frequency range.

The external tuned circuit for the higher frequency mode of oscillation takes the form of a Lecher system connected between the two grid terminals. When so connected the dependence of frequency upon circuit tuning is pronounced, as contrasted with the negligible dependence observed if the Lecher system is connected between the plate and the grid. When oscillating in the higher-frequency mode the spiral grid tube shows only a comparatively small dependence of frequency on grid potential and this may be compensated by a proportional change in plate potential. This, coupled with the fact that the output increases rapidly with increasing grid potential, makes it possible to apply various schemes of amplitude modulation. Characteristics of the type shown in Figs. 21, 22, and 23 for the straight-wire-grid tube cannot be taken except for a limited portion of the range due to the inability of the grid to dissipate the energy required in the upper portion of the grid voltage or grid current ranges.

While the spiral grid tube will also oscillate in the lower-frequency mode, its efficiency and output are considerably lower than the corresponding values for the straight-wire-grid tube previously discussed. Its field of usefulness is, therefore, largely limited to the higher frequency mode of oscillation in the frequency range above 600 megacycles.

The "Magnetron" Oscillator

The "magnetron" in its simplest form consists of a cylindrical diode or 2-electrode tube, with a uniform magnetic field in the direction of the electrode axis. The original type of tube has been largely superseded for ultra-high-frequency generation by the so-called split-plate magnetron, first used by Okabe,[18] in which the cylindrical anode is divided longitudinally into two (or more) segments to the terminals of which is connected the tuned circuit. Such a tube is shown in Fig. 25.

In the frequency range from 300 to 600 megacycles the split-plate magnetron compares favorably with the negative grid tube both in output and in anode efficiency. Its use has been limited because of the complicating factor of the magnetic field, and the attending modulation difficulties. For frequencies higher than 600 megacycles the magnetron provides larger outputs than those so far reported by other means. It has been used at frequencies up to 30,000 megacycles, a value well above that so far reported for any other type of vacuum tube.

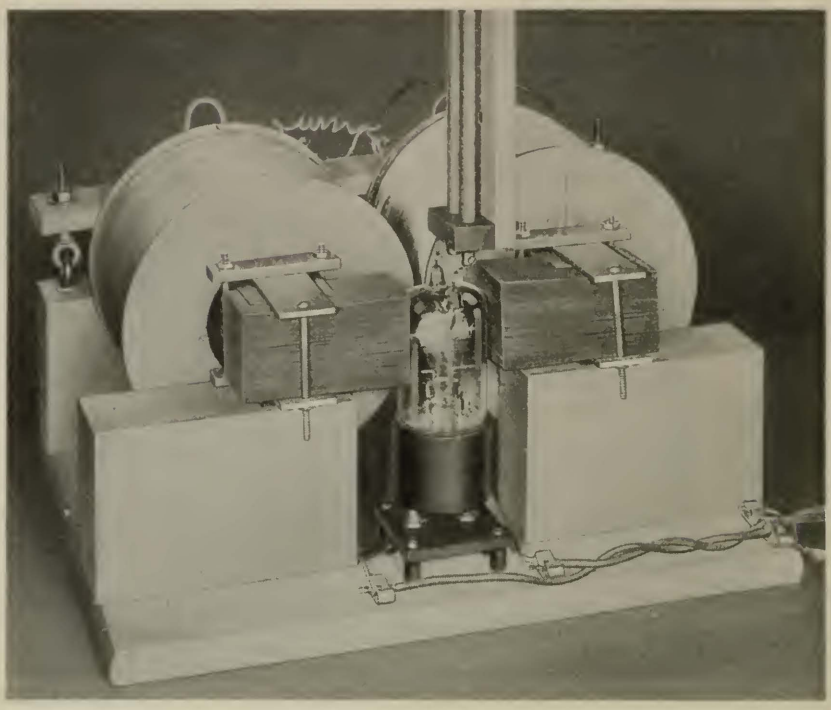


Fig. 25—An experimental model of the split-plate magnetron showing a possible arrangement of the magnetic field.

The magnetron depends for its operation upon the curvature of the electron orbits produced by the magnetic field. As first shown by Hull [15] in 1921, a critical field exists beyond which the anode current falls off rapidly to zero. This field is given by the relationship

$$H = \frac{6.72}{R}\sqrt{V}, \tag{3}$$

where R is the anode radius and V is the potential of the cylindrical anode with respect to an axial filament. Although the original magnetron of Hull and Elder made use of variations in the magnetic field in its operation as a generator, it was soon discovered that oscillators could also be produced with steady fields by two somewhat different mechanisms. The one, first pointed out by Habann,[16] makes use of a negative resistance effect observable in the static characteristics and the other, first described by Žáček,[17] involves the electron transit time in a way quite analogous to the way in which it is involved in the positive grid triode. Both mechanisms have been used to produce oscillations at ultra-high frequencies.

Negative Resistance Type

Data reported by McArthur and Spitzer [1] on a split-plate magnetron tube are illustrative of the negative resistance type of behavior. Static characteristics taken by varying the potential of one anode with the magnetic field held constant for different values of the potential on the other anode are shown in Fig. 26. The pronounced negative

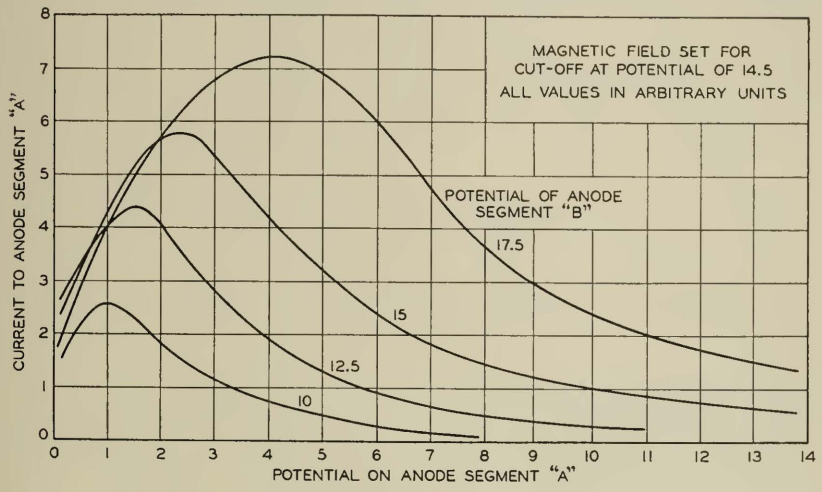


Fig. 26—Static characteristics of a split-plate magnetron.

resistance effect is obvious. This negative resistance characteristic can be utilized in producing oscillations.

Output and efficiency curves for this tube as an oscillator are shown in Fig. 27. These data were obtained by connecting a "tank"

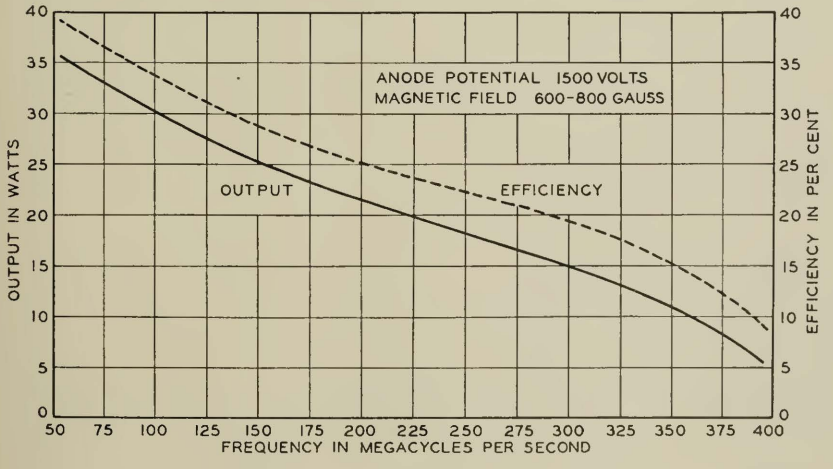


Fig. 27—Output and efficiency curves at different frequencies for split-plate magnetron.

circuit, tuned to the desired frequency, across the two anodes, as shown in Fig. 28. Each anode delivers energy to the oscillating circuit during alternate half-cycles, so that in effect, it is equivalent to a push-pull oscillator. The limiting frequency as set by the inter-electrode capacitances and lead inductances (corresponding to the similar limit for the negative grid tube) is 450 megacycles. The de-

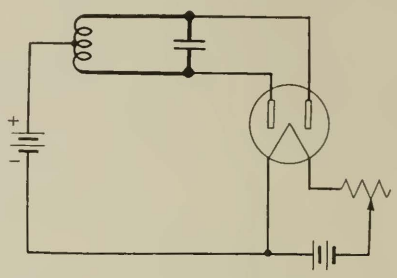


Fig. 28—Oscillation circuit of split-plate magnetron. Data for Fig. 27 taken in this type of circuit.

crease in output before this limit is reached is due to resistance and radiation losses and to the effect of electron transit time.

The magnetron, as contrasted with the negative grid tube, will oscillate with circuits having a high decrement. However, for its most efficient operation the effective anti-resonant impedance of the tuned circuit when loaded must be approximately 10 times the value required by a triode with the same anode dimensions. The load resistance that can be obtained at high frequencies is only a fraction of this value, so that the efficiency becomes increasingly less with higher frequencies. A further limitation is due to the fact that the electron current is concentrated on only a small part of the anode surface. This reduces the safe anode dissipation unless the anode is designed to have a high thermal conductivity. Because of these limitations, the ratio of output to the inter-electrode capacitance may be only slightly more favorable than the corresponding ratio for a triode of the same anode dimensions.

Type Depending on Electron Transit Time

When the magnetic field of a split-plate magnetron is adjusted to near the critical value given by equation (3), oscillations can be produced whose frequency will depend primarily upon the time of flight of electrons between filament and anode in a way closely resembling the behavior of the positive grid oscillator in its lower frequency mode of oscillation. For best output, the field must be above the critical

value. To fix the time of flight and hence the frequency of oscillation, the magnetic field and plate voltage must be adjusted to certain values roughly expressed by the empirical relationship

$$\lambda H = 13{,}100, \tag{4}$$

where λ is the wave-length in centimeters and H is the field strength in gausses, which must also satisfy equation (3). It is found that for best operation the magnetic and electric fields within the tube should not be exactly perpendicular. This lack of perpendicularity may be achieved either by tipping the magnetic field relative to the tube axis or by introducing end plates within the tube and maintaining them at a fixed positive potential.

Kilgore [19] has given complete information concerning this type of oscillator. In Fig. 29, taken from his paper, is shown the dependence

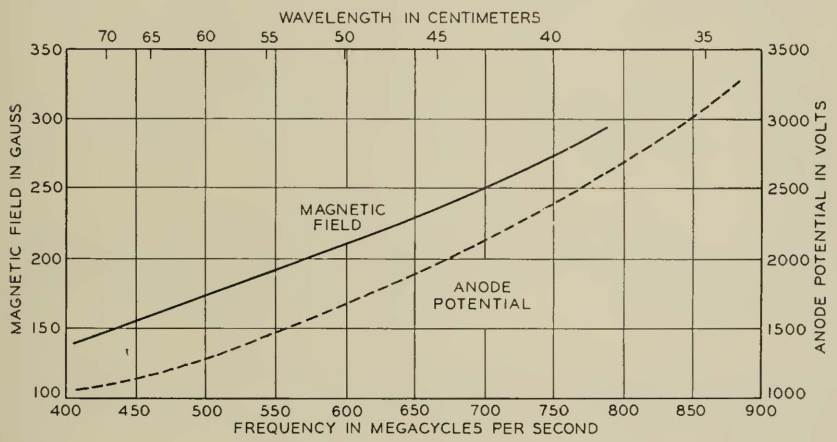


Fig. 29—Relation of magnetic field and anode potential to frequency of oscillation in magnetron oscillator of second type.

of field strength and anode potential on the desired frequency. The existence of a preferred frequency fixed by these values is confirmed by the data in Fig. 30 relating the output and wave-length with the length of the attached Lecher system. The decreased output shown at the second peak is due to the added losses introduced by the extended length of the system. The outputs shown on these curves are not in watts, but represent relative readings of the field strength near the oscillator. With optimum adjustments 7 watts at 715 megacycles is reported, the efficiency being about 8 per cent. The dependence of output and frequency on the applied anode potential

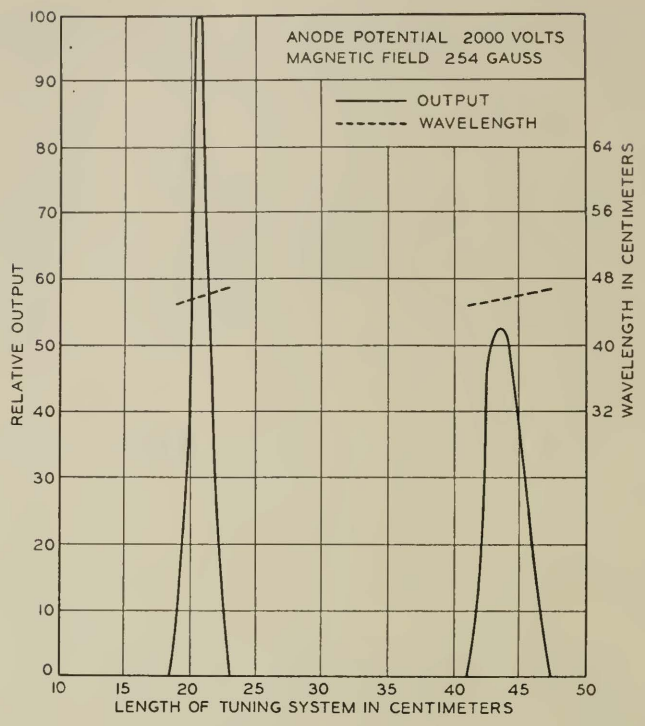


Fig. 30—Relation of wave-length and output to length of tuning system in magnetron oscillator of second type. Anode potential—2000 volts, Magnetic field—254 gausses.

is shown in Fig. 31, and the dependence of frequency on the current in the magnetic field coil in Fig. 32. The importance of the adjustment of the field angle is shown by the data in Fig. 33.

An output of 2.5 watts at 3160 megacycles has been reported by Wolff, Linder, and Braden.[22] They find that the efficiency of the tube is much improved by using end plates in place of the tipped magnetic field. Cleeton and Williams [21] have been able to obtain oscillations at 30,000 megacycles with a magnetron tube.

Amplification

The use of the conventional thermionic triode as an amplifier greatly exceeds its use as an oscillation generator in communication applications. Its ability to amplify has contributed much more to the development of our present-day long distance communication, whether by wire or by radio, than has its ability to oscillate. The complete utilization of ultra-high frequencies as carrier channels in communica-

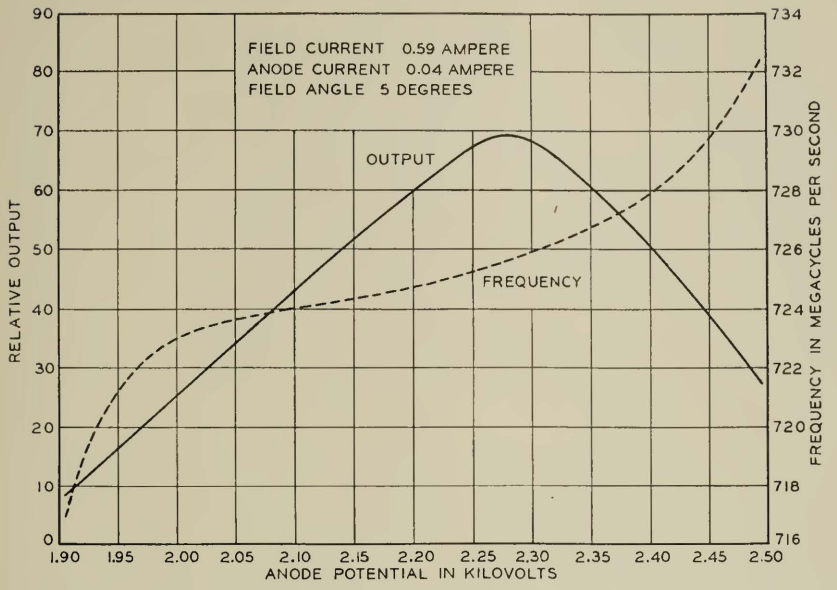


Fig. 31—Relation of output and frequency to anode potential in magnetron oscillator of second type.

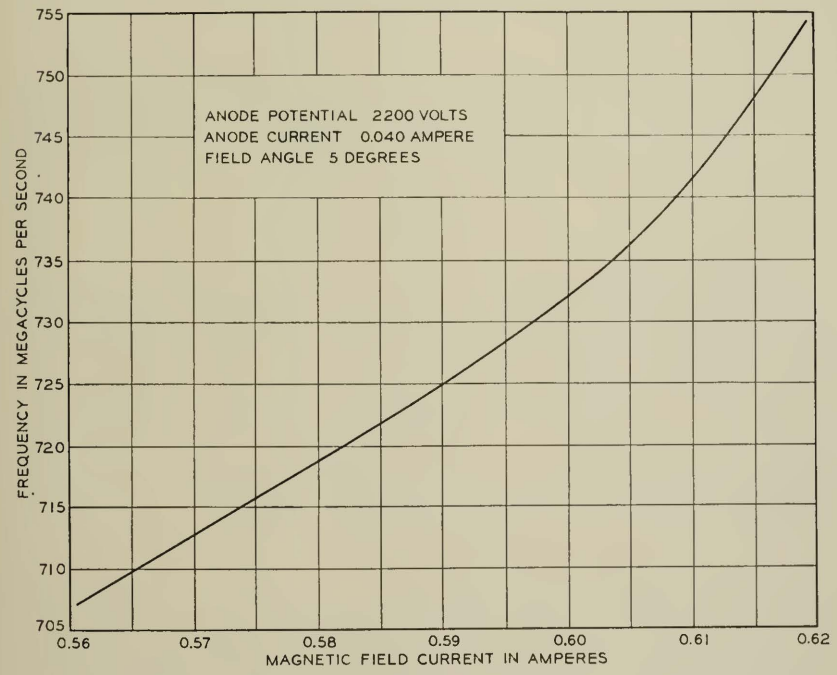


Fig. 32—Relation of frequency to magnetic field current in magnetron oscillator of second type.

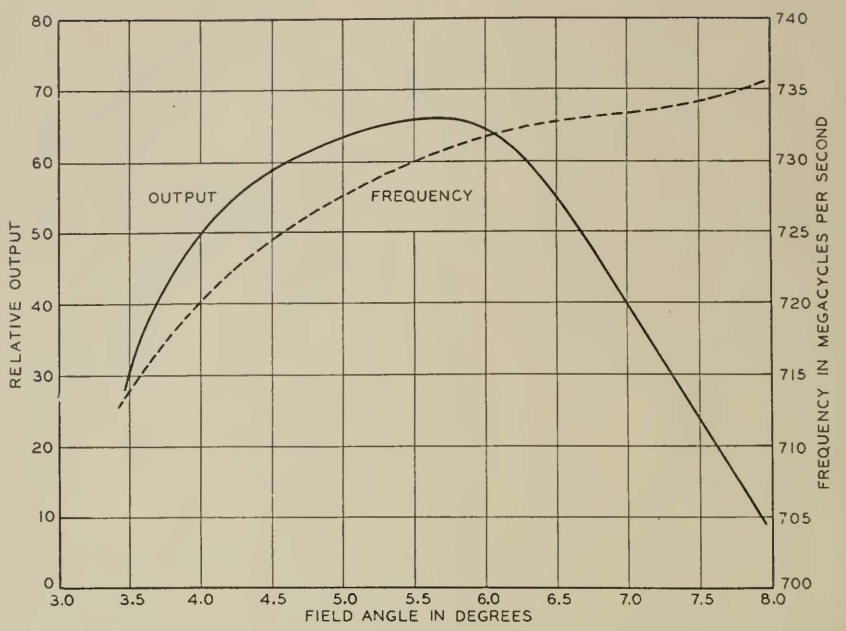


Fig. 33—Relation of output and frequency to field angle in magnetron oscillator of second type.

tion will also, no doubt, be dependent upon the development of suitable amplifiers for this frequency range. Although certain forms of pseudo-amplification are possible with tubes of the Barkhausen and magnetron types, the negative grid triode and multi-element tubes derived therefrom are the only devices available for very high frequencies which will amplify in the sense that the output is an enlarged undistorted replica of the input.

As the frequency of operation of the negative grid triode is increased, difficulties in securing stable operation as an amplifier and in realizing the full gain indicated by the tube constants are encountered. These difficulties, as is well known, are in the main due to the tendency of the amplifier to oscillate or "sing" because of feed-back through the grid-plate capacitance. This may be overcome either by the introduction of a compensating capacitance somewhere in the circuit, so-called neutralization, or by the introduction of an electrostatic shield or screen within the tube envelope between the grid and plate, giving the screen-grid tetrode. Neutralization schemes fail at very high frequencies because of the inductance of the tube leads which makes difficult the correct location of the neutralizing capacity and because of transit-time effects which shift the phase of the needed compensa-

tion. However, conventional screen-grid tetrodes and pentodes are available which function satisfactorily over the major portion of the frequency range covered by the conventional 3-element tube as an oscillator.

For frequencies above approximately 60 megacycles specially designed tubes are required. Because of the similarity in the special frequency requirements, it is expected that there will be found a succession of multi-element tubes for amplification use, each rated for a band of frequencies, patterned after corresponding triode oscillators. The special frequency requirements for the amplifying tube are even more severe than those for the triode oscillator, so that the multi-element amplifying tube will in general cease amplifying at a frequency somewhat lower than the frequency limit of oscillation of the corresponding triode oscillator.

Thompson and Rose [3] have described small screen-grid tubes which will amplify at frequencies of 300 to 400 megacycles. One of these tubes is shown in Fig. 13. Their characteristics are similar to those of the conventional screen-grid tube in many respects. The very great reductions in inter-electrode capacitances, lead inductances, and transit time make possible the construction of receiving circuits using tuned radio frequency amplification at these very high frequencies. The ratio of the frequency limits of the corresponding triode as an oscillator (1000 megacycles) to the frequency at which amplification was reported (400 megacycles) is typical and illustrates the apparently inevitable failure of the amplifier to keep pace with the oscillator in the struggle toward higher and higher frequencies.

References

Negative Grid Tubes

1. E. P. McArthur and E. E. Spitzer, "Vacuum Tubes as High-Frequency Oscillators," *I. R. E. Proc.*, v. 19, 1931, p. 1971. (Bibliography of 28 titles.)
2. C. E. Fay and A. L. Samuel, "Vacuum Tubes for Generating Frequencies Above 100 Megacycles," *I. R. E. Proc.*, v. 22, 1934, p. 679.
3. B. J. Thompson and G. M. Rose, "Vacuum Tubes of Small Dimensions for Use at Extremely High Frequencies," *I. R. E. Proc.*, v. 21, 1933, p. 1707.
4. W. H. Wenstrom, "An Experimental Study of Regenerative Ultra-Short Wave Oscillators," *I. R. E. Proc.*, v. 20, 1932, p. 113.
5. I. E. Mouromtseff and H. V. Noble, "A New Type of Ultra-Short Wave Oscillator," *I. R. E. Proc.*, v. 20, 1932, p. 1328.
6. W. Kroebel, "On the Production of Undamped Oscillations of Wave-Lengths of the Order of a Decimeter in a Reaction Circuit I," *Ann. der Physik*, Series 5, v. 14, 1932, p. 80.

Positive Grid Tubes

7. H. Barkhausen and K. Kurz, "The Shortest Waves Obtainable with Vacuum Tubes," *Phys. Zeits.*, v. 21, 1920, p. 1.
8. E. W. B. Gill and J. H. Morrell, "Short Electric Waves Obtained by Valves," *Phil. Mag.*, v. 44, 1922, p. 161.

9. K. Kohl, "Continuous Ultra-Short Electric Waves," *Ergebnisse der exakten Naturwissenschaften*, v. 9, 1930, p. 275. (Bibliography of 135 titles.)
10. G. Potapenko, "Investigations in the Field of Ultra-Short Electro-Magnetic Waves," *Phys. Rev.*, v. 39, 1932, p. 625; v. 39, 1932, p. 638; v. 40, 1932, p. 988; v. 41, 1932, p. 216. (Many references to the literature.)
11. H. Collenbusch, "Investigations of Very Short Wave Tubes," *Ann. der Physik*, Series 5, v. 13, 1932, p. 191.
12. H. N. Kozanowski, "A New Circuit for the Production of Ultra-Short-Wave Oscillations," *I. R. E. Proc.*, v. 20, 1932, p. 957.
13. E. C. S. Megaw, "Electronic Oscillations," *I. E. E. Jl.*, v. 72, 1933, p. 313. (Bibliography of 47 titles.)
14. H. G. Clavier, "Production and Utilization of Micro-Rays," *Elec. Communication*, v. 12, 1933, p. 3

Magnetron Tubes

15. A. W. Hull, "The Effect of a Uniform Magnetic Field on the Motion of Electrons Between Coaxial Cylinders," *Phys. Rev.*, v. 18, 1921, p. 31.
16. E. Habann, "A New Vacuum Tube Generator," *Zeit. f. Hockfreq.*, v. 24, 1924, p. 115.
17. A. Žáček, "A Method for the Production of Very Short Electromagnetic Waves," *Zeit. f. Hockfreq.*, v. 32, 1928, p. 172.
18. K. Okabe, "Production of Intense Extra-Short Electromagnetic Waves by Split-Anode Magnetron," *I. E. E. Jl.*, Japan, 1928, p. 284. See also *I. R. E. Proc.*, v. 17, 1929, p. 652; and *I. R. E. Proc.*, v. 18, 1930, p. 1748.
19. G. R. Kilgore, "Magnetostatic Oscillators for Generation of Ultra-Short Waves," *I. R. E. Proc.*, v. 20, 1932, p. 1741.
20. E. C. S. Megaw, "An Investigation of the Magnetron Short Wave Oscillator," *I. E. E. Jl.*, v. 72, 1933, p. 326. (13 references.)
21. C. E. Cleeton and N. H. Williams, "A Magnetostatic Oscillator for the Generation of 1 to 3-Cm. Waves," *Phys. Rev.*, v. 44, 1933, p. 421.
22. I. Wolff, E. G. Linder, and R. A. Braden, "Transmission and Reception of Centimeter Waves," *I. R. E. Proc.*, v. 22, 1934, p. 555.

General

23. F. B. Llewellyn, "Vacuum Tube Electronics at Ultra-High Frequencies, *I. R. E. Proc.*, v. 21, 1933, p. 1531.
24. W. E. Benham, "Theory of the Internal Action of Thermionic Systems at Moderately High Frequencies," Part I—*Phil. Mag.*, March 1928, p. 641. Part II—*Phil. Mag. A*, v. 11, 1931, p. 457.
25. L. Rohde, "On Transmitter Tubes for Producing Waves One Meter Long," *Hochfreq. und Elektroakustik*, v. 40, 1932, p. 3.
26. W. C. White, "The Selection of Types of Vacuum Tubes as Sources of High Frequency Power," *G. E. Rev.*, v. 36, 1933, p. 394.
27. W. H. Wenstrom, "Historical Review of Ultra-Short-Wave Progress," *I. R. E. Proc.*, v. 20, 1932, p. 95. (Bibliography of 54 titles.)

Horizontal Rhombic Antennas *

By E. BRUCE, A. C. BECK and L. R. LOWRY

The paper discusses the theoretical methods employed by the authors in dimensioning horizontal rhombic receiving antennas. Experimental proof is given of the engineering accuracy of the directivity calculations on which this work is based. There are included brief descriptions of the antenna-to-transmission line coupling circuits and the resistance terminations for rhombic antennas.

Introduction

AN introductory discussion has been given in a previous paper [1] of a type of antenna which maintains a desirable degree of directivity throughout a broad continuous range of frequencies. This structure was descriptively termed the "diamond-shaped" antenna in that paper, but has since become known as the "rhombic antenna" and will be so designated here.

This paper discusses, in some detail, the theoretical methods employed by the authors in dimensioning horizontal rhombic receiving antennas from data obtained by preliminary surveys of the incident-plane angles of wave-arrival at the proposed receiving site.

Experimental proof is given of the engineering accuracy of the directivity calculations on which this work is based. Checking measurements were made on small-scale rhombic antennas operating at correspondingly short wave-lengths. Confirming data were also obtained on large, adjustable rhombic antennas during the reception of European signals.

The paper also includes a brief discussion of the antenna-to-transmission line coupling circuits employed and, in addition, the resistance terminations, located at the end of the antenna remote from the receiver, used for suppressing standing waves and promoting unidirectivity. Some of the performance curves obtained on these devices are reproduced.

Antenna Dimensions

Before designing any highly directive receiving antenna system, it is desirable to have a knowledge of the direction of arrival, the angular spread, and the angular variation in direction of the waves to be

* Published in *Proc. I. R. E.*, January, 1935.

[1] "Developments in Short-Wave Directive Antennas," by E. Bruce, *Proc. I. R. E.*, August, 1931; *Bell Sys. Tech. Jour.*, October, 1931.

received. Such information has been obtained by employing the methods described by Friis, Feldman and Sharpless.[2]

In general, the principal axis of the antenna departs very little from the true bearing of the station to be received. However, the average vertical direction of the waves in the incident plane and their angular variation have an effect upon the determination of the correct dimensions of a horizontal rhombic receiving antenna.

Without careful consideration, one might be led to believe that the theoretical effectiveness of a horizontal rhombic receiving antenna would increase without limit, for a stable wave-direction, as the properly related dimensions are increased. It will be shown that, for a given incident wave-angle above the horizontal, the optimum dimensions have quite definite values.

The essential dimensions of this type of antenna are its height above ground, the length of a side element, and the inclination angle of these elements in respect to the wave-direction. At frequencies higher than about fifteen megacycles, this type of antenna is relatively so inexpensive that it is usually economically possible to select a combination of these dimensions which will produce the maximum signal output.

At lower frequencies, the increasing cost, particularly of the pole height required for maximum output, demands that a careful analysis be made of effective compromise designs. One such design method involves a sacrifice in antenna height, which sacrifice can be partially compensated for by an increase in element length. Thus the economic balance involves a weighing of supporting structure costs against the cost of land. Again, cases may arise where the recommended antenna height is practical but a sacrifice in element length is essential. The effect of this sacrifice on the directive pattern can be partially compensated for by a readjustment of the tilt-angle of the elements. The directive pattern may also be aligned with the wave-angle when both height and element length are reduced, but only at an output sacrifice.

The designer's first problem is to choose between the maximum output arrangement and the various compromise designs. In the case of a transmitting antenna, the problem is to create as large a field as possible at the distant receiving point, for a given amount of transmitter power. As long as the increase in antenna cost is smaller than the corresponding cost of an increase in transmitter power, no compromise should be made in the antenna dimensions. It is, therefore, obvious that the choice, between optimum and compromise dimensions for a transmitting antenna, is a function of the cost of the transmitter to be employed.

[2] "The Determination of the Direction of Arrival of Short Radio Waves," by H. T. Friis, C. B. Feldman and W. M. Sharpless, *Proc. I. R. E.*, January, 1934.

For a receiving antenna, the dividing line between the optimum and the compromise design methods is often defined by the relative importance, as circuit limiting factors, of static as compared with set-noise. The optimum design results in a large signal output, from the antenna, which enables the over-riding of set-noise and it also results in effective directional discrimination against static. The compromise height arrangement maintains and sometimes improves upon the directional discrimination against static but sacrifices a part of the possible antenna output-signal level. If the frequency is sufficiently low so that static rather than set-noise is practically always the circuit limitation, no real harm will result from using the compromise height design with its accompanying economic saving.

In general, compromise in element length results in a loss in both the antenna output-signal level and in static discrimination. This procedure is recommended only for exceptional cases where a restriction in available land exists, or where a broad directional characteristic is required because of a highly variable wave-direction.

Directivity Equations

In the appendix of this paper will be found the derivation, for the stated assumptions, of the three dimensional directivity equation for a horizontal rhombic receiving antenna terminated by its characteristic impedance. The principal antenna dimensions will be apparent by reference to Fig. 1.

If the horizontal component β, of the angle made by the wave-direction with the principal antenna axis, is set equal to zero in the final resulting equation in the appendix, the directivity equation for the incident plane passing through the principal axis is as follows:

$$I_R = k\left[1 + ae^{-j\left(\frac{4\pi}{\lambda}H\sin\Delta + \alpha\right)}\right] \cdot \left[\frac{\cos\phi}{1 - \sin\phi\cdot\cos\Delta}\right] \cdot \left[1 - e^{-j\frac{2\pi l}{\lambda}(1 - \sin\phi\cdot\cos\Delta)}\right]^2, \quad (1)$$

where I_R = receiver current,
k = proportionality factor,
H = height above ground,
l = wire element length,
ϕ = one-half of side apex angle,
Δ = angle made with ground by wave-direction in incident plane,
a = amplitude ratio of ground reflected field to incident field,
α = apparent phase angle lag caused by ground reflection,
λ = wave-length.

Since the principal axis of the antenna is nearly always directed toward the distant transmitter, the conditions which will maximize the above expression are of prime interest. The determination of these conditions is greatly simplified by the assumption of a perfect ground.

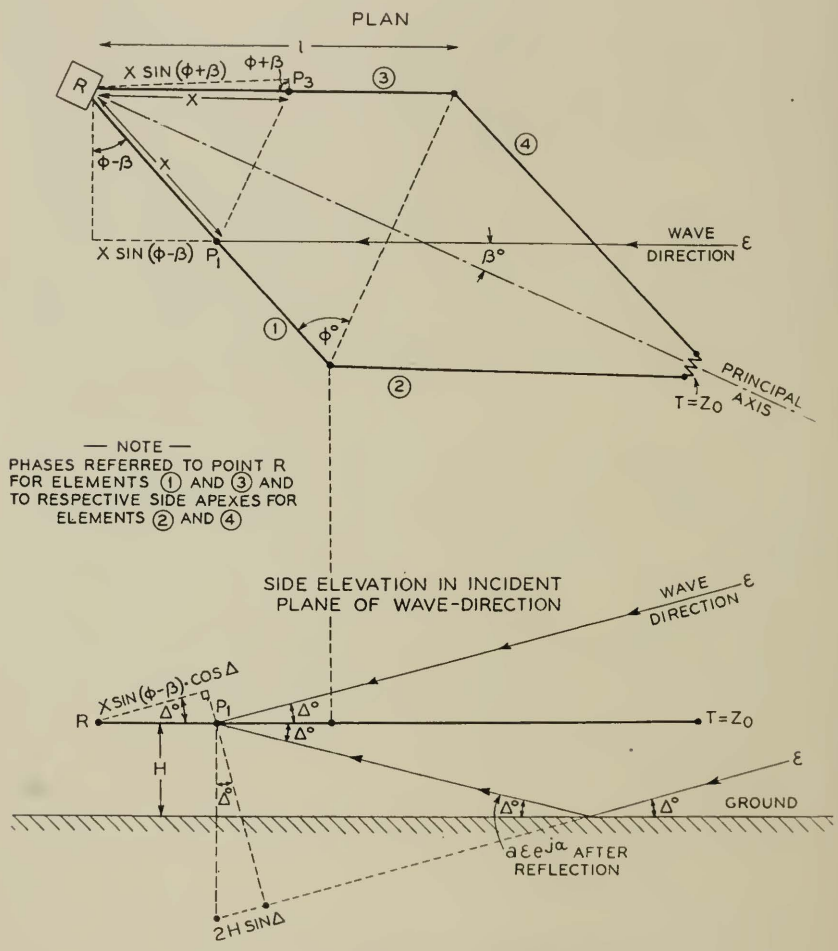


Fig. 1—Horizontal rhombic antenna dimensions.

Fortunately this is not a radical assumption for the case of horizontally polarized waves at the usual incident angles. This may be verified by reference to Fig. 2 where the values of a and α are plotted for several ordinarily encountered ground constants.

Using the perfect ground assumption, the amplitude of equation (1) is given by

$$I_R = k' \left[\sin \left(\frac{2\pi H}{\lambda} \sin \Delta \right) \right] \cdot \left[\frac{\cos \phi}{1 - \sin \phi \cdot \cos \Delta} \right] \cdot \left\{ \sin^2 \left[\frac{\pi l}{\lambda} (1 - \sin \phi \cdot \cos \Delta) \right] \right\}. \quad (2)$$

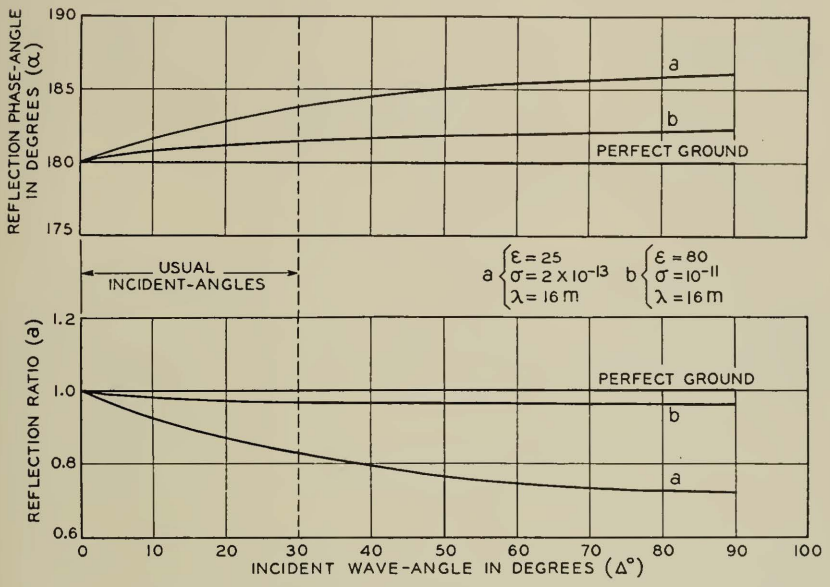


Fig. 2—Characteristics of wave-reflections from ground.

In this equation the first bracketed term may be referred to as the "height" factor, the second as the "asymmetrical directivity" factor, and the third as the "phasing" factor.

Maximum Output Design Method

In maximizing equation (2), it is necessary to deal with three variables, the antenna dimensions H, l and ϕ. Differentiating with respect to each of these variables, while holding the other two constant, and equating to zero, the following three expressions are obtained:

$\frac{\delta I_R}{\delta H} = 0$ when

$H = \frac{\lambda}{4 \sin \Delta}$ for the lowest practical height,

$\frac{\delta I_R}{\delta l} = 0$ when

$$l = \frac{\lambda}{2(1 - \sin \phi \cdot \cos \Delta)} \quad \text{for the major ear of the directive diagram,} \tag{4}$$

$\frac{\delta I_R}{\delta \phi} = 0$ when

$$\tan \left[\frac{\pi l}{\lambda} (1 - \sin \phi \cdot \cos \Delta) \right] = - \frac{2\pi l \cdot \cos^2 \phi \cdot \cos \Delta (1 - \sin \phi \cdot \cos \Delta)}{\lambda (\sin \phi - \cos \Delta)} . \tag{5}$$

Substituting equation (4) into (5), it is found that,

$$\sin \phi = \cos \Delta. \tag{6}$$

This result determines the fact that, regardless of antenna height, the best tilt-angle ϕ is the complement of the wave-angle Δ, where the optimum element length is used. Taking this result and substituting it back into either (4) or (5) results in

$$l = \frac{\lambda}{2 \sin^2 \Delta} . \tag{7}$$

The value of "l" in equation (7) together with the height given by equation (3), and the tilt-angle given by equation (6) determine the dimensions of the horizontal rhombic antenna with maximum output for any given wave-angle Δ.

As an example of the use of the above equations, Fig. 3 is the resulting incident plane directive pattern with the antenna designed for the given wave-angle of 17.5 degrees. Figure 4 is a plot of the antenna output versus the azimuth angle for the fixed incident angle of 17.5 degrees as obtained by using these dimensions in the three-dimensional directivity equation in the appendix.

Note in Fig. 3 that the directive pattern does not have its maximum radius at the line indicating the given wave-direction even though the greatest possible amplitude for that wave-angle has been determined.

It is believed that this method of design has a definite application where over-riding set-noise by the largest possible signal output is paramount. In cases where discrimination against random static is desirable, or where the received wave-direction is unstable, an alignment of the mean wave-direction with the optimum radius of the

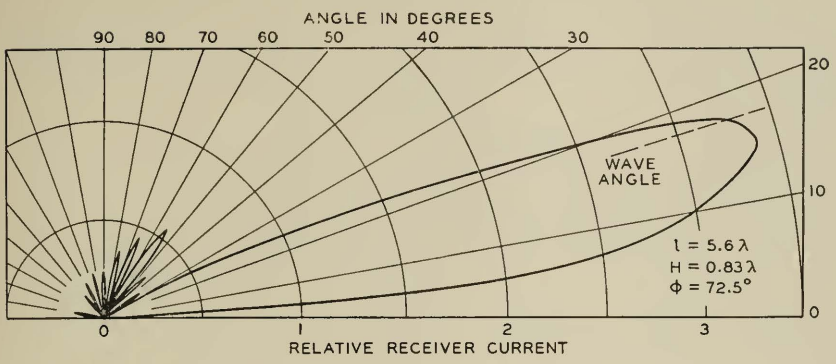


Fig. 3—Maximum output design. Receiver current diagram of incident plane directivity for a 17.5° wave-angle.

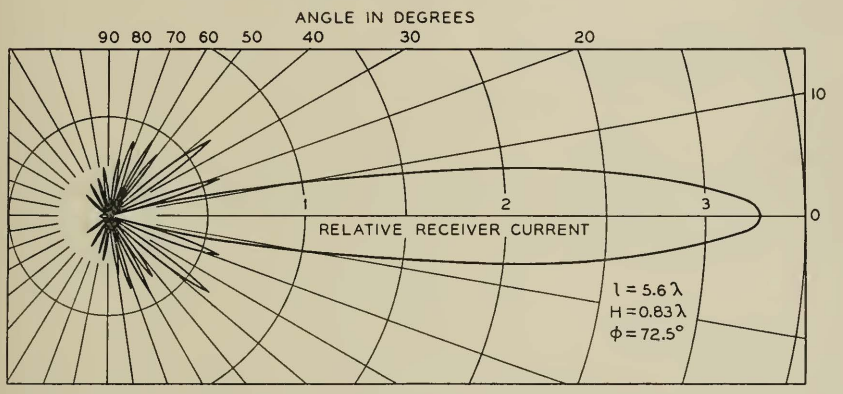


Fig. 4—Maximum output design. Receiver current versus azimuth angle for a 17.5° wave-angle.

directive diagram becomes necessary. Such an alignment can be achieved by the method described below, but only by a small sacrifice in the amplitude [3] of I_R at the required wave-angle.

The Alignment Design Method

Since by our previous method of design the height factor of equation (2) was aligned with the wave-angle, it is necessary to change either or both of the dimensions l and ϕ to secure the best output under the limiting alignment condition. This limiting condition is obtained as

[3] To compare the energy gain of one antenna with another, the amplitudes of I_R, which are proportional to effective voltages, should be corrected for the differences in radiation resistance of the antennas.

follows:

$\frac{\delta I_R}{\delta \Delta} = 0$ when

$$II = \sin\phi \cdot \tan\Delta \cdot \tan\left(\frac{2\pi II}{\lambda}\sin\Delta\right) \cdot \left\{\frac{\lambda}{2\pi(1 - \sin\phi\cdot\cos\Delta)} - \frac{l}{\tan\left[\frac{\pi l}{\lambda}(1 - \sin\phi\cdot\cos\Delta)\right]}\right\}. \quad (8)$$

Substituting (3) into (8) gives

$$\tan\left[\frac{\pi l}{\lambda}(1 - \sin\phi\cdot\cos\Delta)\right] = 2\left[\frac{\pi l}{\lambda}(1 - \sin\phi\cdot\cos\Delta)\right]. \quad (9)$$

This is a transcendental equation of the form $\tan x = 2x$, whose roots, to four significant figures, are

$x = 0, \quad 0.3710\pi, \quad 1.466\pi, \quad 2.480\pi, \quad 3.486\pi, \quad$ etc.

Alignment of the major ear of the directive pattern with the wave-angle occurs only for the first solution greater than zero. Therefore we have for this condition,

$$l = \frac{0.371\lambda}{1 - \sin\phi\cdot\cos\Delta}. \quad (10)$$

To obtain the dimensions for greatest output in the alignment design case, l and ϕ in equation (2) are no longer independent variables, but are related by equation (10). We can, therefore, eliminate either one we choose from equation (2) and maximize this new equation for the required dimensions.

Substituting (10) into (2) gives:

$$I_R' = k''\left[\sin\left(\frac{2\pi II}{\lambda}\sin\Delta\right)\right]\left[\frac{\cos\phi}{1 - \sin\phi\cos\Delta}\right] \quad (11)$$

and

$$\frac{\delta I_R'}{\delta\phi} = 0 \text{ when} \quad (12)$$

$$\sin\phi = \cos\Delta.$$

Since equation (12) is identical with equation (6), the recommended design for directive pattern alignment with the wave-angle is obtained by changing only the antenna length which was given for the maximum

output design by equation (7). By substituting (12) into (10), this becomes

$$l = \frac{0.371\lambda}{\sin^2 \Delta} . \tag{13}$$

Therefore, alignment design is obtained by the use of equations (3), (6) and (13). This shows that to change from maximum output design to alignment design it is only necessary to reduce the length to approximately seventy-four per cent of the value recommended for maximum output.

Figures 5 and 6 are examples of the directive patterns secured by such

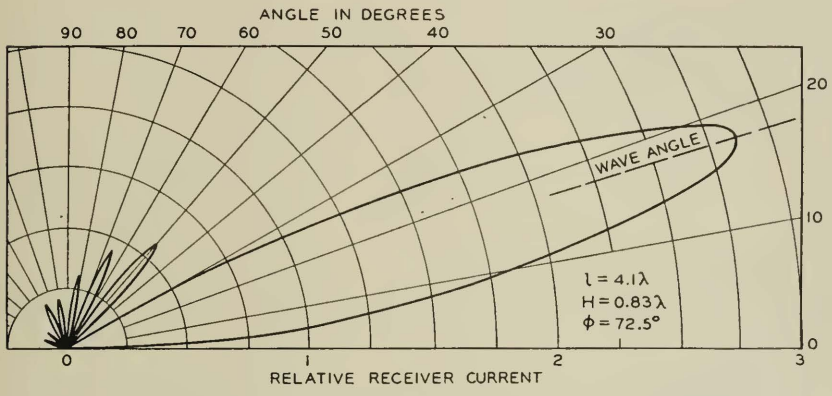


Fig. 5—Alignment design. Receiver current diagram of incident plane directivity for a 17.5° wave-angle.

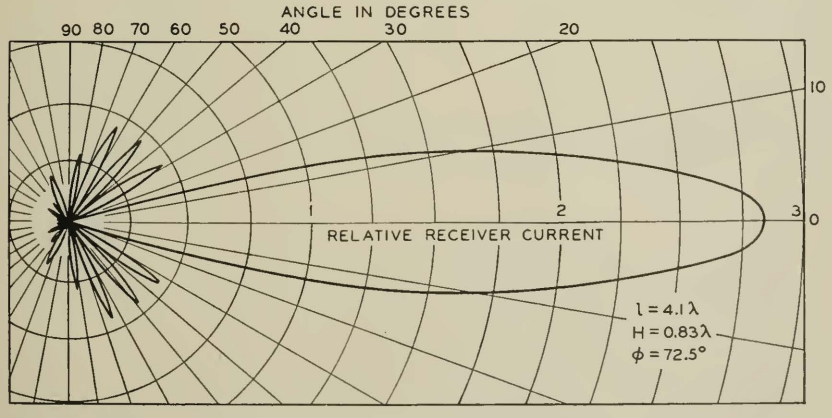


Fig. 6—Alignment design. Receiver current versus azimuth angle for a 17.5° wave-angle.

alignment design for a wave-angle of 17.5 degrees. Comparison of these diagrams with Figs. 3 and 4 shows a small loss in I_R resulting from such design. Figures 7 and 8 are similar alignment design plots for a wave-angle of 11 degrees.

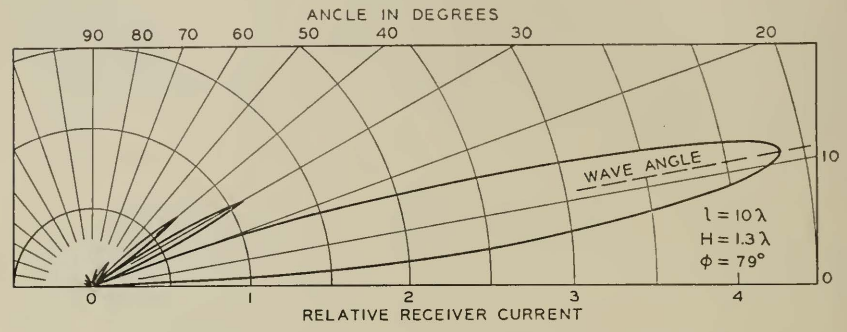


Fig. 7—Alignment design. Receiver current diagram of incident plane directivity for an 11° wave-angle.

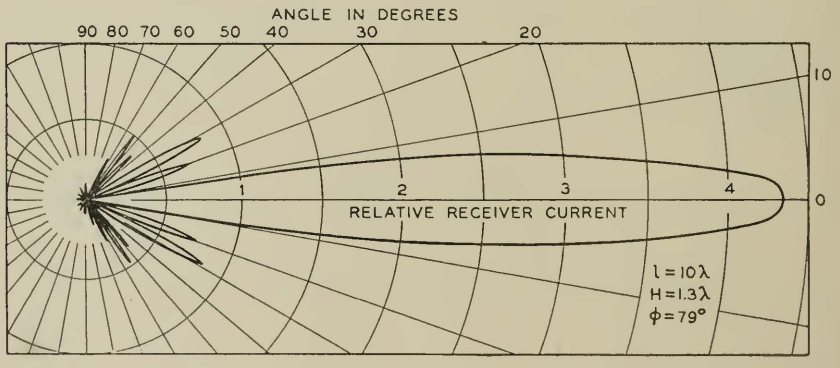


Fig. 8—Alignment design. Receiver current versus azimuth angle for an 11° wave-angle.

Compromise Design Methods

When an antenna output sacrifice must be tolerated to obtain smaller dimensions, it is often desirable to design the system for directive pattern alignment with the wave-angle, as mentioned above. The methods of accomplishing this result will now be given.

The effect on the directive pattern of reducing the height can be largely compensated for by increasing the length.

Substituting (6) into (8) gives

$$\frac{H}{\tan\left(\frac{2\pi H}{\lambda}\sin\Delta\right)} = \frac{\lambda}{2\pi\sin\Delta} - \frac{l\cdot\sin\Delta}{\tan\left(\frac{\pi l}{\lambda}\sin^2\Delta\right)}. \qquad (14)$$

When both Δ and H are specified, this equation gives the element length l to be used with the reduced height. Figures 9 and 10 are

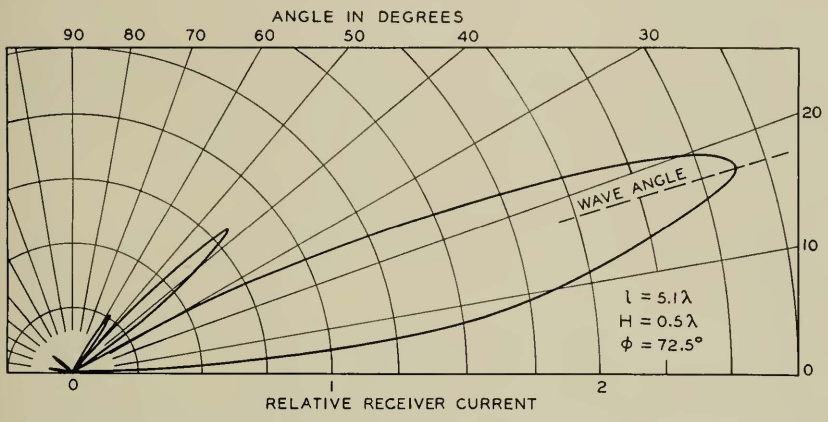


Fig. 9—Compromise height design. Receiver current diagram of incident plane directivity for a 17.5° wave-angle and a height of one-half wave-length.

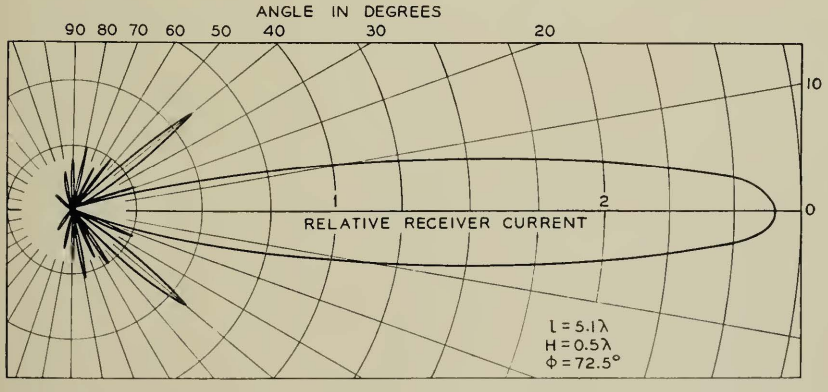


Fig. 10—Compromise height design. Receiver current versus azimuth angle for a 17.5° wave-angle and a height of one-half wave-length.

examples of the directive diagrams resulting from a reduced height of one-half wave-length and an incident wave-angle of 17.5 degrees. These last mentioned figures may be directly compared with Figs. 5

and 6 to determine what effect the sacrifice in height and the increased length has had.

The effect on the directive pattern of reducing the length can be minimized by changing the tilt-angle ϕ. When the height given by equation (3) is used, the new tilt-angle can be obtained by writing equation (10) as

$$\sin \phi = \frac{l - 0.371\lambda}{l \cos \Delta} . \tag{15}$$

If it is desired to align the directive pattern with the wave-angle when the height and length are both reduced from the recommended values, equation (8) can be solved for the new tilt-angle. Any of these compromise designs will, of course, reduce the antenna output.

Limitations of Results

The accuracy of the calculations in this paper is based on the following assumptions:

1. Negligible effect by wire attenuation.
2. Negligible mutual coupling between elements.
3. The rate of change of the effective voltage with varying dimensions is large as compared with the rate of change of radiation resistance.

The validity of the assumptions has been checked to engineering accuracies in the ultra-short wave experiments to be described, where the shape of the directivity predicted by these equations agreed well with experiment. However, theory and experiment have both shown that ordinary ground has an appreciable influence on the wire attenuation when the antenna heights are of the order of only small fractions of a wave-length. This effect increases as the antenna dimensions are extended. In such unusual cases, this behavior should be carefully investigated even though calculations including large wire attenuations have shown surprisingly little effect upon the directive diagrams. Approximate values for the variation of radiation resistance with dimensions, for terminated rhombic antennas, indicate that assumption 3 is probably warranted provided that the antenna is not extraordinarily small.

Finally, attention should be called to these important facts:

The suggested dimensions in this paper not only do not restrict the broad frequency range of this type of antenna but are in many cases

the broadest arrangements under the prescribed conditions. This statement is based on a large number of calculated directive diagrams, for various frequencies, and has been substantiated by operating tests.

Also, in certain applications, as with ultra-short waves where the direction of wave propagation is substantially horizontal ($\Delta = 0$), the derived equations place no restrictions on the antenna size.

Experimental Checks of Directivity Calculations

One method employed for experimentally checking the calculations of directivity was by means of a model small enough to permit continuous rotation of the antenna in respect to a fixed wave-direction. A simple transmitter and a double-detection receiving set, which employed a calibrated intermediate frequency attenuator, both operating at about four meters, were utilized for this work. Even at these wave-lengths, the physical dimensions of the antenna were rather large so that it was necessary to build an antenna smaller in wave-length dimensions than is usually recommended for commercial installations. As previously stated, any antenna for regular service would be designed for its own specific conditions, so that this antenna is not to be taken as a recommended design. However, this fact had no bearing on the results of these tests, for there was no intention of using this movable antenna for any purpose other than experimentally checking the directivity equations.

A photograph of the test installation is shown in Fig. 11. The antenna wires were mounted on a lattice work cross of light weight wood suspended by a system of ropes from two wooden poles. This permitted the hoisting of the entire antenna system to any height up to several wave-lengths, and rotating it to any desired angular position. The antenna is shown in more detail in the photograph of Fig. 12. It was constructed of two space-tapered wires in parallel, spaced in order to lower the antenna impedance to permit the matching of the open-wire transmission line which was attached to one end. Each element was three and a quarter wave-lengths long. The receiving set was placed in a small push-cart at the other end of the transmission line so that the line length did not need to be changed with antenna movements. An antenna of this size had a measured gain of fourteen decibels over a half-wave horizontal dipole at the same height.

The horizontal plane directive pattern was experimentally determined by measuring the antenna angle and signal output at each maximum and minimum and at fixed angles on the major lobe, as the antenna was rotated about its vertical axis. The pattern measured

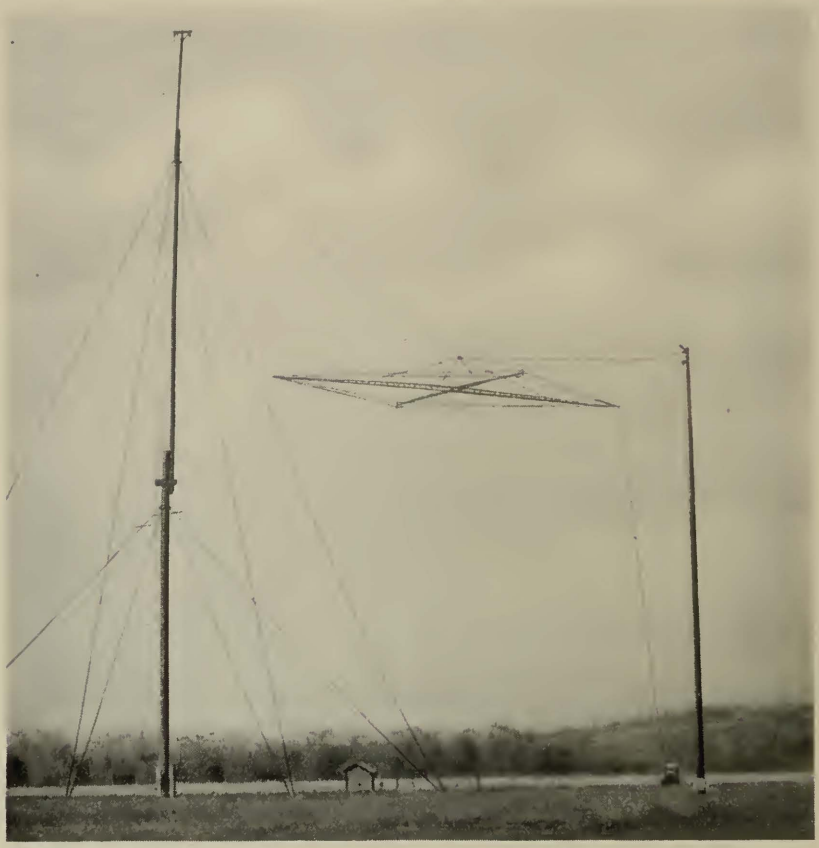


Fig. 11—Test antenna used at four meters for experimental checks of directivity calculations.

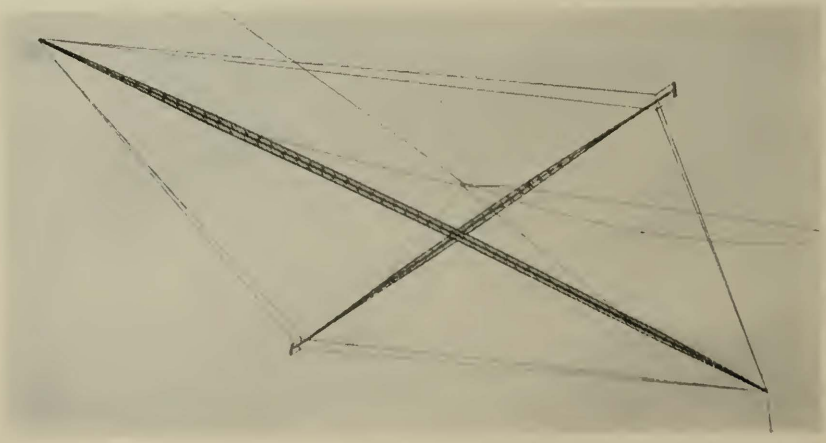


Fig. 12—A detail view of the antenna shown in Fig. 11.

in this manner is shown in Fig. 13. In this figure is also shown, for comparison purposes, the pattern calculated for these conditions by the equations derived in the appendix. The agreement between these two patterns is quite evident.

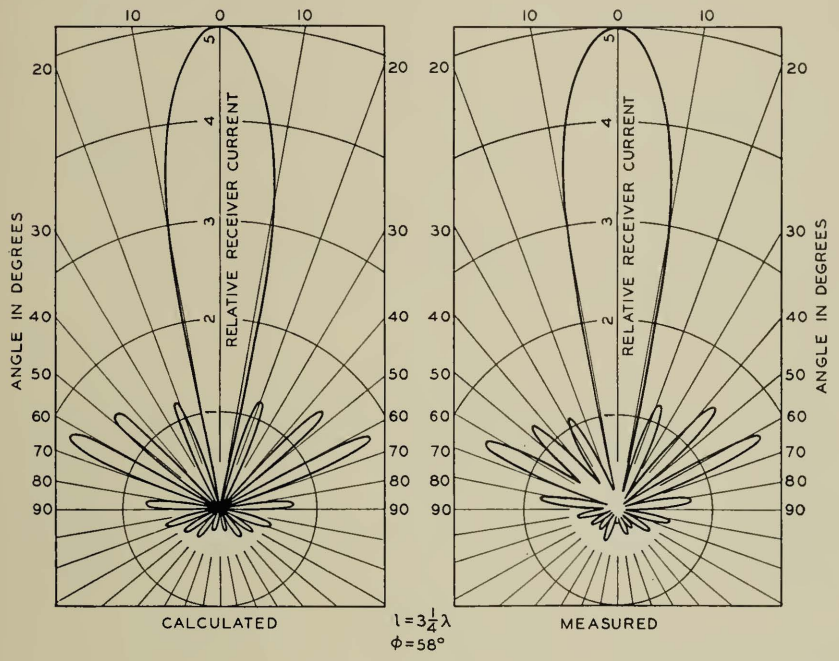


Fig. 13—Receiver current diagrams of test antenna output current versus azimuth angle for a 2.7° wave-angle.

To check the vertical directivity of the system, a portable oscillator was raised from the ground to the top of a one-hundred foot pole in front of the antenna. Due to the limited range of arrival angles so obtained, only the major lobe of the directive pattern could be traced out. Figure 14 gives the calculated major lobe for these conditions (equations corrected to actual path length difference between direct and reflected waves) with the antenna one-half wave-length above ground, and Fig. 15 is the pattern at a height of one wave-length. The circles on these curves are the measured values. These patterns show that the directivity does not change very rapidly with changes in antenna height.

Experimental checks of these directivity calculations have also been obtained on the larger antennas ordinarily used for long distance reception. A large horizontal rhombic antenna was constructed with a

motor driven winch to adjust its interior angles, thus providing control of the directivity of the antenna. This antenna could be "steered" so that either the major lobe or the first null could be used to receive or

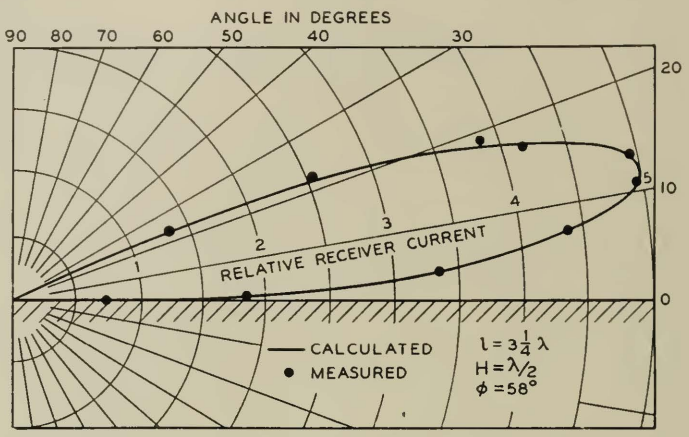


Fig. 14—Receiver current diagram of incident plane directivity with test antenna one-half wave-length above ground. Major lobe only.

discriminate against the components of the received signal. The angle of arrival of these components was measured by the methods described

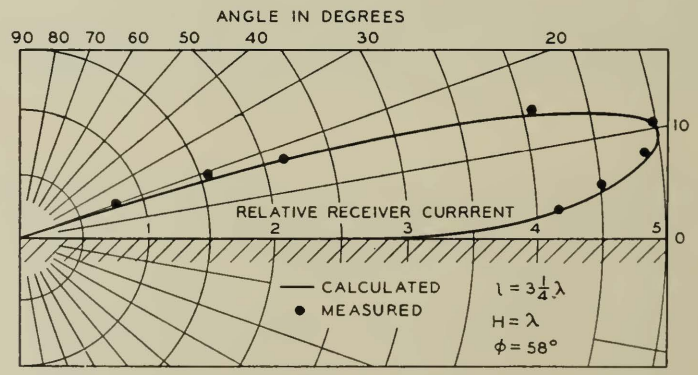


Fig. 15—Receiver current diagram of incident plane directivity with test antenna one wave-length above ground. Major lobe only.

by Friis, Feldman and Sharpless.[2] The agreements were at all times close. This work was incidental to some fading studies the results of which also checked qualitatively the directivity calculations.

These substantial agreements between calculated and measured values are an indication of the accuracy secured in theoretically predicting antenna directivity when using the stated assumptions. It is believed, therefore, that they establish the validity of the design methods now being used for commercial applications of this type of antenna.

Standing Wave Suppression

To obtain the uni-directional characteristics usually desired and to provide a constant terminal impedance over a broad frequency range, the end of the rhombic antenna, remote from the receiver, is terminated in its equivalent characteristic impedance to prevent wave reflections.

The termination was originally adjusted by determining the greatest front-to-back signal reception ratios which could be obtained with the terminating graphite resistors or the other circuits employed. This method of testing proved both cumbersome and erratic for the following reasons:

Radiating field oscillators had to be placed at sizable distances both in front and in the rear of the antenna. Manipulating these oscillators, for frequency runs, proved inconvenient. Furthermore, properly elevating the oscillators was impractical because of the large heights required to simulate the actual average wave-angles usually encountered. When the oscillators were at more reasonable heights above ground, the front-to-back signal ratios could rarely be reproduced since the wave-direction was unstably situated on the lower steep edge of the incident plane directive diagram.

The procedure for testing the termination has been simplified through the measurement, by a substitution method, of the antenna terminal impedance as the frequency is varied over the required frequency band. By readjusting the termination, until these measured values are constant, the required value is obtained.

The schematic diagram of the impedance measuring equipment employed is shown in Fig. 16. In the actual apparatus, the individual

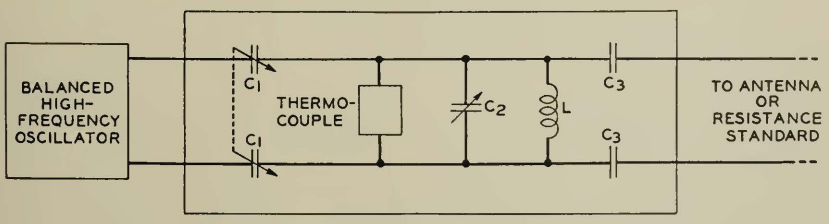


Fig. 16—Schematic diagram of impedance measuring equipment.

parts are separately shielded and symmetrically arranged so that the assembly can be attached to a portable balanced oscillator used for other work. Such a combination is shown in the photograph of Fig. 17.

Again referring to Fig. 16, the condensers marked C_1 control the degree of coupling while maintaining an accurate balance to ground.

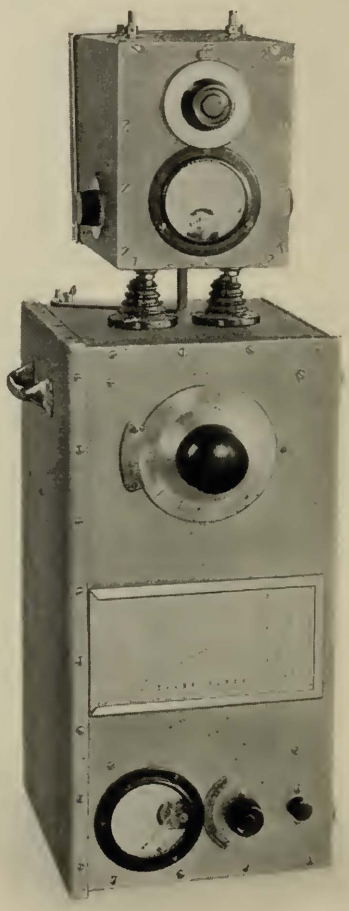


Fig. 17—Impedance measuring equipment and balanced oscillator.

C_2 is a condenser used for anti-resonating the complete circuit, this anti-resonance being indicated by a thermocouple deflection. The condensers labelled C_3 are large blocking condensers which have been added to the circuit to permit the measurement of the value of the variable substitution resistance without the necessity of disconnecting that resistance. A pencil lead is usually employed as this resistor.

The measurement consists in replacing the antenna by a value of resistance R which gives the same thermocouple deflection. The reactive components involved are tuned out in each case by adjusting the condenser C_2. The change in the dial setting of this condenser, as the resistance is substituted for the antenna, permits the determination of the reactive component of the antenna impedance.

Curve 3, in Fig. 18, shows the terminal resistance versus frequency of a properly terminated experimental antenna constructed by the authors. Curve 1, of that figure, results from a termination resistance which is too large while that in curve 2 is too small. Incidentally, it is interesting to note that the value of the measured terminal resistance of this antenna is lower than the resistance required, at the other end of the antenna, for the correct termination. This behavior is probably an effect of the radiation resistance of the system.

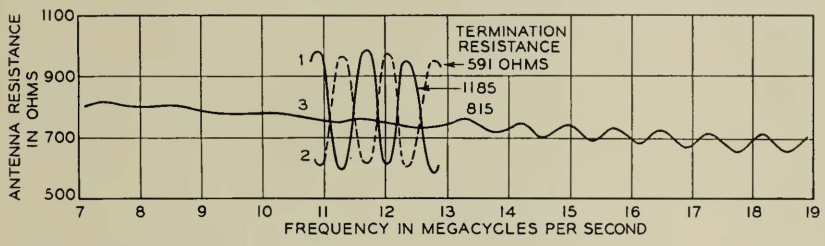


Fig. 18—Rhombic antenna resistance versus frequency for three values of termination resistance.

When a lumped termination resistance is employed, measurements reveal that it operates as if paralleled by a small effective capacitance, a capacitance probably due to the proximity of the adjacent converging antenna wires. This capacitance prevents the resistance from acting as an entirely satisfactory termination. Its effect can be minimized through a proper distribution of the termination resistance over a short length of the converging part of the antenna. This distribution tends to isolate the harmful capacitances by means of the resistances and also introduces a small effective series inductance which is helpful in neutralizing the residual capacitive reactance. Other circuits have been devised, but this one is often employed because of its simplicity and ruggedness.

Antenna-to-Pipe Line Coupling Circuits

Concentric pipe transmission lines are usually favored for receiving antenna installations as they can be buried in the ground to give a substantial and weather-proof construction, free from reactions on adjacent antennas. This pipe-line runs up one pole of a rhombic

antenna and terminates adjacent to the output terminals of the antenna.

Large terminated rhombic antennas, built with number 12 A.W.G. wire, present a terminal impedance, balanced to ground, of nearly 800 ohms resistance. Where the ratio of the inner diameter of the outer pipe to the outer diameter of inner pipe of the transmission line is the optimum value [5] of 3.6, the characteristic impedance of the line is about 77 ohms. The problem is, therefore, to transform properly a balanced resistance of about 800 ohms into an unbalanced resistance of about 80 ohms without appreciable losses over the entire operating range of the antenna.

Figure 19 is a photograph of a simple type of coupling circuit developed for experimental use. Figure 20 is the measured loss versus frequency curve obtained when the device is connected between the stated impedances. The broad frequency range is primarily the result of a high coefficient of coupling. Balance is obtained by symmetrically disposing two primary windings in series about two secondary windings in parallel. A still greater operating range may be obtained through a more elaborate arrangement. Note in Fig. 19 that the boxes and pipe line are gas tight so that nitrogen pressure can be applied, if desired, to prevent moisture absorption through "breathing."

Conclusion

The original paper [1] on rhombic antennas, in discussing the situation regarding short-wave communication, stated that the main limitations in this field were three in number:

(*a*) Inherent receiver noise.
(*b*) External noise (static, man-made noises, etc.)
(*c*) Signal fading.

That paper attempted to show that the design of the receiving antenna system has an important bearing upon overcoming all three of these difficulties.

In the present paper, the authors have indicated how progress in overcoming the first two limitations, through refinements in the horizontal rhombic type of antenna, has been made.

Improvements in the selection of the receiving antenna dimensions have greatly helped the noise situation. Termination and coupling circuit developments now allow many of these antennas to operate one circuit or several circuits simultaneously within a four-to-one frequency range.

[5] E. J. Sterba and C. B. Feldman, "Transmission Lines for Short-Wave Radio Systems," *Proc. I. R. E.*, July, 1932; *Bell Sys. Tech. Jour.*, July, 1932.

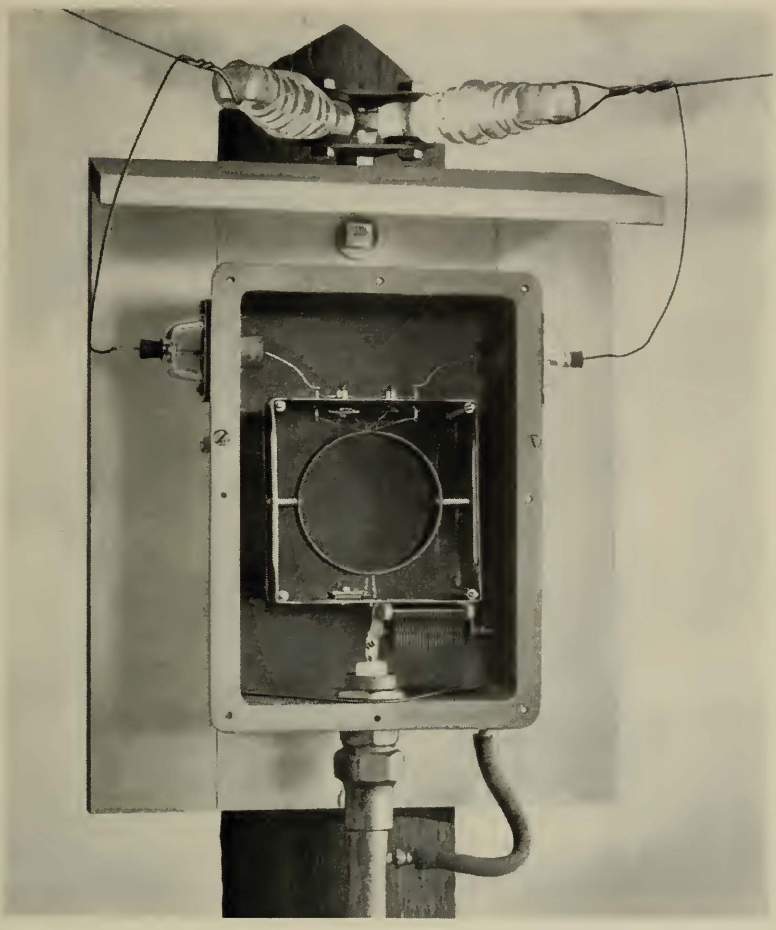


Fig. 19—Experimental equipment for coupling a rhombic antenna to a concentric pipe transmission line.

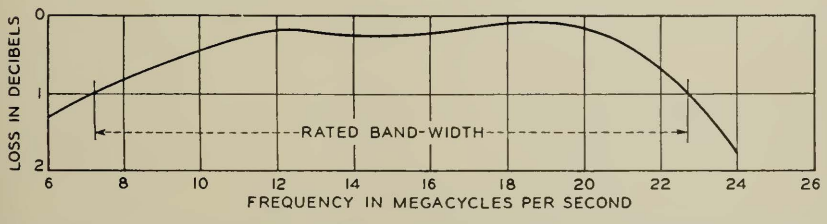


Fig. 20—Loss versus frequency of the experimental coupling equipment, shown in Fig. 19, when used to connect an antenna balanced impedance of 778 ohms in parallel with 5 mmf. to a pipe-line unbalanced resistance of 80 ohms.

APPENDIX

Calculations of Three-Dimensional Directivity

Antenna: Horizontal Rhombus.
Wave Polarization: Horizontal.
Far End Termination: Characteristic Impedance.
Ground Constants: General Case.

Assumptions: Negligible Wire Attenuation and Leakage.
Negligible Mutual Coupling.
Uniform Characteristic Impedance.

Nomenclature: (See Fig. 1)

P = point under consideration.

l = element length.

ϕ = one-half of side apex angle.

Δ = angle made with ground by wave-direction in incident plane.

β = angle made by horizontal component of wave-direction with principal axis drawn through receiver and termination.

H = height above ground.

X = wire distance of point P from R for element (2) as well as for element (1).

ϵ = r.m.s. space voltage.

E = r.m.s. space voltage and phase at P due to ϵ, direct propagation.

E' = r.m.s. space voltage and phase at P due to ϵ, after ground reflection.

E_{P1} and E_{P2} = wire voltage at P in element (1) or element (2), direct propagation.

E_{P1}' and E_{P2}' = same as above but for ground reflection.

a = amplitude ratio of ground reflection to direct space voltage.

α = apparent phase angle change after ground reflection.

λ = wave-length.

I_R = receiver current.

Z_0 = characteristic impedance.

R = receiver impedance.

T = far end termination impedance.

Voltages Induced in Wire:

For a directly propagated wave in element (1), R being the reference point for phases,

$$E_{P1} = \epsilon \cdot \cos(\phi - \beta) \cdot e^{+j\frac{2\pi}{\lambda} \cdot \sin(\phi - \beta) \cdot \cos\Delta \cdot X} dX.$$

For reflected wave in element (1) ($X = 0$ to l)

$$E_{P1}' = a \cdot \epsilon \cdot \cos(\phi - \beta) \cdot e^{+j\left\{\frac{2\pi}{\lambda}[X \cdot \sin(\phi - \beta) \cdot \cos\Delta - 2H\sin\Delta] - \alpha\right\}} dX.$$

Similarly for element (3) ($X = 0$ to l)

$$E_{P3} = \epsilon \cdot \cos(\phi + \beta) \cdot e^{+j\frac{2\pi}{\lambda} \cdot \sin(\phi + \beta) \cdot \cos\Delta \cdot X} dX.$$

$$E_{P3}' = a \cdot \epsilon \cdot \cos(\phi + \beta) \cdot e^{+j\left\{\frac{2\pi}{\lambda}[X \cdot \sin(\phi + \beta) \cdot \cos\Delta - 2H\sin\Delta] - \alpha\right\}} dX.$$

Also for elements (2) and (4) ($X = 0$ to l) where respective side apexes are now reference points for phases,

$$E_{P2} = -E_{P3} \cdot e^{+j\frac{2\pi l}{\lambda} \cdot \sin(\phi - \beta) \cdot \cos\Delta},$$

$$E_{P2}' = -E_{P3}' \cdot e^{+j\frac{2\pi l}{\lambda} \cdot \sin(\phi - \beta) \cdot \cos\Delta},$$

and

$$E_{P4} = -E_{P1} \cdot e^{+j\frac{2\pi l}{\lambda} \cdot \sin(\phi + \beta) \cdot \cos\Delta},$$

$$E_{P4}' = -E_{P1}' \cdot e^{+j\frac{2\pi l}{\lambda} \cdot \sin(\phi + \beta) \cdot \cos\Delta}.$$

Receiver Current where $T = Z_0$ and $R = Z_0$

$$I_R = \int_0^l \frac{E_{P1} + E_{P1}' + E_{P3} + E_{P3}'}{2Z_0} \cdot e^{-j\frac{2\pi}{\lambda}X} + \int_0^l \frac{E_{P2} + E_{P2}' + E_{P4} + E_{P4}'}{2Z_0} \cdot e^{-j\frac{2\pi}{\lambda}X} \cdot e^{-j\frac{2\pi}{\lambda}l},$$

$$I_R = \int_0^l \frac{\epsilon \cdot \cos(\phi - \beta)}{2Z_0} \cdot \left[1 + ae^{-j\left(\frac{4\pi}{\lambda}H\sin\Delta + \alpha\right)}\right] \cdot e^{-j\frac{2\pi}{\lambda}[1 - \sin(\phi - \beta) \cdot \cos\Delta]X} dX + \int_0^l \frac{\epsilon \cdot \cos(\phi + \beta)}{2Z_0} \cdot \left[1 + ae^{-j\left(\frac{4\pi}{\lambda}H\sin\Delta + \alpha\right)}\right] \cdot e^{-j\frac{2\pi}{\lambda}[1 - \sin(\phi + \beta) \cdot \cos\Delta]X} dX - \int_0^l \frac{\epsilon \cdot \cos(\phi - \beta)}{2Z_0} \cdot \left[1 + ae^{-j\left(\frac{4\pi}{\lambda}H\sin\Delta + \alpha\right)}\right] \cdot e^{-j\frac{2\pi l}{\lambda}[1 - \sin(\phi + \beta)\cos\Delta]} \cdot e^{-j\frac{2\pi}{\lambda}[1 - \sin(\phi - \beta)\cos\Delta]X} dX$$

$$- \int_0^l \frac{\epsilon \cdot \cos(\phi + \beta)}{2Z_0} \cdot \left[1 + ae^{-j\left(\frac{4\pi}{\lambda} H \sin\Delta + \alpha\right)}\right] \cdot e^{-j\frac{2\pi l}{\lambda}[1 - \sin(\phi - \beta)\cos\Delta]} \cdot e^{-j\frac{2\pi}{\lambda}[1 - \sin(\phi + \beta)\cos\Delta]X} dX,$$

$$I_R = \frac{\epsilon \cdot \cos(\phi - \beta)}{2Z_0} \cdot \left[1 + ae^{-j\left(\frac{4\pi}{\lambda} H \sin\Delta + \alpha\right)}\right] \cdot \left[1 - e^{-j\frac{2\pi l}{\lambda}[1 - \sin(\phi + \beta)\cdot\cos\Delta]}\right] \cdot \int_0^l e^{-j\frac{2\pi}{\lambda}[1 - \sin(\phi - \beta)\cdot\cos\Delta]X} dX$$

$$+ \frac{\epsilon \cdot \cos(\phi + \beta)}{2Z_0} \cdot \left[1 + ae^{-j\left(\frac{4\pi}{\lambda} H \sin\Delta + \alpha\right)}\right] \cdot \left[1 - e^{-j\frac{2\pi l}{\lambda}[1 - \sin(\phi - \beta)\cdot\cos\Delta]}\right] \cdot \int_0^l e^{-j\frac{2\pi}{\lambda}[1 - \sin(\phi + \beta)\cdot\cos\Delta]X} dX,$$

$$I_R = +j\frac{\epsilon\lambda}{4\pi Z_0} \cdot \frac{\cos(\phi - \beta)}{1 - \sin(\phi - \beta)\cdot\cos\Delta} \cdot \left[1 + ae^{-j\left(\frac{4\pi}{\lambda} H \sin\Delta + \alpha\right)}\right] \cdot \left[1 - e^{-j\frac{2\pi l}{\lambda}[1 - \sin(\phi + \beta)\cdot\cos\Delta]}\right]\left[1 - e^{-j\frac{2\pi l}{\lambda}[1 - \sin(\phi - \beta)\cdot\cos\Delta]}\right]$$

$$+ j\frac{\epsilon\lambda}{4\pi Z_0} \cdot \frac{\cos(\phi + \beta)}{1 - \sin(\phi + \beta)\cdot\cos\Delta} \cdot \left[1 + ae^{-j\left(\frac{4\pi}{\lambda} H \sin\Delta + \alpha\right)}\right] \cdot \left[1 - e^{-j\frac{2\pi l}{\lambda}[1 - \sin(\phi - \beta)\cdot\cos\Delta]}\right]\left[1 - e^{-j\frac{2\pi l}{\lambda}[1 - \sin(\phi + \beta)\cdot\cos\Delta]}\right],$$

$$I_R = j\frac{\epsilon\lambda}{4\pi Z_0} \cdot \left[\frac{\cos(\phi - \beta)}{1 - \sin(\phi - \beta)\cdot\cos\Delta} + \frac{\cos(\phi + \beta)}{1 - \sin(\phi + \beta)\cdot\cos\Delta}\right] \cdot \left[1 + ae^{-j\left(\frac{4\pi}{\lambda} H \sin\Delta + \alpha\right)}\right] \cdot \left[1 - e^{-j\frac{2\pi l}{\lambda}[1 - \sin(\phi + \beta)\cdot\cos\Delta]}\right] \cdot \left[1 - e^{-j\frac{2\pi l}{\lambda}[1 - \sin(\phi - \beta)\cdot\cos\Delta]}\right],$$

which is the final equation including phase relations.

Extraneous Frequencies Generated in Air Carrying Intense Sound Waves *

By A. L. THURAS, R. T. JENKINS and H. T. O'NEIL

The exact equation of the propagation of plane sound waves in air is not linear and consequently harmonics and combination tones are generated. The pressure of these extraneous frequencies in terms of the fundamental pressure, frequency, and distance from the source has been mathematically determined by Rayleigh, Lamb and others. These equations have been applied to an exponential horn.

Measurements of the second harmonic and combination tones have been made at various points within a long tube, and in front of an exponential horn. Measurements, in general, agree with theory, but the absolute values are lower than the calculated values.

RECENT developments in horn type loud speakers [1] for high quality reproduction of intense sounds necessitate a consideration of the more exact equations of wave motion if distortion due to the generation of extraneous frequencies in the air of the horn itself is to be avoided. Similar considerations may be of some importance in connection with the pick-up of intense sounds.

The propagation of waves of finite displacement has interested physicists for more than a century. In 1808 Poisson derived an equation which shows that, in general, a sound wave cannot be propagated without a change in form and consequent generation of additional frequencies. This distortion is caused by the non-linearity of air; that is, if equal positive and negative increments of pressure are impressed on a mass of air the changes in volume of the mass will not be equal; the volume change for the positive pressure will be less than the volume change for the equal negative pressure. An idea of the nature of the distortion can be obtained from the adiabatic curve AB for air as given in the familiar volume pressure indicator diagram (Fig. 1a). The undisturbed pressure and specific volume of air are indicated by point P_0V_0. Any deviation from the tangent through this point causes distortion and consequent generation of extraneous frequencies. The theoretical magnitudes of the waves of extraneous frequencies are obtained from a solution of the exact differential equation of wave propagation in air. The solution shows that the pressure of the second harmonic frequency, which is generated in the air, increases with the frequency and the magnitude of the fundamental

* Published in the January 1935 issue of the *Jour. Acous. Soc. Am.*

[1] E. C. Wente and A. L. Thuras, "Loud Speakers and Microphones," *Bell Sys. Tech. Jour.*, 13, 259 (1934).

pressure and also with the distance from the sound source. The solution also gives the magnitudes of the waves of sum and difference frequencies generated when two tones are simultaneously impressed

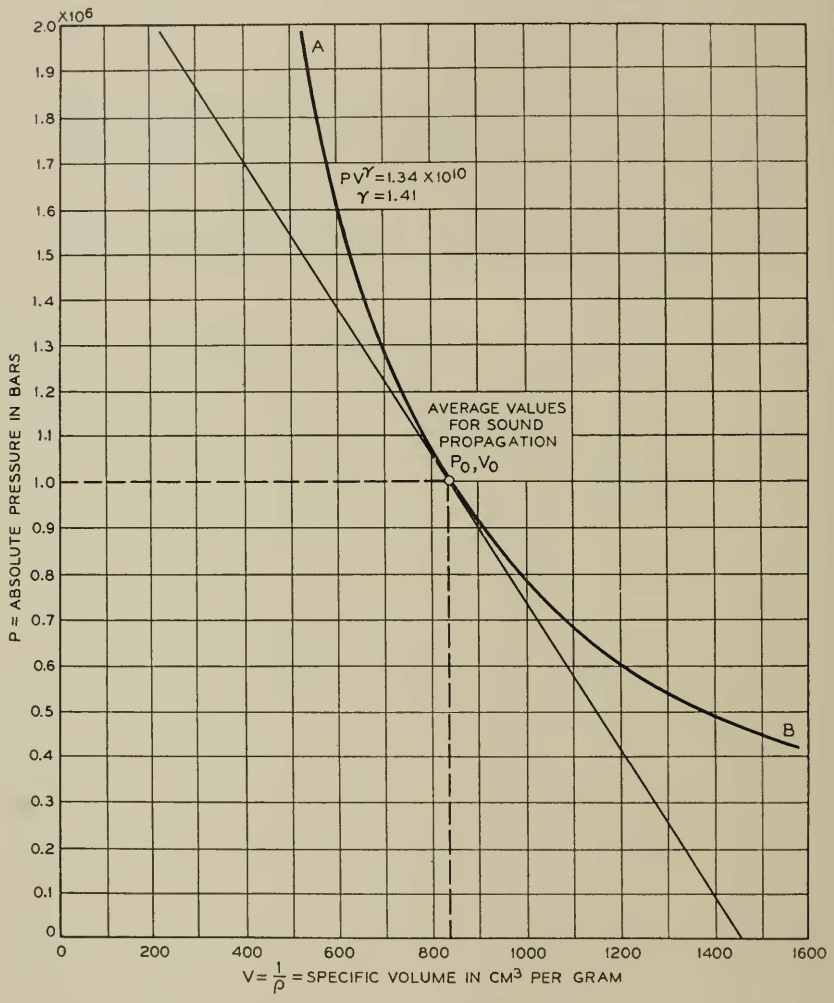


Fig. 1a—Adiabatic curve for air.

on the air; these magnitudes also increase with distance from the source and with the product of the fundamental pressures and, respectively, with the sum and difference frequencies.

Theory of Propagation of Plane Waves of Finite Amplitude

The derivation of the exact differential equation for sound wave propagation in air involves the continuity equation, Newton's force equation and the equation expressing the relation between pressure and specific volume in a gas. Since there may be some question as to the accurate definition of the density and force in the equation of motion a somewhat detailed discussion of this subject will be given.

Following Rayleigh,[2] let y and $y + (\partial y/\partial x)dx$ be the actual distances at time t from the plane $x = 0$ to neighboring layers of air whose undisturbed positions are defined by x and $x + dx$, respectively, Fig. 1b.

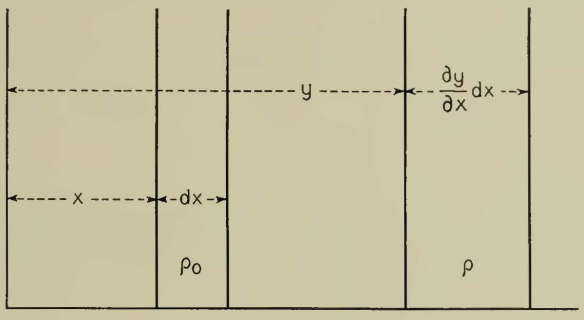


Fig. 1b.

The displacement corresponding to y is thus $\xi = y - x$ and the equation of continuity of the fluid is

$$\rho = \rho_0(\partial y/\partial x)^{-1} = \rho_0(1 + \partial\xi/\partial x)^{-1}, \tag{1}$$

where ρ and ρ_0 are the densities of the fluid in the disturbed and undisturbed states, respectively. If the effect of viscosity is neglected the exact equation of motion of the element of mass $\rho(\partial y/\partial x)\cdot dx$ is

$$\frac{\partial^2 y}{\partial t^2}\cdot\rho\frac{\partial y}{\partial x}dx = \frac{\partial^2 y}{\partial t^2}\cdot\rho_0 dx = -\frac{\partial p}{\partial y}\cdot\frac{\partial y}{\partial x}dx,$$

or

$$\rho_0(\partial^2\xi/\partial t^2) = -\partial p/\partial x, \tag{2}$$

p is the pressure at the point y (Fig. 1b) which moves with the air particle, not the pressure at a fixed point. Except for very large displacements these pressures are nearly the same. From equations (1) and (2)

$$\partial^2\xi/\partial t^2 = (dp/d\rho)\cdot(1 + \partial\xi/\partial x)^{-2}\cdot(\partial^2\xi/\partial x^2). \tag{3}$$

[2] Lord Rayleigh, "Theory of Sound," 2nd Ed., Vol. II, p. 31.

By virtue of equation (1), equation (3) is linear in ξ only if $dp/d\rho = K\rho^{-2}$ or $dp/dv = -K$, where $v = 1/\rho$ = specific volume and K is a constant. This condition is not satisfied during any ordinary variations of state of a gas, but is approximately satisfied when the variations are very small. For isothermal changes we have $pv = p_0v_0$ and for adiabatic changes:

$$p/p_0 = (v_0/v)^\gamma = (\rho/\rho_0)^\gamma, \tag{4}$$

where γ is the ratio of the specific heats and p_0 is the undisturbed atmospheric pressure. In either case, for very small variations, the pv curve is practically identical with the tangent to the curve, hence dp/dv is practically constant (Fig. 1a).

From equations (1), (3), (4) we obtain the exact equation of adiabatic plane wave motion in a non-viscous fluid:

$$\partial^2\xi/\partial t^2 = c^2(1 + \partial\xi/\partial x)^{-\gamma-1}(\partial^2\xi/\partial x^2), \tag{5}$$

where $c^2 = \gamma p_0/\rho_0$. This equation is given by Rayleigh.[2, 3] Rocard [4] was first to call attention to the generation of harmonics in the air within an exponential horn. His theoretical solution is based on a plane wave equation in which the term $\partial^2\xi/\partial t^2$ was replaced by

$$\frac{\partial^2\xi}{\partial t^2} + \frac{\partial\xi}{\partial t} \cdot \frac{\partial}{\partial x}\left(\frac{\partial\xi}{\partial t}\right). \tag{6}$$

In support of this substitution Rocard cites Riemann's [5] treatment of the problem. However, Riemann's analysis is based on the Eulerian form of the hydrodynamical equations whereas equation (5) is derived from the Lagrangian equations. (For a comparison of these systems of equations see Lamb.[6]) In the Lagrangian notation $\partial\xi/\partial t$ and $\partial^2\xi/\partial t^2$ are the *exact* values of the velocity and acceleration, respectively, of the particle whose displacement from its equilibrium position (x) is ξ. It is to be noted that in equation (2) the term ρ_0, or undisturbed density, does not represent an approximation.

A rigorous solution of (5) for the displacement ξ as an explicit function of x and t has not been obtained. As a first approximation to equation (5) we take

$$\frac{\partial^2\xi}{\partial t^2} = c^2\frac{\partial^2\xi}{\partial x^2} - (\gamma + 1)c^2\frac{\partial\xi}{\partial x} \cdot \frac{\partial^2\xi}{\partial x^2}. \tag{7}$$

[3] Lamb, "Dynamical Theory of Sound," 2nd Ed., p. 182.

[4] Y. Rocard, "Sur la propagation des ondes sonores d'amplitude finie," *Comptes rendus*, 196, 161 (1933).

[5] Riemann, "Ueber die Fortpflanzung ebener Luftwellen von endlicher Schwingungsweitte," *Göttingen Abhandlungen*, No. 8, 1860.

[6] Lamb, "Hydrodynamics," 6th Ed., Chapter I.

This approximation restricts the dilatation $\partial\xi/\partial x$ to values small compared with unity or the excess pressure to values small compared with γp_0, but the restriction need not be as severe as would be required for the linear approximation:

$$\partial^2\xi/\partial t^2 = c^2\partial^2\xi/\partial x^2.$$

By a method of successive approximations, carried to the second approximation, Lamb [3] derives the solution of (7):

$$\xi = a\cos\omega(t - x/c) + \frac{\gamma+1}{8}\frac{\omega^2}{c^2}a^2x[1 - \cos 2\omega(t - x/c)] \quad (8)$$

corresponding to a motion $\xi = a\cos\omega t$ imposed on the air at $x = 0$, and assuming complete absence of reflection. By virtue of (4) and (1) we have for the pressure:

$$p = p_0(1 - \gamma\partial\xi/\partial x + \cdots).$$

Neglecting the terms of small amplitude, in the region where $4\pi x$ is large compared with the wave-length λ we have

$$p = P_{dc} + p_1 + p_2, \quad (9)$$

where

$$P_{dc} = p_0 - \gamma p_0\cdot((\gamma+1)/8)\cdot(\omega^2/c^2)a^2,$$
$$p_1 = -\gamma p_0\cdot(\omega a/c)\sin\omega(t - x/c),$$
$$p_2 = \gamma p_0\cdot((\gamma+1)/4)\cdot(\omega^3/c^3)\cdot a^2x\sin 2\omega(t - x/c),$$

or

$$p_1 = 2^{\frac{1}{2}}P_1\cos[\omega(t - x/c) + \pi/2],$$
$$p_2 = 2^{\frac{1}{2}}P_2\cos[2\omega(t - x/c) - \pi/2], \quad (10)$$

where

$$P_1 = \gamma p_0\omega a/2^{\frac{1}{2}}c, \quad (11)$$

$$P_2 = \frac{\gamma+1}{2(2)^{\frac{1}{2}}}\cdot\frac{P_1^2}{\gamma p_0}\cdot\frac{\omega x}{c}. \quad (12)$$

P_1 and P_2 are thus the r.m.s. fundamental and second harmonic pressures, respectively.

Lamb [3] also gives a solution of (7) for the case when the forced motion at $x = 0$ is: $\xi = \xi_A\cos\omega_A t + \xi_B\cos\omega_B t$. In addition to the two fundamentals and two second harmonics, the pressure now includes components whose frequencies are, respectively, the sum and difference of the two primary frequencies:

$$p_s = 2^{\frac{1}{2}}P_s\cos[(\omega_A + \omega_B)(t - x/c) - \pi/2],$$
$$p_d = 2^{\frac{1}{2}}P_d\cos[(\omega_A - \omega_B)(t - x/c) + \pi/2],$$

where

$$P_s = \frac{\gamma + 1}{2(2)^{\frac{1}{2}}} \cdot \frac{P_A P_B}{\gamma p_0} \cdot \frac{(\omega_A + \omega_B)x}{c}, \tag{13}$$

$$P_d = \frac{\gamma + 1}{2(2)^{\frac{1}{2}}} \cdot \frac{P_A P_B}{\gamma p_0} \cdot \frac{(\omega_A - \omega_B)x}{c}, \tag{14}$$

and P_A, P_B are the two r.m.s. fundamental pressures.

If we extend Lamb's method of solution of equation (7) to the third approximation and again consider the case $\xi = a \cos \omega t$ at $x = 0$ we find that in the region $kx \gg 1$, the r.m.s. third harmonic pressure is

$$P_3 = \frac{3}{16}\left(\frac{\gamma + 1}{\gamma p_0} \cdot \frac{\omega x}{c}\right)^2 P_1^3. \tag{15}$$

In the case of the greatest r.m.s. fundamental pressure used in the experiments, $P_1 = 8000$ bars at 600 c.p.s., equation (15) indicates that the third harmonic at 400 cm from the source is about 10 db below the second harmonic.

An approximate correction for the effect of attenuation in a tube caused by viscosity and heat conduction can be obtained by assuming that each of the extraneous frequencies and the fundamental is attenuated as if it were the only wave present. Thus the r.m.s. value of the fundamental at any point x is assumed to be

$$P_1 = P_0 e^{-\alpha_1 x} \qquad \text{or} \qquad dP_1/dx = -\alpha_1 P_1,$$

where P_0 is the r.m.s. value of the fundamental at the point $x = 0$ and α_1 is the measured attenuation factor for the fundamental. If P_2 is the r.m.s. value of the second harmonic at the point x and α_2 is the measured attenuation factor for the second harmonic in the absence of the fundamental, we have by using equation (12):

$$dP_2/dx = KP_1^2 - \alpha_2 P_2 = KP_0^2 e^{-2\alpha_1 x} - \alpha_2 P_2, \tag{16}$$

where

$$K = \frac{\gamma + 1}{2(2)^{\frac{1}{2}}} \cdot \frac{1}{\gamma p_0} \cdot \frac{\omega}{c}.$$

When $\alpha_1 = 0$ and $\alpha_2 = 0$, equation (16) is equivalent to (12). The solution of (16) which is consistent with the fact that the second harmonic vanishes at $x = 0$ is

$$P_2 = [KP_0^2/(2\alpha_1 - \alpha_2)][e^{-\alpha_2 x} - e^{-2\alpha_1 x}].$$

Hence

$$\frac{P_2}{P_1} = \frac{\gamma + 1}{2(2)^{\frac{1}{2}}} \cdot \frac{P_0}{\gamma p_0} \cdot \frac{\omega}{c} x \cdot R, \tag{17}$$

where

$$R = \frac{e^{-(\alpha_2-\alpha_1)x} - e^{-2\alpha_1 x}}{(2\alpha_1 - \alpha_2)x} = 1 - \frac{\alpha_2 x}{2} + \cdots.$$

For the tube used in these experiments α_2 may be taken as $2\frac{1}{2}\alpha_1$ which gives

$$R = 1 - \alpha_1 x/2\tfrac{1}{2} + \cdots.$$

Similar correction factors were derived for the other extraneous frequencies measured in the experiments.

Measurements of Plane Waves of Finite Amplitude in a Tube

The experimental work consisted of measuring the second harmonic generated along a tube. Measurements were also made of the sum and difference tones when two fundamental frequencies of equal pressure were simultaneously impressed on the air of the tube. For the fundamental pressures and distances used in the experiments, the magnitudes of the other harmonics and higher order sum and difference frequencies were probably small. For high fundamental frequencies and pressures, however, these other tones are important, since they increase more rapidly with frequency and pressure than the second harmonic; for instance, the third harmonic pressure increases as the square of the fundamental frequency and as the third power of the fundamental pressure.

A sinusoidal displacement, uniform over the cross section, was impressed on the air at one end of a long tube. The tube had an inside diameter of 3.8 cm and was 1566 cm long. Measurements were made in the first 705 cm only and the remainder of the tube was used for obtaining a non-reflective termination.

A search transmitter, comprising a small tube of 0.08 cm inside diameter and 7.5 cm long, coupled to a small condenser microphone,[7] was used for the measurements. Attenuation in this search tube was sufficiently high to prevent either overloading the microphone or altering the sound wave propagated within the long tube. The search transmitter was connected to a stage of amplification so operated as to preclude non-linear distortion. This was followed by a band-pass filter which selected the frequency desired in the measurements. The filter was terminated by a measuring circuit consisting of a high-gain amplifier and a vacuum tube voltmeter. A diagram of the arrangement is shown in Fig. 2.

[7] H. C. Harrison and P. B. Flanders, "An Efficient Miniature Condenser Microphone System," *Bell Sys. Tech. Jour.*, **11**, 451 (1932).

To obtain reliable measurements throughout the length of the test tube it was necessary to reduce standing waves to a negligible magnitude. This was accomplished by laying a strip of felt in the last 761 cm of the tube, terminating the end with an acoustic resistance approximately equal to the characteristic impedance of the tube, and carefully sealing up all joints along the tube. The pressure variation in the standing wave was ± 0.3 db, which corresponds to a reflection coefficient of 0.035.

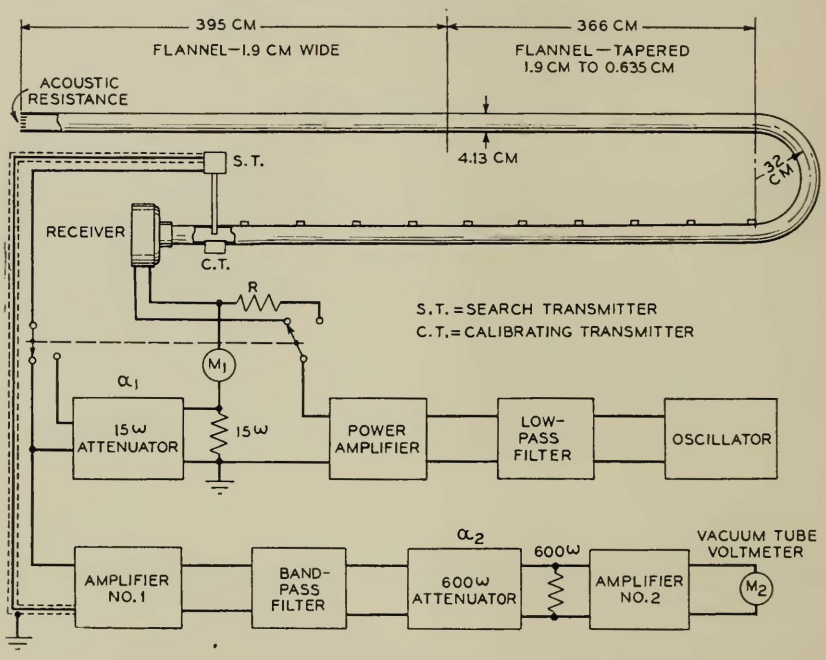


Fig. 2—Apparatus for measuring extraneous frequencies generated in air carrying intense sound waves. *R*, resistance substitute for receiver.

The oscillator current was supplied to the loud speaker through a low-pass filter and the measured harmonic content was found to be 73 db below the fundamental. Pressure measurements in the tube close to the loud speaker indicated that the harmonics generated in the measuring circuit and loud speaker were more than 50 db below the fundamental pressure at 2000 bars.

A calibration of the search transmitter was obtained by comparison with a small condenser microphone whose diaphragm was exposed directly to the sound wave at the same position on the test tube, see Fig. 2. The calibrating microphone (*C.T.*, Fig. 2) had been previously

calibrated by a thermophone.[8] The measuring circuit following the search transmitter was calibrated for each frequency measured by introducing the oscillator current into the search transmitter circuit through the attenuator (α_1, Fig. 2).

The ratio of the pressure of the frequency generated along the tube to the fundamental pressure was measured by the attenuator α_2, Fig. 2, at various holes along the tube in which the search transmitter was inserted. If an appreciable fraction of the harmonic generated along the tube is reflected at the end of the tube, the magnitude of the reflected component near the source may be comparable with or larger than the harmonic generated between the source and the point in question. This was found to be the case when several measurements were made near the source over a distance covering a wave-length. The measured variation in the total pressure of the harmonic over this distance was $\pm$ 2.5 db whereas the variation for a fundamental of this frequency, as previously stated, was $\pm$ 0.3 db. Therefore the measurements close to the receiver may be inaccurate.

Figure 3 shows the measured pressure ratio of the generated second harmonic to the fundamental along the tube and the theoretical curve

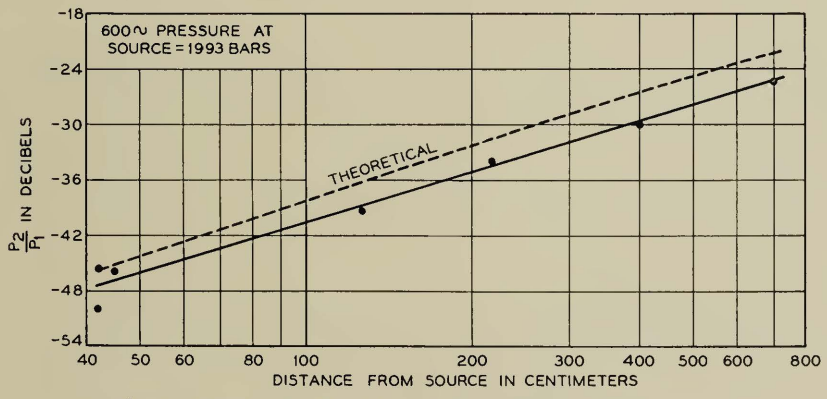


Fig. 3—Magnitude of 2nd harmonic *vs.* distance from source.

calculated from equation (17). Each of the three experimental points plotted at about 45 cm from the receiver is the average pressure ratio for a series of readings taken over a distance of a wave-length along the tube.

The measured and theoretical pressure ratios of the second harmonic to the fundamental are shown as a function of frequency and pressure in Figs. 4 and 5.

[8] L. J. Sivian, "Absolute Calibration of Condenser Transmitter," *Bell Sys. Tech. Jour.*, **10**, Jan. (1931).

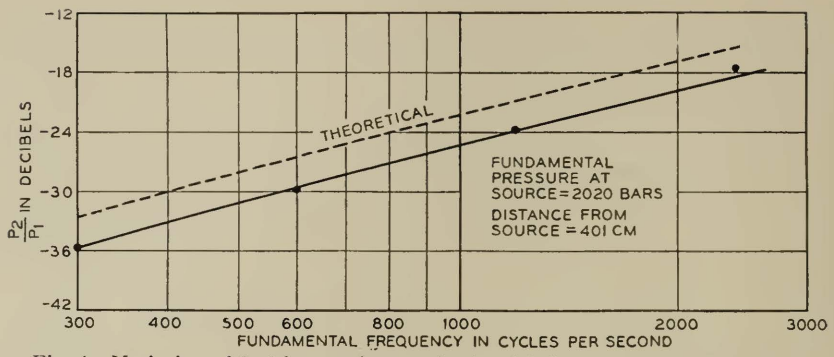


Fig. 4—Variation of 2nd harmonic magnitude with frequency of fundamental.

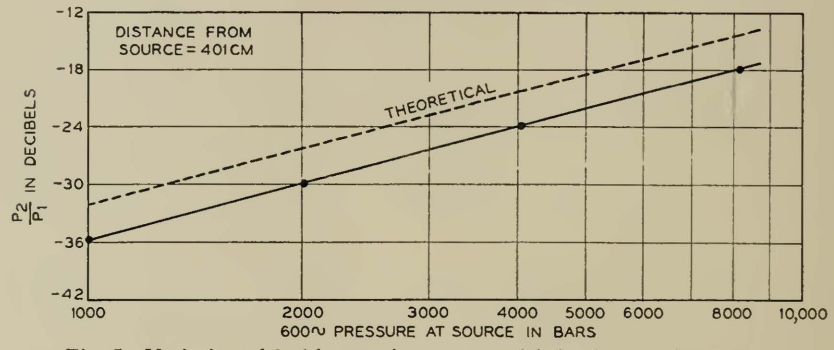


Fig. 5—Variation of 2nd harmonic pressure with fundamental pressure.

Figure 6 shows the magnitude of the first order sum and difference frequencies along the tube when two frequencies of equal pressure are simultaneously impressed on the air in the tube.

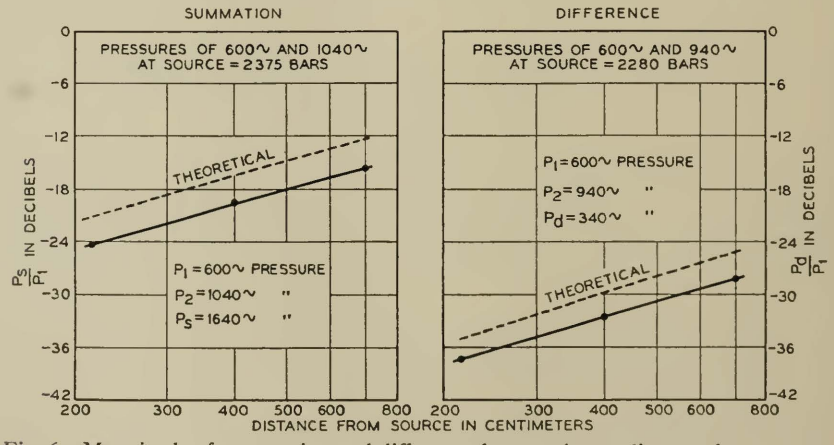


Fig. 6—Magnitude of summation and difference frequencies *vs.* distance from source.

Exponential Horn Theory

The second harmonic generated in any short section of a horn is approximately the same as that generated in a tube of area equal to the mean area of the section of the horn. Therefore, from the tube equation and the expression for the change in pressure due to the divergence of the horn the magnitude of the generated second harmonic pressure at any point along the horn can be obtained.

The r.m.s. value of a small excess pressure in an exponential horn of section $S = S_0 e^{mx}$ is attenuated according to the law

$$P = P_t e^{-mx/2} \qquad \text{or} \qquad dP/dx = -mP/2,$$

where P_t is the r.m.s. pressure in the throat of the horn (at $x = 0$). The index of taper m is equal to $4\pi f_c/c$ where f_c is the cut-off frequency of the horn.

From equation (12), the rate at which the r.m.s. value of the second harmonic increases along a tube is

$$\frac{dP_2}{dx} = \frac{\gamma + 1}{2(2)^{\frac{1}{2}}} \frac{P_1^2}{\gamma p_0} \cdot \frac{\omega}{c} = KP_1^2.$$

If it is assumed that the same expression represents the rate of *generation* of second harmonic along a horn, and that both the fundamental and second harmonic diverge in the same manner, the complete differential equation for P_2 becomes

$$\frac{dP_2}{dx} = KP_1^2 - \frac{m}{2} P_2 = KP_{1t}^2 e^{-mx} - \frac{m}{2} P_2,$$

where P_1 and P_{1t} are, respectively, the r.m.s. fundamental pressures at the point in question (x) and in the throat of the horn. The solution consistent with the condition $P_2 = 0$ at $x = 0$ is

$$P_2 = [KP_{1t}^2/(m/2)][e^{-mx/2} - e^{-mx}]. \tag{18}$$

Since $P_1 = P_{1t} e^{-mx/2}$, the ratio of the second harmonic pressure to the fundamental pressure at any point x in the horn is thus

$$\frac{P_2}{P_1} = KP_{1t} \cdot \frac{1 - e^{-mx/2}}{m/2} = \frac{\gamma + 1}{2\sqrt{2}} \cdot \frac{P_{1t}}{\gamma p_0} \cdot \frac{\omega}{c} x', \tag{19}$$ *

* According to this equation the magnitude of the second harmonic pressure generated in the air of a horn is 6 db lower than that given by Rocard's equation previously published in the paper "Loud Speakers and Microphones" (Reference 1), page 264, equation 1.

where

$$x' = (1 - e^{-mx/2})/(m/2). \tag{20}$$

Equations (19) and (20) indicate that the second harmonic at the mouth of an exponential horn of length x is equivalent to the second harmonic at the end of a tube of length x' and of uniform section, equal to the area of the throat of the horn. Thus the second harmonic in the mouth of a horn having an index of taper $m = 0.075$ cm^{-1}, cut-off frequency 200 c.p.s. and length 78 cm is equal to the second harmonic at the end of a straight tube 25 cm long and of the same diameter as the throat of the horn.

Exponential Horn Measurements

Measurements of the output of a horn attached to a moving coil receiver were made in an acoustically damped room. The horn had a throat diameter of 3.8 cm, a length of 78 cm and a cut-off frequency of 200 c.p.s. The diaphragm of the loud speaker was coupled to the throat of the horn through a straight tube 13 cm long. A filtered single frequency tone was impressed on the loud speaker and the sound was picked up by a small microphone in front of the horn. The fundamental and second harmonic voltages from the microphone amplifier were separated by means of a band-pass filter and measured.

The approximate acoustic power at the throat of the horn was calculated from the known efficiency of the loud speaker and the electrical voltage and current supplied.

The measured and calculated ratios of the second harmonic pressure to the fundamental pressure at the mouth of the horn, including the effects of generation of second harmonic in both the horn and the straight tube coupling the receiver and the throat of the horn, are shown in Fig. 7 in terms of the sound output in watts. See equations

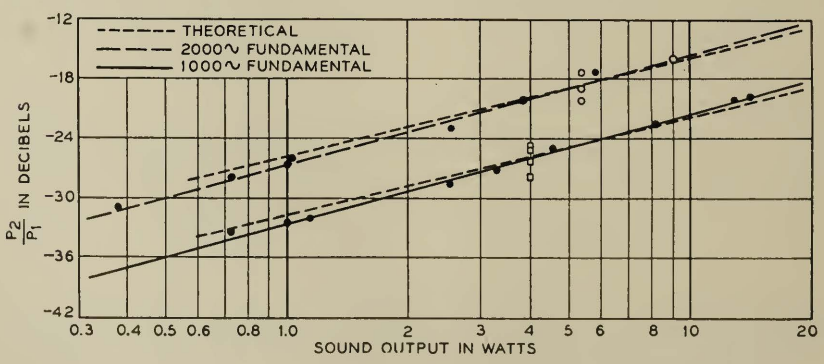


Fig. 7—2nd harmonic generated in an exponential horn *vs.* sound output.

(17) and (19). The attenuation due to viscosity and heat loss in the horn has been neglected.

A number of measurements at various microphone positions in front of the horn, shown by the dotted circles, indicate the difficulty of obtaining accurate results in a room. The average of the measurements at a number of random positions in front of the horn at a constant sound power output and the measurements at a single position for various sound outputs gives the plotted curve which is probably not greatly different from that which would be obtained in open air.

Figure 8 shows the measured and calculated ratios of the second harmonic pressure to the fundamental pressure for various fundamental frequencies.

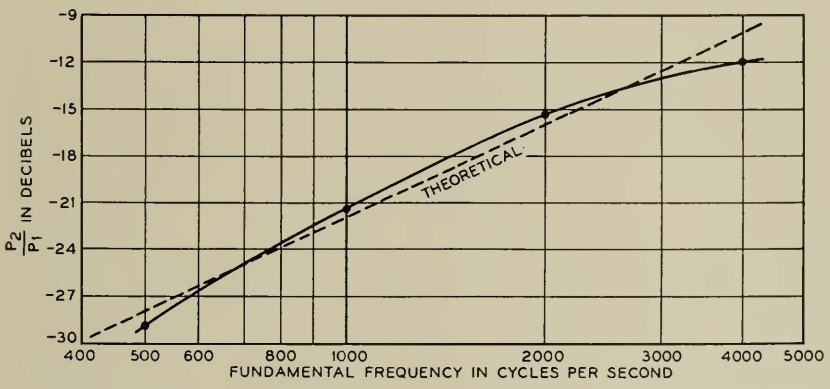


Fig. 8—2nd harmonic generated in an exponential horn *vs.* fundamental frequency (sound output = 10 watts).

Demonstration of Extraneous Frequencies

Almost two hundred years ago Sorge, a German organist, and Tartini, an Italian violinist, discovered independently, apart from all theory, that the union of two loud independent tones produced a difference tone. That this is not entirely a subjective tone produced by the ear but is actually present in the air was demonstrated by others some years later by the use of a tuned resonator. With modern apparatus consisting of power amplifiers, oscillators and tuned electromechanical vibrators it is a relatively simple matter to show not only the difference tone but the summation and harmonic tones as well.

An exponential horn was attached to the open end of the 1566 cm tube previously described but with the damping material removed. A moving coil microphone placed in front of the horn picked up the complex tone produced when two equal pure tones of 600 and 940

c.p.s. were impressed on the air. The microphone voltage was amplified and impressed on six torsional vibrators tuned to 340, 600, 940, 1200, 1540 and 1880 c.p.s. Spherical mirrors attached to the vibrators produced on a screen bands of light the amplitudes of which were approximately proportional to the relative pressures of the various frequencies in the complex tone.

For the higher power inputs to the loud speaker used in the experiment, the presence of the sum and difference frequencies and the harmonic frequencies was easily observed. At these power outputs the quality of the sound was very disagreeable and the fundamental tones could hardly be distinguished.

Conclusion

The theoretical and experimental determinations of the extraneous frequency waves generated in the air within a tube are in good agreement as regards the variation in magnitude with frequency, distance from the source and magnitude of the primary tones. The magnitude of the second harmonic is very nearly proportional to the distance from the source, to the fundamental frequency and to the square of the amplitude of the fundamental pressure. When there are two primary tones, the extraneous frequencies generated in the air include, as well as the harmonics of the primary tones, frequencies which are, respectively, the sum and the difference of the primary frequencies and also other higher order tones. The magnitudes of the summation and difference tones are very nearly proportional to the distance from the source, to the product of the magnitudes of the two primary pressures, and in each case, to the frequency of the particular combination tone. As regards the absolute magnitudes of the generated tones in the tube all of the measured values are about 3 db lower than the theoretical values.

Good agreement was obtained also between the experimental and theoretical determinations of the second harmonic generated in the air within an exponential horn, as regards proportionality to the fundamental frequency and power and also as to absolute magnitude. In fact the agreement in absolute magnitude was closer for the horn than for the tube, but not much significance should be attached to this fact, as the horn theory is developed from the theoretical solution for the tube and the horn measurements are known to be less reliable than the tube measurements.

Abstracts of Articles from Bell System Sources

North Atlantic Ship-Shore Radio Telephone Transmission During 1932–1933.[1] CLIFFORD N. ANDERSON. This paper extends the analysis of ship-shore radio transmission data for an additional two-year period beyond that reported on in a previous paper. Contour diagrams show the variation of signal field with time of day and distance for the winter, summer, spring, and fall seasons and for the approximate frequencies 4, 8, and 13 megacycles.

A comparison is made with the data obtained during 1930 and 1931. In general, transmission during 1932–1933 tends to be somewhat better on frequencies below about 9 megacycles and somewhat poorer on frequencies above 9 megacycles. At 4 megacycles the increase is of the order of 10 decibels, and for 13 and 17 megacycles, the decreases are about 6 and 10 decibels, respectively.

Loudness, Pitch and the Timbre of Musical Tones and Their Relation to the Intensity, the Frequency and the Overtone Structure.[2] HARVEY FLETCHER. It is generally supposed that the three psychological characteristics of a musical tone, namely, loudness, pitch and timbre, are each directly dependent upon the corresponding physical characteristics of the sound wave producing the tone, namely, intensity, frequency of vibration and overtone structure. In this paper it is shown that each of the psychological characteristics is not dependent upon a single one but upon all three of the above mentioned physical characteristics. Quantitative measurements of these relationships have been made using a large number of observers. For example, it was found that one tone may have the same pitch and loudness as another and yet have an intensity 100 times as great, the difference being due to the difference in the overtone structure; or two tones may have the same pitch and overtone structure and yet have frequencies of vibration that differ as much as 5 per cent, the difference being due to the different intensity of the two tones. An empirical formula showing the dependence of loudness upon the three physical characteristics mentioned has been formulated although no such formula has been found for showing a similar relationship for pitch. Such a relationship is shown graphically for pure tones. Also some very marked effects upon

[1] *Proc. I. R. E.*, October, 1934.
[2] *Jour. Acous. Soc. Am.*, October, 1934.

the pitch of a musical tone due to the changes of overtone structure are described.

Overvoltages on Transmission Lines.[3] C. L. GILKESON and P. A. JEANNE. Observations of line-to-ground voltages have been made under routine operating conditions on an isolated neutral system, a Petersen coil system, 3 neutral resistance grounded systems, and 2 directly grounded systems. Results of these observations are given in this paper. Measurements were made with oscillographs supplemented, on all but 2 systems, by surge recorders.

An Acoustic Spectrometer.[4] C. N. HICKMAN. A series of tuned reeds are mounted so that they may be electromagnetically driven. Each reed carries a small concave mirror with which light from an illuminated slit is brought to a focus on a screen. These slit images are lined up in the order of the reed frequencies. When a current having a complex wave, such as the speech current from a microphone, is passed through the electromagnet, the reeds and in consequence the slit images on the screen will oscillate. The driving system and the reeds are so designed that the amplitude of oscillation of each image is proportional to the strength of the corresponding harmonic component in the driving current. Therefore, by observing or photographing the slit image amplitudes, the frequencies and the relative energy content of the components of a complex current may be determined. A spectrometer of this type covering a small frequency range (50 to 3109 cycles) was built for demonstration purposes. The range of such an instrument can be extended to higher and lower frequencies.

The Measurement of Harmonic Power Output of a Radio Transmitter.[5] P. M. HONNELL and E. B. FERRELL. A method of determining the harmonic power output of a high-frequency radio transmitter is described. It is a method for measuring the power delivered by the transmitter to the antenna system, as distinguished from the more common method of measuring harmonic field strengths at specified locations. It is essentially a comparison method. The unknown harmonic power, present with the fundamental, is compared by means of a sufficiently selective receiving set with a known comparison power which is supplied in the absence of the fundamental. The method in practice seems to be accurate within about one decibel. It is applicable to the measurement of power other than harmonic power.

[3] *Elec. Engg.*, September, 1934.
[4] *Jour. Acous. Soc. Am.*, October, 1934.
[5] *Proc. I. R. E.*, October, 1934.

Piezoelectric Frequency Control.[6] F. R. LACK. This paper discusses the use made of the piezoelectric effect in designing sub-standard timekeepers and frequency generators. The nature of the piezoelectric effect is outlined and mention is made of the various classes of crystals in which it is found. The technic of setting up, electrically, various types of mechanical vibrations in piezoelectric crystals is described together with methods of obtaining very high frequencies. Other applications of the piezoelectric effect, such as to loud speakers, submarine signaling, etc., are briefly reviewed.

Reverberation Measurements in Auditoriums.[7] G. T. STANTON, F. C. SCHMID and W. J. BROWN, JR. This paper presents some of the problems encountered in attempts to measure the reverberation time in auditoriums, and indicates certain procedures found helpful in their solution. In addition, comparison is offered between results obtained with two step-by-step methods and a continuous decay.

[6] *Jour. S. M. P. E.*, October, 1934.
[7] *Jour. Acous. Soc. Am.*, October, 1934.

Contributors to this Issue

Alfred C. Beck, E.E., Rensselaer Polytechnic Institute, 1927. Test Department, New York Edison Company, summers 1926–27. Instructor in mathematics, Rensselaer Polytechnic Institute, 1927–28. Member, Technical Staff, Bell Telephone Laboratories, 1928–.

Edmond Bruce, B.S., Massachusetts Institute of Technology, 1924. Radio service, U. S. Navy, 1917–19. Western Electric Company, 1924–25; Bell Telephone Laboratories, 1925–. Mr. Bruce has been engaged in the development of short-wave radio receivers and field-strength measuring equipment, and has specialized in directive antenna systems for short-wave radio communication. He was awarded the Morris Liebmann Memorial Prize by the Institute of Radio Engineers in 1932.

A. B. Clark, B.E.E., University of Michigan, 1911. American Telephone and Telegraph Company, 1911–34; Toll Transmission Development Engineer, 1928–34. Bell Telephone Laboratories, 1934–. As Toll Transmission Development Director, Mr. Clark is engaged in work largely concerned with toll telephone and telegraph systems.

W. C. Ellis, Ch.E., Rensselaer Polytechnic Institute, 1924; Ph.D., 1927. Bell Telephone Laboratories, 1927–. Dr. Ellis has been engaged in metallurgical studies on magnetic materials and copper alloys.

R. T. Jenkins, B.S., in Electrical Engineering, Cooper Union, 1917; E.E., 1920. Engineering Department, Western Electric Company, 1916–25; Research Department, Bell Telephone Laboratories, 1925–. Mr. Jenkins' work has been largely in connection with transmission testing methods and apparatus, and with electro-acoustic calibrations.

J. B. Johnson, B.S.. University of North Dakota, 1913; M.S., University of North Dakota, 1914; Ph.D., Yale, 1917. Engineering Department, Western Electric Company, 1917–25; Bell Telephone Laboratories, 1925–. As Research Physicist, Dr. Johnson has been engaged in the development and application of special vacuum tubes and gaseous discharge devices.

M. J. KELLY, B.S., Missouri University, 1914; M.S., University of Kentucky, 1915; Ph.D., University of Chicago, 1919; Instructor of Physics, University of Kentucky, 1914–15; Research Assistant, University of Chicago, 1915–18. Western Electric Company, 1918–25; Bell Telephone Laboratories, 1925–. As Vacuum Tube Development Director, Dr. Kelly has worked on the development of thermionic and photoelectric devices.

FREDERICK B. LLEWELLYN, M.E., Stevens Institute of Technology, 1922; Ph.D., Columbia University, 1928. Western Electric Company, 1923–25; Bell Telephone Laboratories, 1925–. Dr. Llewellyn has been engaged in the investigation of special problems connected with radio and vacuum tubes.

LEWIS R. LOWRY, B.S. in Electrical Engineering, University of Washington, 1927. Pacific Telephone and Telegraph Company, summers 1922–26. Member, Technical Staff, Bell Telephone Laboratories, 1927–.

H. T. O'NEIL, B.S. in Electrical Engineering, State College of Washington, 1923. Engineering Department, Western Electric Company, 1923–25. Bell Telephone Laboratories, 1925–. Mr. O'Neil has been engaged in theoretical acoustic research.

H. A. PIDGEON, B.S., Ohio University, 1911; M.S., 1912; Ph.D., Cornell University, 1918. Western Electric Company, Engineering Department, 1918–25. Bell Telephone Laboratories, 1925–. As Vacuum Tube Development Engineer, Dr. Pidgeon is engaged in research and development work on vacuum tubes.

A. L. SAMUEL, A.B., College of Emporia (Kansas), 1923; S.B. and S.M. in Electrical Engineering, Massachusetts Institute of Technology, 1926. Instructor in Electrical Engineering, Massachusetts Institute of Technology, 1926–28. Bell Telephone Laboratories, 1928–. Mr. Samuel has been engaged in research and development work on vacuum tubes.

E. E. SCHUMACHER, B.S., University of Michigan; Research Assistant in Chemistry, 1916–18. Engineering Department, Western Electric Company, 1918–25; Bell Telephone Laboratories, 1925–. As Assistant Research Metallurgist, Mr. Schumacher is in charge of a group whose work relates largely to research studies on metals and alloys.

A. L. THURAS, B.S., University of Minnesota, 1912; E.E., 1913. Laboratory assistant with U. S. Bureau of Standards, 1913–16. Graduate student in physics, Harvard, 1916–17. Engineering Department, Western Electric Company, 1920–25. Bell Telephone Laboratories, 1925–. Mr. Thuras has been engaged in studies and development of electro-acoustic devices and instruments.

VOLUME XIV APRIL, 1935 NUMBER 2

THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

Cable Crosstalk—Effect of Non-Uniform Current Distribution in the Wires—*R. N. Hunter and R. P. Booth* 179

Experiments with Directivity Steering for Fading Reduction—*E. Bruce and A. C. Beck* 195

A General Theory of Electric Wave Filters—*H. W. Bode* 211

Ideal Wave Filters—*H. W. Bode and R. L. Dietzold* . 215

Ultra-Short-Wave Propagation: Mobile Urban Transmission Characteristics—*C. R. Burrows, L. E. Hunt and A. Decino* 253

An Application of Number Theory to the Splicing of Telephone Cables—*H. P. Lawther, Jr.* 273

Contemporary Advances in Physics, XXIX — The Nucleus, Fourth Part—*Karl K. Darrow* 285

Ferromagnetic Distortion of a Two-Frequency Wave—*Robert M. Kalb and William R. Bennett* 322

Abstracts of Technical Papers 360

Contributors to this Issue 366

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

50c per Copy *$1.50 per Year*

THE BELL SYSTEM TECHNICAL JOURNAL

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

EDITORIAL BOARD

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

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, 1935

PRINTED IN U. S. A.

The Bell System Technical Journal

Vol. XIV *April, 1935* *No. 2*

Cable Crosstalk—Effect of Non-Uniform Current Distribution in the Wires

By R. N. HUNTER and R. P. BOOTH

When wires are close to each other as they are in cable, the mutual inductance coupling between pairs is not a simple number, constant at all frequencies. Because of the non-uniform and non-symmetrical distribution of current over the cross-sections of the conductors the "effective mutual inductance" is of the form $M = M_a + jM_b$ where both M_a and M_b vary with frequency. This paper discusses the results of certain measurements which have been made of the effective mutual inductance between straight wires and between cable pairs over a wide range of frequencies extending up to a million cycles. This is of interest in connection with crosstalk problems in cable carrier telephone systems.

FOR many years it has been recognized that non-uniform distribution of current over the cross-section of a conductor reduces the efficiency of transmission in either power or communication circuits. With direct current or with alternating current of very low frequency the current is distributed almost uniformly. As the frequency increases, the current distribution becomes more and more non-uniform.

If the two conductors of a circuit are remote from each other the high-frequency current distribution in either conductor is practically symmetrical with respect to its center, the density of the current being lowest in the center of the conductor and highest near the surface of the conductor. If, however, the conductors are close together, the high-frequency current distribution in either conductor is unsymmetrical due to the proximity of the other conductor. This is known as the proximity effect.

It is probably not well known that this proximity effect may have an important bearing on the crosstalk between communication circuits.[1] While the effect is negligible in open-wire circuits, it is quite marked in cable circuits. This paper describes an investigation of the influence

[1] This effect was mentioned in the Carson-Hoyt paper on "Propagation of Periodic Currents Over a System of Parallel Wires," in the *Bell System Technical Journal* of July, 1927.

of the proximity effect on crosstalk between long non-loaded cable circuits which are being studied in connection with the development of high frequency carrier systems suitable for toll telephone cables. More specifically, the paper covers tests made to determine the influence of the proximity effect on the mutual inductance between circuits; data are given both for the case of two isolated non-twisted pairs and for the case of pairs in a quadded 19-gauge cable.

In cable carrier systems it is not practicable to operate like frequency bands in opposite directions on different pairs in the same cable without heavy shields between the pairs. The relatively large level differences that may exist between pairs transmitting in opposite directions would result in excessive crosstalk of the near-end type. Like carrier frequency bands are, therefore, transmitted in the same direction in a cable and the crosstalk between pairs used for carrier systems is of the far-end type.

It has been shown [2] that far-end crosstalk at carrier frequencies between long non-loaded cable pairs can be considerably reduced by the use of simple networks connected between the two pairs at one point in their length. The crosstalk balanced out by such networks is of the "transverse" [3] type. Crosstalk of the interaction type varies in a complicated way with frequency, and cannot, therefore, be annulled by a simple network. For any two similar circuits all the elements of transverse crosstalk, due to the unbalances occurring at various points along the line, arrive at the same time at the far end of the line. The crosstalk currents due to unbalances of the same type such as capacitance unbalances arrive in the same or opposite phase (if the circuits are perfectly smooth). It will be seen, therefore, that a properly designed network connected at one point in the line may be used to practically annul the far-end transverse crosstalk. In order to design the most effective type of network for balancing transverse crosstalk it is necessary to know the manner in which the crosstalk coupling in any elementary length varies with frequency.

The crosstalk coupling between two pairs in an elementary length may be represented by a mutual admittance and a mutual impedance. It can be considered that the voltage between the two wires of the disturbing circuit drives crosstalk currents into the disturbed circuit through the mutual admittance. The currents in the disturbing circuit acting through the mutual impedance also cause crosstalk

[2] As discussed in the Clark-Kendall paper on "Carrier in Cable" in the *Bell System Technical Journal* of July, 1933.

[3] The various types of crosstalk are discussed in the paper on "Open-Wire Crosstalk" by A. G. Chapman in the *Bell System Technical Journal* of January and April, 1934.

currents. The mutual admittance is due almost entirely to capacitive coupling, the leakance ordinarily being negligible in its effect on crosstalk coupling. This capacitive coupling varies but little with frequency and its effect on crosstalk may be balanced out by means of a simple condenser. If the proximity effect were negligible, the mutual impedance would be substantially that of a simple mutual inductance constant with frequency. The crosstalk due to this coupling would, therefore, be balanceable by means of a simple inductance coil. If, however, the proximity effect is not negligible the mutual impedance is due to a complex mutual inductance both of whose components vary considerably with frequency. This is the case in cable circuits and a complex balancing unit must be designed if the complex magnetic coupling is to be accurately simulated.

The mutual impedance, Z_M, between two circuits is by definition the negative ratio of the induced series voltage [4] (e) in the disturbed circuit to the current (I) in the disturbing circuit. Thus,

$$Z_M = -\frac{e}{I}.$$

Since the induced voltage is proportional to the time variation of the magnetic field set up by the disturbing current, it is important to visualize how this field may be altered by changes in the distribution of the current, I, over the cross section of the disturbing conductor. Four types of current distribution will be considered and the effects on Z_M noted.

In order to simplify the following qualitative explanation of the effect of current distribution on mutual impedance it will be assumed that in all cases the disturbed wire is a filament. When the disturbed wire is finite in cross section the effect is generally similar, but more complicated.

Case I—Current Concentrated in a Filamentary Disturbing Wire

In the case of a wire of infinitely small cross section the magnetic field due to a sinusoidal current, I, induces a voltage in another filamentary wire located in this field as expressed by the familiar equation

$$e = -j\omega MI.$$

The mutual impedance is a pure reactance equal to $j\omega M$, where M, the coefficient of mutual inductance, is a pure number and independent of frequency.

[4] This voltage is defined as being the negative of the value of an inserted electromotive force such as to bring the total current in the disturbed circuit to zero.

Case II—Current Uniformly Distributed in a Solid Cylindrical Disturbing Wire

Consider next the case where the total disturbing current is uniformly distributed over the cross section of a solid cylindrical wire. Such a distribution exists exactly with direct current only, but is closely approximated at very low frequencies. Since the magnetic field outside of a conductor carrying a uniformly distributed current is the same as would exist if the total current were concentrated in the center filament, the total induced voltage in a filamentary wire located in this field is again equal to $-j\omega MI$, where M is the same as in the case of two filaments similarly located in space.

Case III—Current Symmetrically Distributed in a Solid Cylindrical Disturbing Wire

The a.-c. distribution in a solid cylindrical wire is not uniform. However, when the wire is at a considerable distance from its return, the current distribution is practically symmetrical about the axis of the wire although its density varies from a minimum value at the center to a maximum value at the surface. Such a distribution is caused by the fact that the counter-electromotive force induced in a filament near the center of the wire due to the current in all of the other filaments is greater than that induced in a filament at the surface. This is the well-known skin effect.

In this case the total current, I, may be considered as distributed in infinitely thin concentric rings in any one of which the current is the same in phase and magnitude at all points. Since the field outside of one such ring is the same as would exist if all of the ring current were concentrated in a filament at the center, the total field due to the sum of the currents in all the concentric rings is the same as would exist if the total current were concentrated at the center of the wire. Thus, the total voltage induced in a filamentary wire by the field set up by a symmetrically distributed current in the disturbing wire is again expressed by $-j\omega MI$, where M is again a pure number as in the case of two filaments.

Case IV—Current Unsymmetrically Distributed in a Solid Cylindrical Disturbing Wire

If a solid wire and its return are placed close together, as in a cable pair, the a.-c. distribution is neither uniform nor symmetrical about the axis of the wire. In this case the magnetic field set up by the current in the return wire contributes to the counter-electromotive

force acting in each filament of the other conductor and causes a further redistribution of the current in that conductor over and above that due to the above mentioned skin effect. This additional alteration in current distribution is known as the proximity effect.

The resultant current distribution can no longer be symmetrical about the axis of either wire. The current in the return wire sets up greater back-electromotive forces in the filaments of the other wire which are close to it than in the more remote filaments. These back-electromotive forces tend to act in opposition to those set up by the current in the wire itself since the current in the return wire is opposite in sign. Hence, the proximity of the return wire reduces the counter-electromotive force acting in the filaments closest to it in the other wire more than it does in the filaments farther away. This results in higher current density in the sides of the wires adjacent to each other.

The current distribution due to the combined action of skin and proximity effects is shown for a pair of round copper wires in space in Figs. 1-A and 1-B.[5] The wires are No. 19 A.W.G. and are separated a distance equivalent to that between wires in 19-gauge cable pairs. The current distribution at 56 kilocycles is shown in Fig. 1-A and at 112 kilocycles in Fig. 1-B. It is seen that the tendency at the higher frequencies is for the current to concentrate on the sides of the wires adjacent to each other. With perfect conductors the current would all be on the surface of the wires and for this wire spacing would be distributed as shown in Fig. 1-C.[6] With actual conductors this distribution is approached as the frequency increases toward the highest conceivable wire communication frequency.

In addition to this unsymmetrical distribution of current with respect to magnitude the currents in various filaments in the conductor may be considerably out of phase with the current at the center. This phase shift may be quite unsymmetrical as indicated for three wire diameters in Figs. 2-A and 2-B. While similar phase shifts occur when only skin effect is present, such shifts are symmetrical about the center of the wire so that the currents at all points in a thin concentric ring have the same phase. Figure 2-A shows the phase shift at 56 kilocycles and Fig. 2-B the phase shift at 112 kilocycles. It is seen that the tendency at the higher frequencies is for the currents at different points on the surface to become in phase with each other. At infinite frequency the surface currents would be in phase.

[5] The current distribution and phase change at 56 and 112 kilocycles were computed from formulas given by Harvey L. Curtis in Bureau of Standards Scientific Paper No. 374, entitled, "An Integration Method of Deriving the Alternating Current Resistance and Inductance of Conductors."

[6] This distribution was calculated by Ray S. Hoyt.

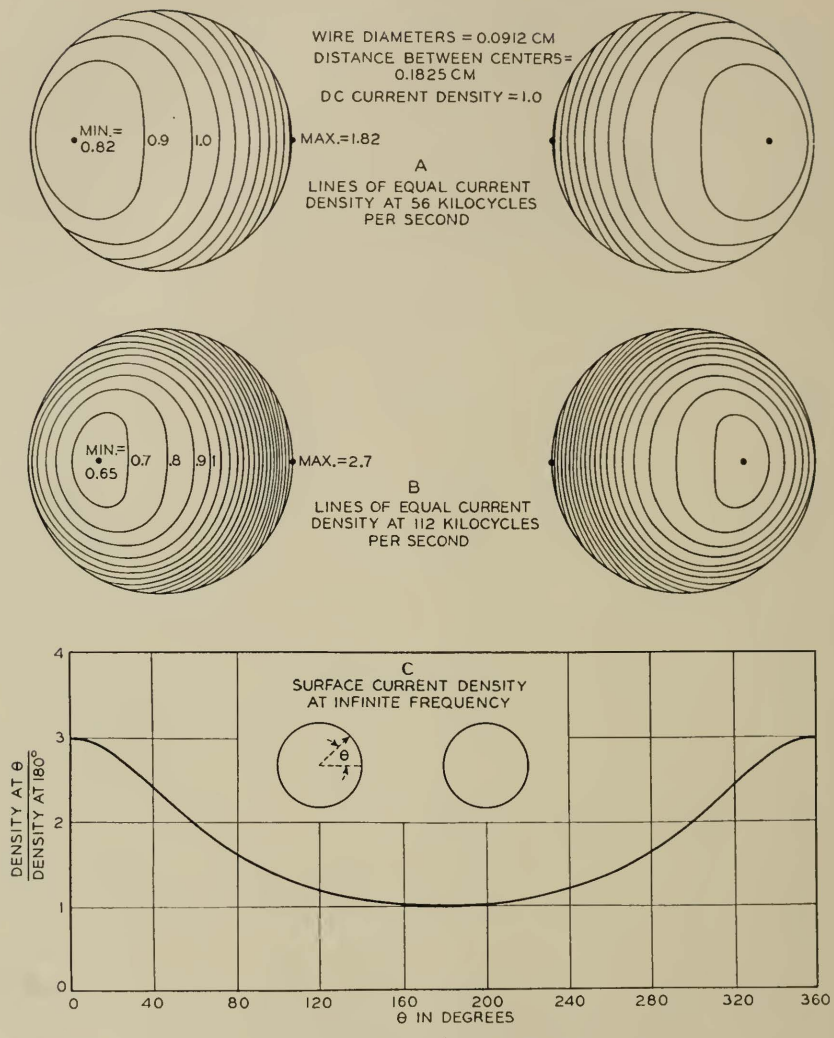


Fig. 1—Current distribution in parallel 19-ga. wires.

If a wire is carrying a total current, I, distributed unsymmetrically in phase and magnitude as in Figs. 1 and 2, the magnetic field surrounding the wire can no longer be the same as would be produced by the same current flowing in the center filament of the wire. The voltage induced in a disturbed filamentary wire located in this field must therefore be different from that induced by the field set up by any of the preceding types of current distribution in the disturbing wire. In each of those cases the induced voltage, e, was exactly in

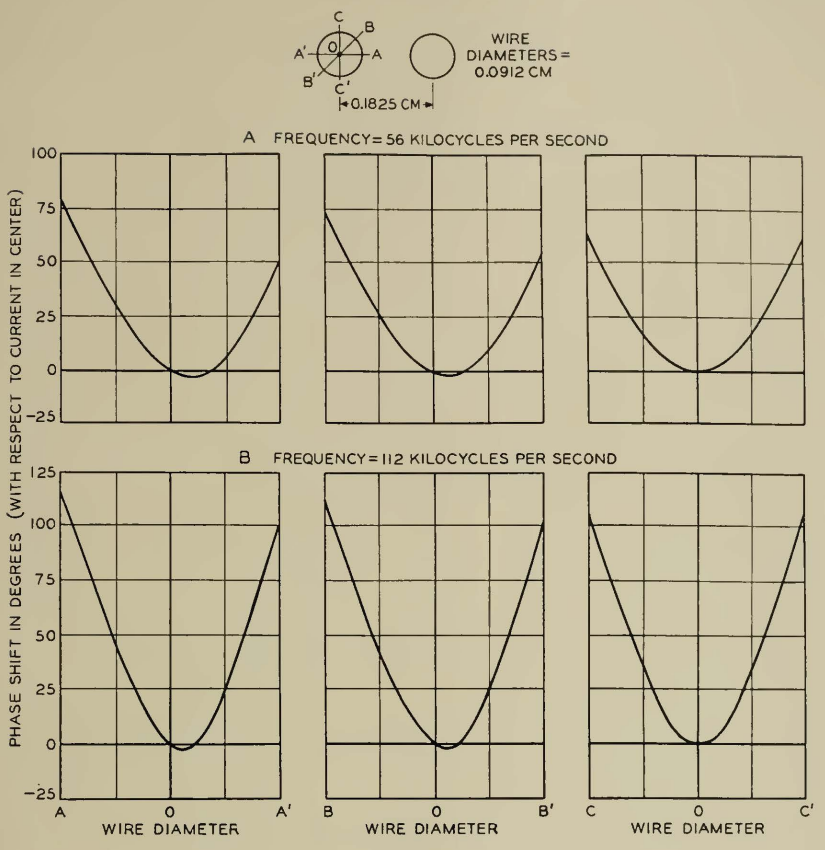


Fig. 2—Phase shift in parallel 19-ga. wires.

phase quadrature with the total disturbing current, I, and the mutual impedance was equal to $j\omega M$, a pure imaginary. For unsymmetrical current distribution in the disturbing conductor, the following discussion shows that the induced voltage can no longer be exactly in phase quadrature with the disturbing current and that the mutual impedance is complex.

The total voltage induced in a disturbed filamentary wire by the total current flowing in a solid wire is the vector sum of the induced voltages due to all of the currents in the various filaments of the disturbing wire. Thus, if $i_1, i_2, i_3, \cdots i_n$ are the vector currents in the various filaments of the disturbing wire and if $m_1, m_2, m_3, \cdots m_n$ are the corresponding coefficients of mutual inductance between each of these filaments and the disturbed filamentary wire, the total induced

voltage in the disturbed circuit is

$$e = -j\omega(m_1 i_1 + m_2 i_2 + m_3 i_3 + \cdots m_n i_n).$$

The mutual impedance, Z_M, may therefore be written

$$Z_M = -\frac{e}{I} = \frac{j\omega(m_1 i_1 + m_2 i_2 + m_3 i_3 + \cdots m_n i_n)}{I},$$

where $I = i_1 + i_2 + i_3 + \cdots i_n$. This is a general expression and holds for any type of current distribution in the disturbing conductor.

In the case of symmetrical current distribution (Case III) all filamentary currents having the value i_1 lie in a ring concentric with the center of the wire. The voltage induced in a disturbed filamentary wire due to all the currents in one such ring is the same as if their total value, I_1, was concentrated in the center of the ring. This voltage is equal to $-j\omega M_1 I_1$ where M_1 is the coefficient of mutual inductance between the center filament of the disturbing wire and the disturbed filamentary wire. The same reasoning holds for currents having values i_2, i_3, $\cdots$ i_n and the mutual impedance may be written

$$Z_M = -\frac{e}{I} = \frac{j\omega(M_1 I_1 + M_2 I_2 + M_3 I_3 + \cdots M_n I_n)}{I}.$$

But $M_1 = M_2 = M_3 = \cdots M_n = M$ since all are computed from the center filament in the disturbing wire to the disturbed filamentary wire. Then

$$Z_M = j\omega M \frac{(I_1 + I_2 + I_3 + \cdots I_n)}{I}$$
$$= j\omega M$$

since $I_1 + I_2 + I_3 + \cdots I_n = I$. This is the same expression for Z_M as given in the discussion on symmetrical current distribution.

However, when the current distribution in the disturbing wire is unsymmetrical in phase and magnitude it is impossible to make the above simplifications. In the general expression

$$Z_M = -\frac{e}{I} = \frac{j\omega(m_1 i_1 + m_2 i_2 + m_3 i_3 + \cdots m_n i_n)}{I}$$

there is no correspondingly simple way to separate the m's from the i's in the complex expression in brackets and the phase angle of the expression may be quite different from that of I. Therefore, e cannot be in phase quadrature with respect to I. In order to put the equation

for Z_M in the same form as in preceding cases the bracketed expression may be arbitrarily rewritten as

$$m_1 i_1 + m_2 i_2 + m_3 i_3 + \cdots m_n i_n = I(M_a + jM_b).$$

Then,

$$Z_M = j\omega(M_a + jM_b) = -\omega M_b + j\omega M_a,$$

where the mutual inductance is now considered complex and as having two components such that

$$M = M_a + jM_b.$$

The total current in the disturbing circuit acting through the component M_b of the mutual inductance sets up an induced voltage in the disturbed circuit in quadrature with the induced voltage due to M_a, and in the same or opposite phase as the total current in the disturbing circuit. Ordinarily the phase will be opposite and the actual values of M_b will be negative with respect to M_a.

Both M_a and M_b vary with frequency. While the total current in the disturbing wire is assumed constant, the unsymmetrically distributed currents in the various filaments change in relative magnitude and phase as the frequency changes. At very low frequencies the current is distributed nearly uniformly in phase and magnitude over the cross-section of the wire. The mutual inductance between this wire and the disturbed filamentary wire is nearly the same as the d.-c. value since M_a cannot be appreciably changed from the d.-c. value and M_b must be very nearly zero. At very high frequencies the major part of the current flows unsymmetrically on the surface of the disturbing wire but the filamentary surface currents are practically in phase with each other. This results again in a low value of M_b because the total induced voltage will be practically in phase quadrature with the total disturbing current. However, due to the unsymmetrical current distribution, the value of M_a is considerably altered from its d.-c. value. At intermediate frequencies the current distribution lies between these two extremes and produces corresponding values of M_a and M_b. Since M_b is zero for both zero and infinite frequency it is evident that a maximum value must be reached at some intermediate frequency.

As noted at the outset of this discussion, the disturbed circuit is assumed to be a filament. In all practical cases the wires involved are finite in cross section, and the reasoning outlined above must be applied to each filament of the disturbed conductor in order to get the total effect.

Discussion of Test Results

In order to study the variation with frequency of the mutual inductance between cable circuits, measurements were made on various combinations of pairs in a 55-foot length of No. 19 A.W.G. toll cable. To obtain information on the performance of the measuring apparatus, measurements were also made on the calculable case of two non-twisted pairs, six feet in length (approximately). Various separations between the two wires of a pair and three different wire gauges were used to show the change in mutual inductance for various degrees of proximity effect.

The measurements were made with a "crosstalk bridge" or admittance unbalance measuring set which permits the measurement of crosstalk in both phase and magnitude. Although the mutual inductance between two pairs may be determined from either near-end or far-end crosstalk tests, it was found that greater accuracy in M_b could be obtained from far-end tests. The computation of M_b from near-end tests involves two terms of opposite sign and of nearly the same magnitude. Consequently a small error in the reading of the crosstalk bridge may result in a considerably greater error in M_b.

The results of the tests on the six-foot non-twisted pairs are shown in Figs. 3 to 9. The data cover a range of 1 to 1000 kilocycles.

In Figs. 3 and 4 the variation with frequency of M_a and M_b is shown

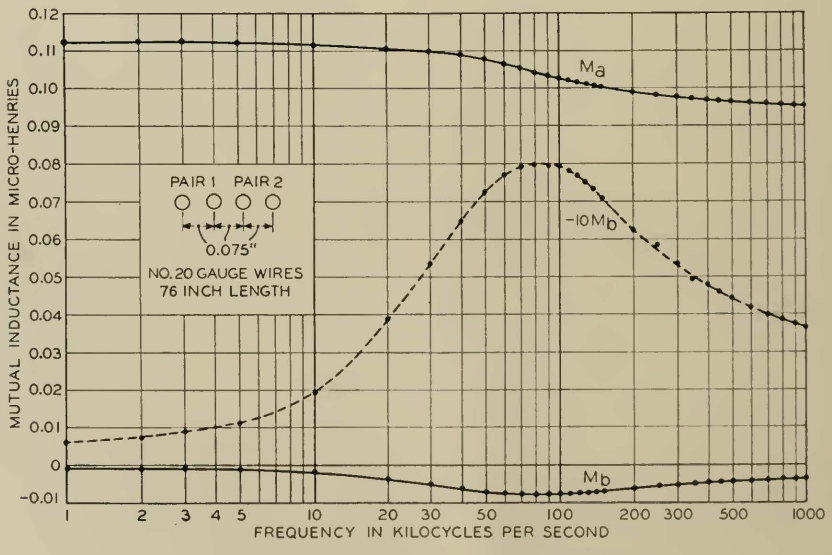


Fig. 3—Mutual inductance between pairs of parallel wires.

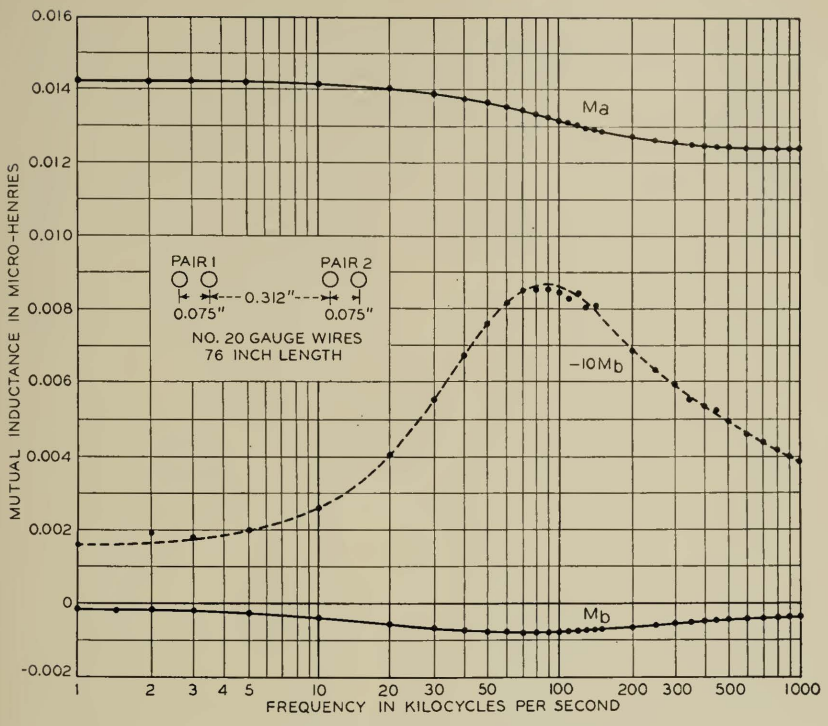


Fig. 4—Mutual inductance between pairs of parallel wires.

for two arrangements of pairs in a horizontal plane. In both cases the axial separation of the two wires of a pair was about 0.075 inch, but in Fig. 3 the axial separation between the nearest wires of the two pairs was 0.075 inch and in Fig. 4 it was 0.312 inch. The wires were No. 20 A.W.G. cotton-covered and were pulled taut to maintain accurate spacing. Two plots are shown for M_b, one actual and the other after multiplying by -10 to show the values more clearly.

The above data are replotted in Fig. 5 to show the frequency variation of M_a and M_b in terms of the values of M_a at one kilocycle. The fact that the frequency characteristics for the two cases are so nearly alike despite the difference in the magnitude of the coupling indicates that the effect depends primarily on the spacing between the wires of a pair and not so much on the relative positions of the pairs.

The proximity effect may be reduced by separating the wires of each pair as shown in Fig. 6. In this case the frequency characteristic of M_a is nearly flat and M_b is so small that it could not be plotted on the same scale as M_a; the curves shown are M_a and $100M_b$. A comparison

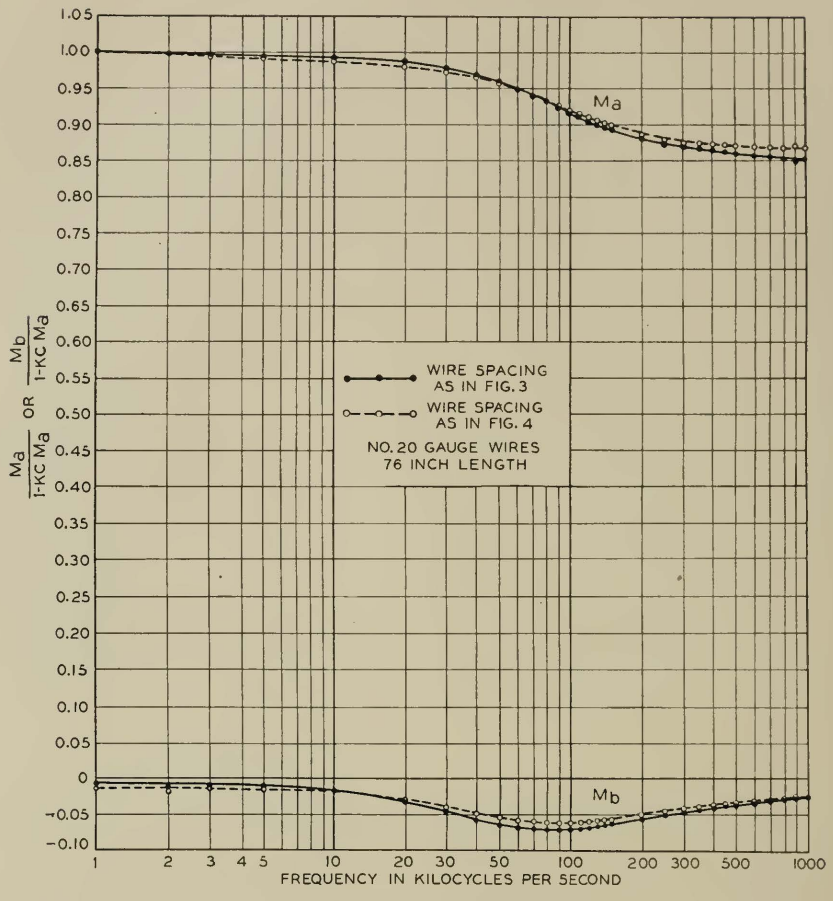


Fig. 5—Mutual inductance between pairs of parallel wires in terms of value for M_a at one kilocycle.

of the curves of M_a and M_b in Figs. 3 and 4 with the curves of Fig. 6 illustrates the relative importance of the proximity effect on magnetic crosstalk in cable pairs and in open-wire pairs. In cable pairs the wires are close together as in Figs. 3 and 4, while in open wire the separation is much greater than that shown in Fig. 6.

The effect of wire gauge on the variation of M_a and M_b is shown in Figs. 7, 8, 9-A and 9-B. Two gauges of wire (No. 10 and No. 18 A.W.G.) were used with centers located at the corners of a 0.14-inch square. The mutual impedance between the vertically adjacent pairs was measured. The corresponding values of M_a and M_b are shown in

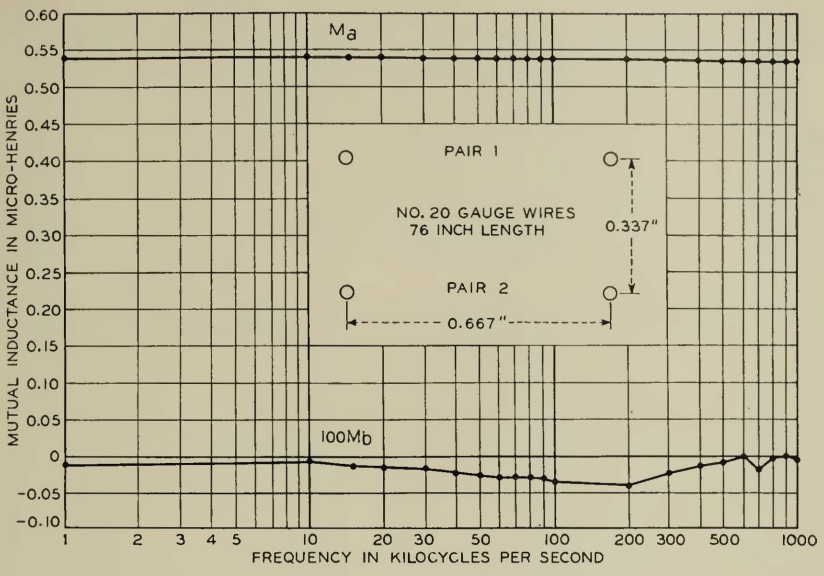


Fig. 6—Mutual inductance between pairs of parallel wires.

Fig. 7 for the 10-gauge wires and in Fig. 8 for the 18-gauge wires.[7] The values of $-10M_b$ are also plotted in Fig. 7 and $-100M_b$ in

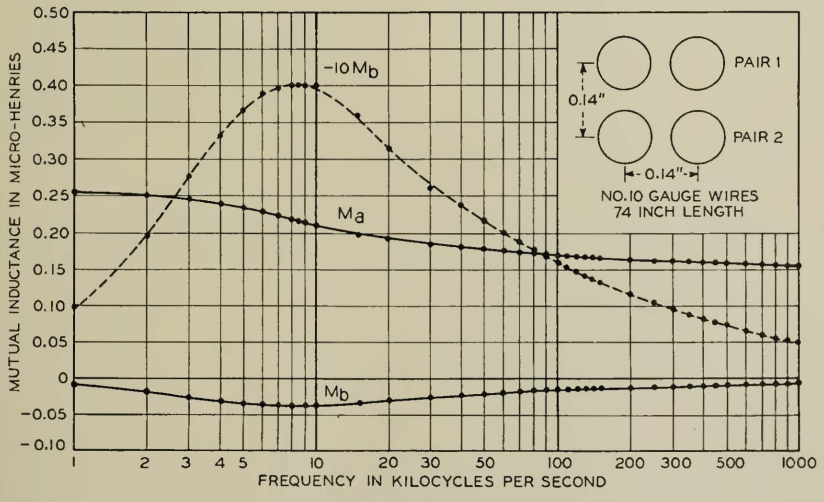


Fig. 7—Mutual inductance between pairs of parallel wires.

[7] These measured values are checked very closely by values calculated by Sallie Pero Mead from the complicated theoretical considerations involved in even the simple case of straight wires.

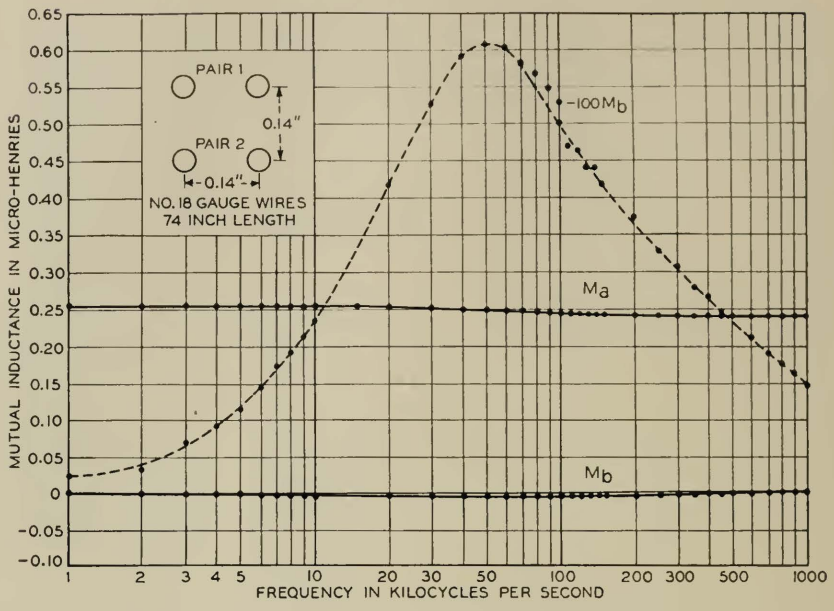


Fig. 8—Mutual inductance between pairs of parallel wires.

Fig. 8 to show the shapes more clearly. A comparison of the frequency characteristics of M_a and M_b for the two wire gauges is shown in Figs. 9-A and 9-B in terms of the one-kilocycle value of M_a for each case. In Fig. 9-A the frequency scale is logarithmic and in Fig. 9-B is linear. As would be expected, the use of smaller wires (No. 18 gauge) decreases the effect of proximity on the magnetic coupling and shifts the frequency at which M_b reaches a maximum value.

The results of tests on the 55-foot length of No. 19 A.W.G. quadded cable are shown on Figs. 10-A and 10-B. The data cover a range of 10 to 480 kilocycles. These figures show the variation with frequency of the average values of M_a and M_b in terms of the ten-kilocycle average value of M_a which is taken as unity. In Fig. 10-A the frequency scale is logarithmic and in Fig. 10-B it is linear. It will be seen that M_b is negative with respect to M_a as in the case of two pairs in space. As in Figs. 3, 4 and 7, curves are given both for M_b and for $-10M_b$, the purpose of the latter curve being to show the shape of M_b more clearly. The value of M_a decreases with frequency, becoming nearly constant above 300 kilocycles at a value 22 per cent less than the value at 10 kilocycles. The component M_b is of negative sign and at 56 kilocycles reaches a maximum value which is 13.4 per cent of M_a at this frequency.

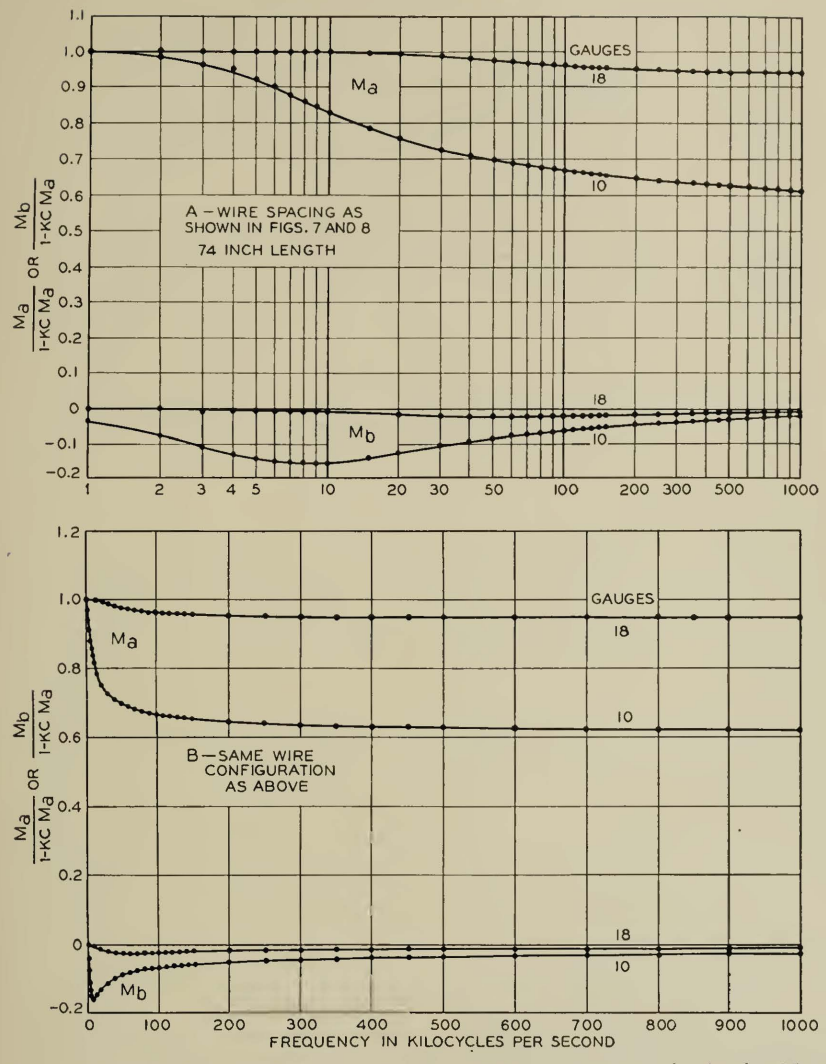


Fig. 9—Mutual inductance between pairs of parallel wires in terms of value for M_a at one kilocycle.

The frequency characteristics of M_a and M_b for individual pair combinations are about the same as shown on Fig. 10, although there are occasional differences such as a positive value of M_b or a change in sign in M_a or M_b at some frequency. Values for pairs in the outside layer did not appear to be much affected by eddy currents in the sheath. As in the case of measurements on parallel wires in space the values of M_a and M_b are very small. For example, between two pairs

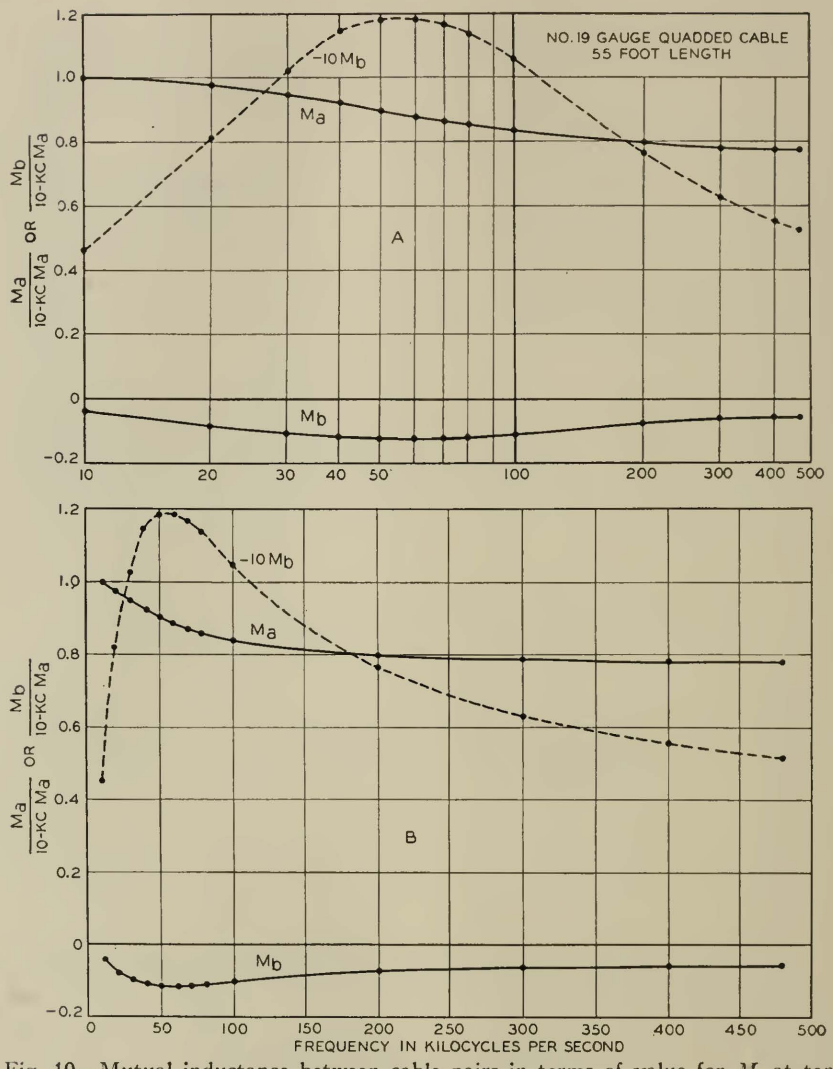


Fig. 10—Mutual inductance between cable pairs in terms of value for M_a at ten kilocycles.

in the same quad the average values at 10 kilocycles are 0.056 and — 0.0030 microhenries. For non-adjacent pairs the values are, of course, much smaller.

Acknowledgment

In this work the writers received much help from Ray S. Hoyt in the matter of general circuit theory. As previously noted, the accuracy of the measurements was established by the work of Sallie Pero Mead in developing a calculation formula for the case of straight wires.

Experiments with Directivity Steering for Fading Reduction *

By E. BRUCE and A. C. BECK

Short-wave fading is largely due to phase interference between multiple path signals of varying path length. Fortunately, stable angular differences usually exist between these paths at the point of reception. It is therefore desirable to employ antenna directivity which is "steerable" and sufficiently sharp to accept only one of the several paths in order to reduce this fading.

This paper describes experiments made with a "steerable" directive antenna during reception of transoceanic short-wave signals. The results demonstrate that sharp angular discrimination is a basically sound method of combating fading which is due to phase interference.

Introduction

RAPID fading in radio communication has been recognized for some time as being due to the interaction of distinct components having different transmission times. The possibility that these components might arrive from slightly different directions was suggested by various observed facts, among which was the behavior of sharply directive antennas.

It has been noticed in the past that fading was affected by the directivity of the receiving antenna. An example is given in the oscillograph records of Fig. 1 showing observations made by the authors some years ago at Cliffwood, New Jersey. These illustrate a condition of less fading on a large "inverted vee" [1] antenna than on a small non-directional antenna, using telegraph signals received from station GBK in England. Beating the signal with a local oscillator provided the audio frequency which was recorded. The directive antenna output was recorded on the upper trace while the lower strip recorded the output of the substantially non-directive, comparison antenna.

Such observations as these suggest the possibility of controlling and reducing fading by a systematic use of sharp directivity. The present paper reports some experiments in which changes in fading are correlated with changes in the directive pattern of a rhombic antenna [1] made by mechanically changing its shape.

It may be reasoned that, where the total differences in the path

* Published in April, 1935 issue of *I. R. E. Proc.* Presented at meeting of I.R.E., April 3, 1935.

[1] E. Bruce, "Developments in Short-Wave Directive Antennas," *Proc. I. R. E.*, Vol. 19, pp. 1406–1433, August, 1931; *Bell Sys. Tech. Jour.*, October, 1931.

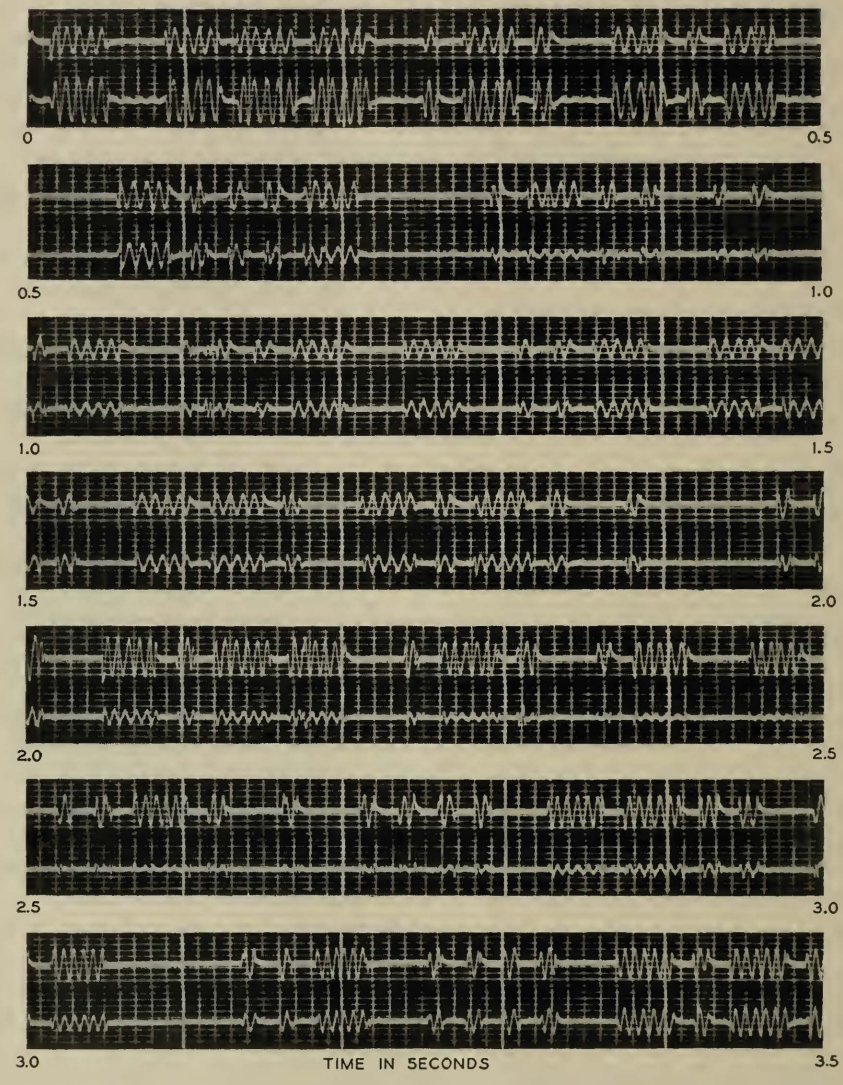


Fig. 1—Oscillographic record of carrier fading reduction. The upper trace is proportional to the output of a large "inverted vee" antenna, and the lower trace to the output of a half-wave vertical antenna when receiving station GBK on 16.6 meters. Taken at Cliffwood, N. J., November 16, 1927, 4:00 P.M., E. S. T.]

lengths are small, variations can result only in the carrier and side bands fading in and out together or in other words "general" fading. In such cases, there either may or may not be appreciable angular separation between the multiple waves at the point of reception. However, there is little question that, where multiple waves cause a "selective" fade over a speech band which is, of course, a very small percentage of the carrier frequency, a material path length difference must exist. Where this is the case, it is difficult to conceive of wave routes which do not possess appreciable angular separation between them at the place of reception. The truth of this latter point is of vital importance in this discussion.

The hope of success in fading reduction through directivity rests on the possibility of a continuous, stable angular separation between the interfering waves during times when fading is really troublesome. Fortunately this possibility is reasonably existent; therefore it should be possible to reject all but one of the interfering paths, by means of sharp directivity, with a consequent reduction in selective fading.

Description of Equipment

Tests have shown [2] that a greater degree of angular spread between the multiple waves exists in the incident vertical plane than in the horizontal plane. It might be expected, then, that such a scheme as that illustrated in Fig. 2 would be worth trying. Here the steep edge

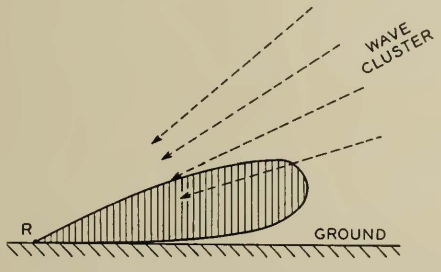


Fig. 2—Edge system for achieving fading reduction with moderate antenna directivity.

of a moderately sharp directional characteristic is moved just far enough into the wave cluster, assumed directively stable, to accept the first wave. Obviously it is possible to approach the wave cluster from the bottom as illustrated or we may approach the cluster from above.

[2] H. T. Friis, C. B. Feldman, and W. M. Sharpless, "The Determination of the Direction of Arrival of Short Radio Waves," *Proc. I. R. E.*, Vol. 22, pp. 47–78, January, 1934.

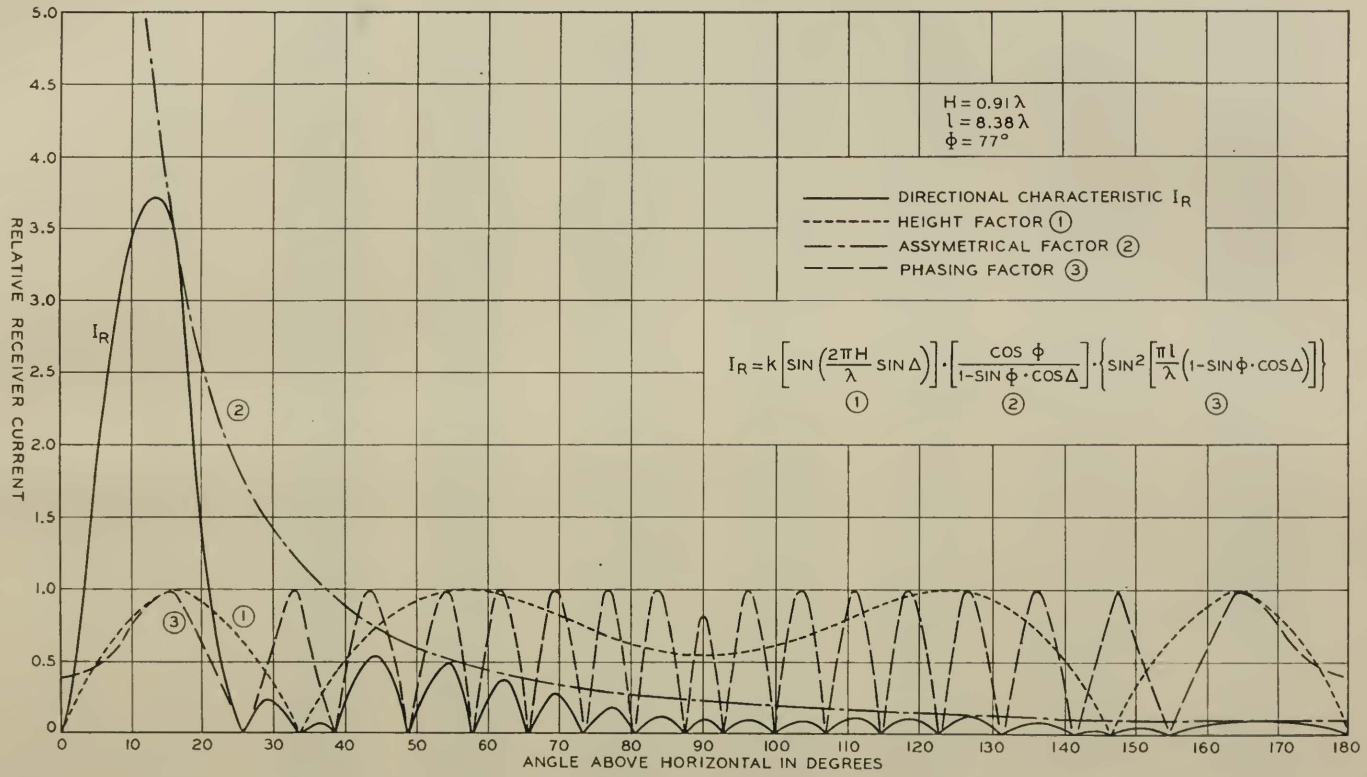


Fig. 3—Receiver current diagram of incident plane directivity for one adjustment of a steerable rhombic antenna. The factors in the equation are plotted separately.

A primary essential in this scheme is that no minor ears of the directive diagram be of appreciable size.

Using this scheme, it is not necessary to discriminate completely against the adjacent waves for practical benefit. A discrimination of ten decibels between two adjacent waves of equal amplitude will make improbable a fade deeper than 5.7 decibels from their sum. Fading of this depth would be relatively unimportant for ordinary speech transmission.

An edge wave may at times be much smaller in amplitude than the adjacent waves. The scheme under discussion may be usefully operative even in this situation since the very smallness of the edge wave means that it cannot be seriously harmful. When signals are weak, the edge of the directive diagram should be advanced until a large amplitude wave is encountered. Some fading of small depth would then exist.

It was stated above that the antenna system used should have no minor ears of appreciable size. At the same time, the edge position of the major loop must be continuously adjustable. A simple method of meeting these requirements is that of mechanically moving the elements of a "long-wire" antenna in space so as to alter the manner of its exposure to the space waves.

Figure 3 is a rectangular plot of the incident plane directive diagram of a large horizontal rhombic antenna when used for GBW on 20.78 meters. The essential antenna dimensions are indicated on that figure as well as the equation for the directive diagram.

Each bracketed quantity in the directive equation of Fig. 3 is separately plotted on that figure together with the final resulting product. Factor 3, known as the "phasing" factor, exerts the greatest influence on the shape of the major lobe. This factor contains only the variables of length l and the angle ϕ, defined as half of the side interior angle. The length cannot be made easily variable but the angle ϕ can be readily adjusted. When an adjustment in ϕ is available for this antenna, Fig. 4 gives the directions of the major lobe maxima, and the first nulls, above the horizontal for a series of wave-lengths. It is evident that a useful degree of steering is provided without limiting the desirable variable wave-length features of the antenna. In all cases, the minor ears remain small.

In Fig. 5 is shown a remote controlled power-winch system for altering the interior angles. This experimental system in slightly modified form was in operation at Holmdel, New Jersey, for some time, without any antenna breakages. This was primarily possible because the angles of flexing were very small and copper-clad steel wire was

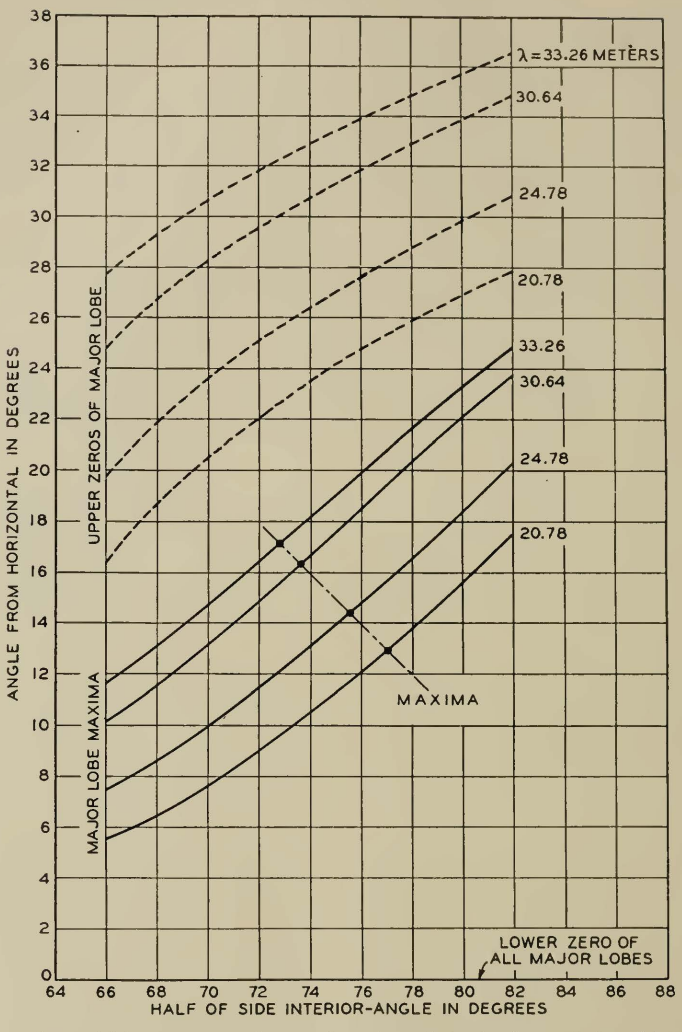


Fig. 4—Steerability, at several wave-lengths, of the horizontal rhombic antenna used for the fading reduction studies. The antenna element lengths were 184 meters and their height 19 meters.

employed in the antenna. The power-winch was equipped with automatic safety stops at the extreme positions, also with a potentiometer which was coupled to the winch to permit the use of a voltmeter as an antenna position indicator. This position indicator was located at the operator's position. By using counterweights, the required size of the winch motor is reduced.

The adopted system for observing selective fading required a fre-

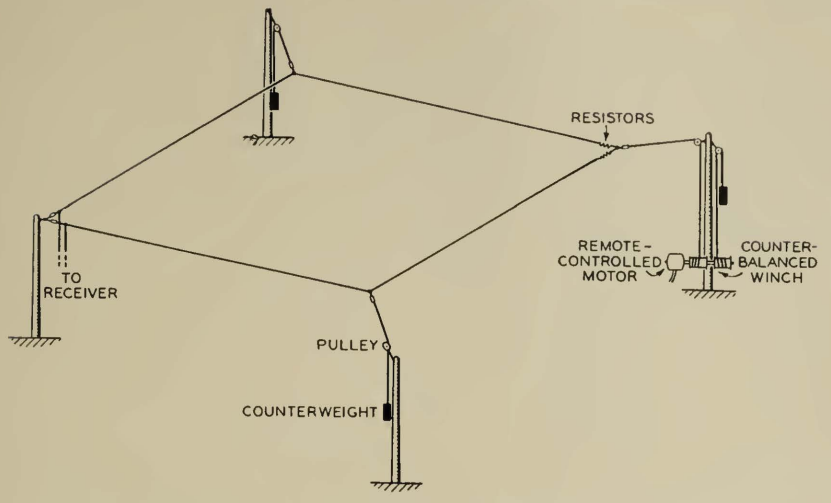


Fig. 5—Mechanical layout of the steerable horizontal rhombic antenna.

quency wobbled carrier from the transmitting station. By beating this frequency-wobbled carrier with a local fixed frequency, a wobbled audio note was obtained after detection. This audio output was impressed on the horizontal plates of a cathode ray tube, after being amplified by an audio amplifier. This produced a horizontal spot deflection on the tube screen which was directly proportional to the field strength of the signal. The vertical plates had a locally adjusted sweep circuit voltage impressed on them to produce vertical spot deflections. The sweep frequency was synchronized with the wobble rate so that the extreme upper and lower deflections occurred at the same instant as the respective upper and lower frequencies of the wobble. Figure 6 indicates the cathode ray picture of a signal without selective fading while that of Fig. 7 shows a severe case of selective fading. It is apparent that general fading was revealed by the horizontal collapse of the rectangle of Fig. 6.

It is an interesting fact that upon the first appearance of the cathode ray figure, with the wobble rates employed, it is a horizontal line moving up and down, but after a few seconds, the traced solid figure stands out clearly, due to the persistence of vision.

One of the surprising results of experience with this system was that, at times of severe fading, eight or ten depressions were occasionally seen within a sweep of a few hundred cycles.

For comparison purposes, there were two complete outfits, as described, with their cathode ray tubes mounted side by side. One outfit

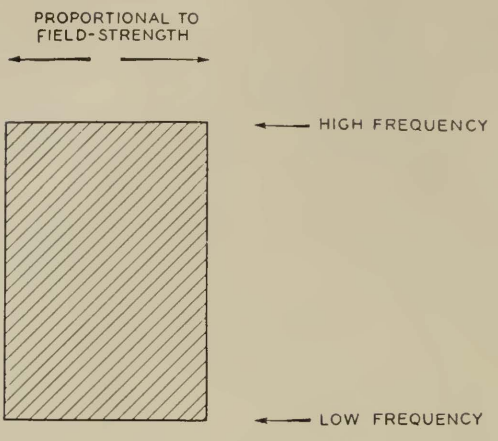


Fig. 6—Cathode ray oscillograph figure for no selective fading when observed with wobbled carrier.

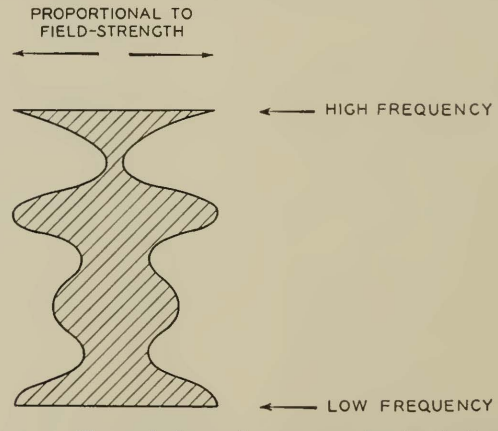


Fig. 7—Cathode ray oscillograph figure for severe selective fading when observed with wobbled carrier.

operated on a simple antenna system, as a standard of comparison, while the second was connected to the adjustable directive antenna. Fig. 10 is a photograph of this apparatus.

Other tests also going on at Holmdel, N. J., were concerned with the measurement of the comparative delay times and the respective angles of the various paths of the waves.[2] To permit this, the British Post Office transmitter sent pulses of very short duration. At the receiving point, a single transmitted pulse frequently appeared as several spaced pulses when a sweep circuit was employed. The spacing enabled the measurement of the relative time delays. It was found to

be the apparently invariable fact that the earlier arriving pulses are the lower in angle with the horizontal and are relatively stable in direction. These tests suggested that a somewhat similar scheme of observations would be useful to the present work since, if pulses were similarly employed, one would actually see the effect on each individual path of steering the antenna.

Accordingly, cathode ray equipment was constructed employing a

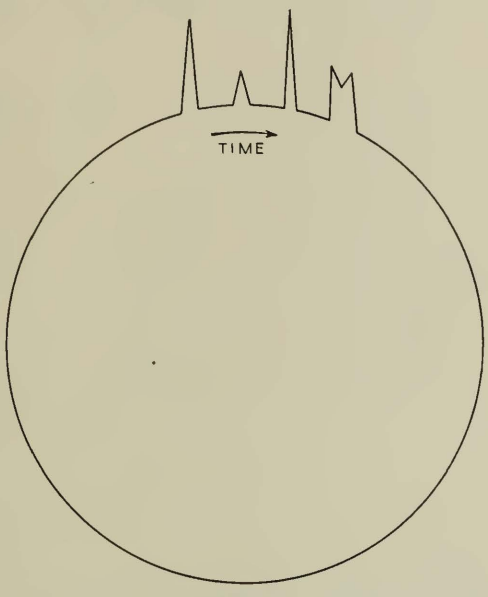


Fig. 8—Cathode ray oscillograph pulse figures when using the circular sweep circuit. The circumference is traversed by the spot in twenty milliseconds.

circular sweep system, in place of the usual linear sweep, thus making the entire time interval always in view. Figure 8 illustrates how the pulses sometimes appeared during this sweep. Since the pulses were always vertical, their definition was lost if permitted to slide down into

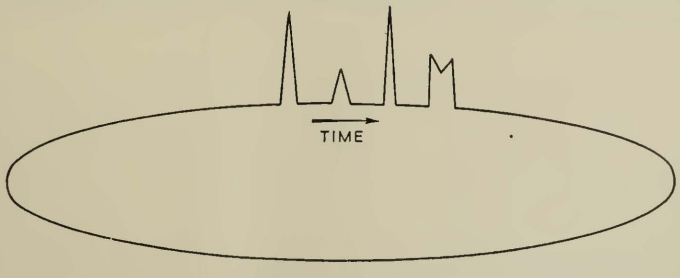


Fig. 9—Cathode ray oscillograph pulse figures when using the elliptical sweep circuit.

the "3 o'clock" or "9 o'clock" positions of the circle. This possibility was considerably reduced by employing the ellipse in Fig. 9 instead of the circle. For general observation purposes the ellipse was used but for more accurate time delay measurements the circle was employed.

The British Post Office station transmitted pulses at intervals of 0.02 second. In order to synchronize with them, an oscillator variable about 50 cycles was used to keep the pulse position stationary. A split-phase circuit feeding the four cathode ray plates produced the circular or elliptical sweep. This equipment is also shown in Fig. 10.

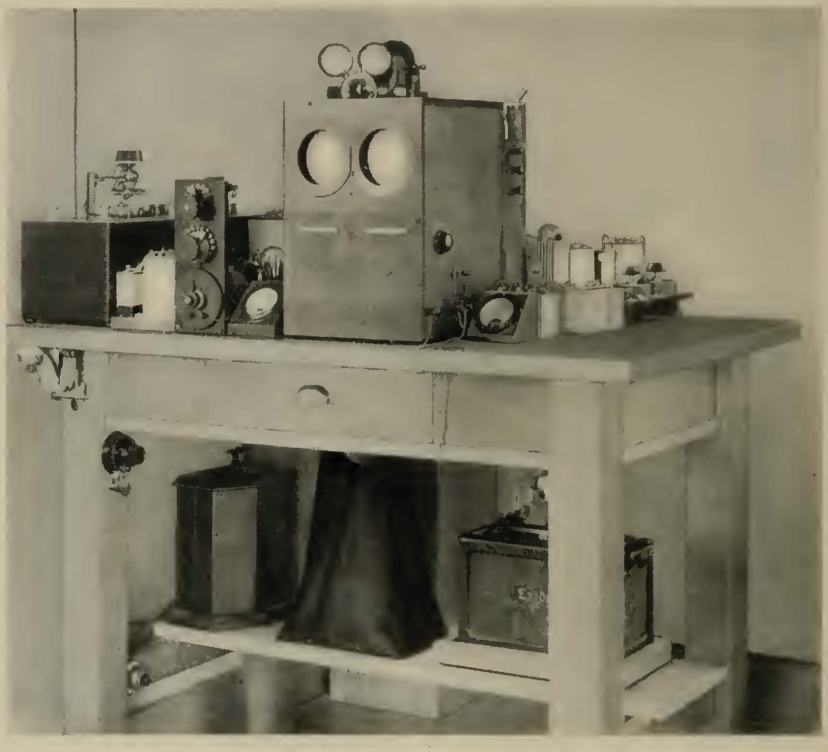


Fig. 10—Cathode ray oscillographs, their amplifiers, and the sweep circuit installation. The meter in the center of the table is the antenna position indicator.

Some studies of general carrier fading were made with a pair of magnetic counters actuated by trigger gas tubes. These fading counters were operated together with automatic recorders so as to maintain the same integrated average signal output. Since, in the recorder integration, ten-second intervals elapsed between gain readjustments, the fading counters operated to record all quick faces, during these inter-

vals, which fell below the average output level by any prescribed amount. A photograph of this equipment is shown in Fig. 11.

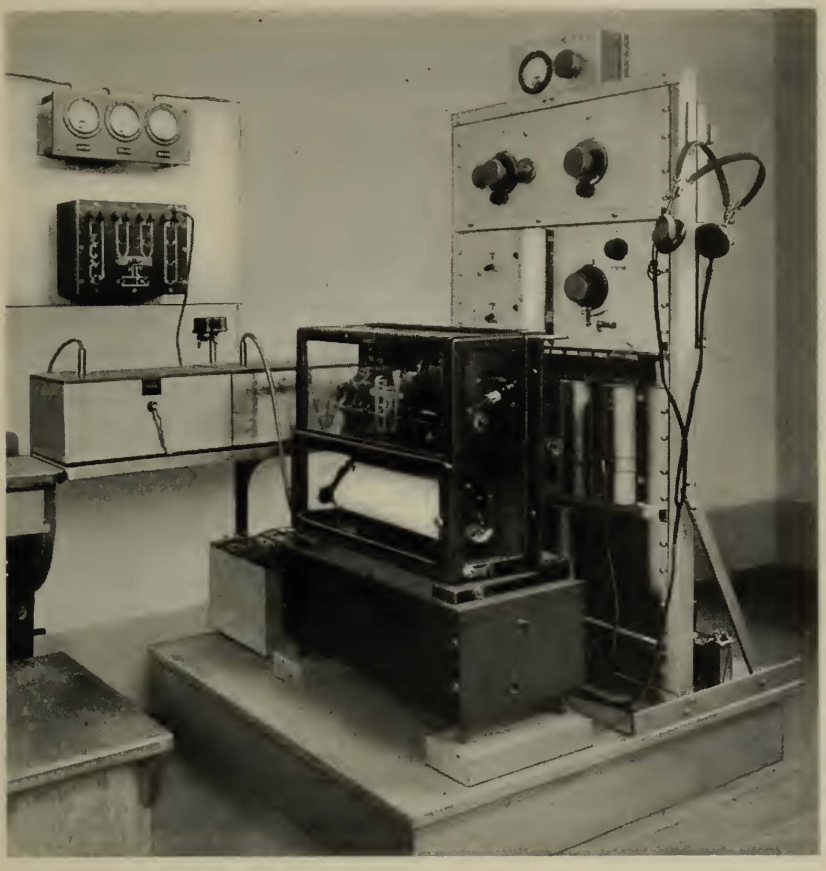


Fig. 11—Field strength recorder and fading counters used for fading reduction studies.

Results

Cathode ray tube observations of selective and also general fading were made on the British Post Office stations GBW and GBU using wobbled carrier. Whenever possible, these observations were made at half-hourly intervals. For record purposes, arbitrary numbers ranging from 0 to 4 were adopted. Zero meant very little fading (five per minute or less) and the most severe cases were represented by 4. These figures were recorded separately for the standard antenna and for the rhombus. The difference between the numbers assigned to each antenna gave an indication of the fading reduction accomplished.

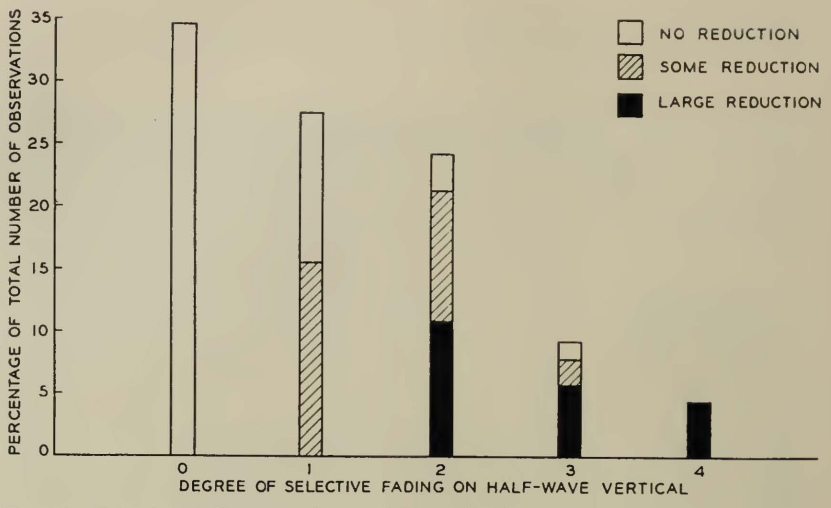


Fig. 12—Selective fading severity and its reduction at the best positions of the rhombic antenna. Stations GBU and GBW, March and April, 1933.

Figure 12 is a summary of results of these half-hourly observations made during the working hours of March and April, 1933. Disregarding the fact that portions of that figure are shaded, the total lengths of the vertical bars represent percentage of the total number of observations plotted against the degree of the selective type of fading, observed on the comparison antenna, as indicated on the abscissas.

During each of the above observation intervals, the rhombus was steered over its available range to determine the best position for reduction in selective fading. Each of the vertical bars in Fig. 12 is subdivided by shading into the various degrees of fading reduction obtainable at the best position of the adjustable rhombus. The solid sections represent large selective fading reductions, the cross-hatched sections are fair reductions, while the unshaded portions indicate that the reductions were not of appreciable magnitude.

Analyzing Fig. 12, the results show that 51 per cent of the readings gave no reduction in selective fading; however, for 35 per cent of the readings there was practically no selective fading to be reduced. On the other hand, if one disregards the rather mild and therefore relatively harmless fading cases, graded 0, 1, and 2, rhombic fading reductions were possible 89 per cent of the remaining time, so that when selective fading on the comparison antenna was really troublesome, it is important to note that an appreciable rhombic selective fading reduction was nearly always accomplished. By deliberately steering the rhombus to a disadvantageous angle, it was possible four per cent

of the time to make the selective fading worse on the rhombic antenna than on the comparison antenna, but no case has been observed where, at an ordinary rhombic antenna setting, the selective fading was not at least equal to or less than that on the comparison antenna.

While the cathode ray tube figures indicated some degree of general fading, where all frequencies fade together, it was evident that this type of fading is of far less importance than the selective type of fading, in fact it was rarely noticeable except when the selective fading was almost absent.

Figure 13 is a photograph of permanent wobble records of selective

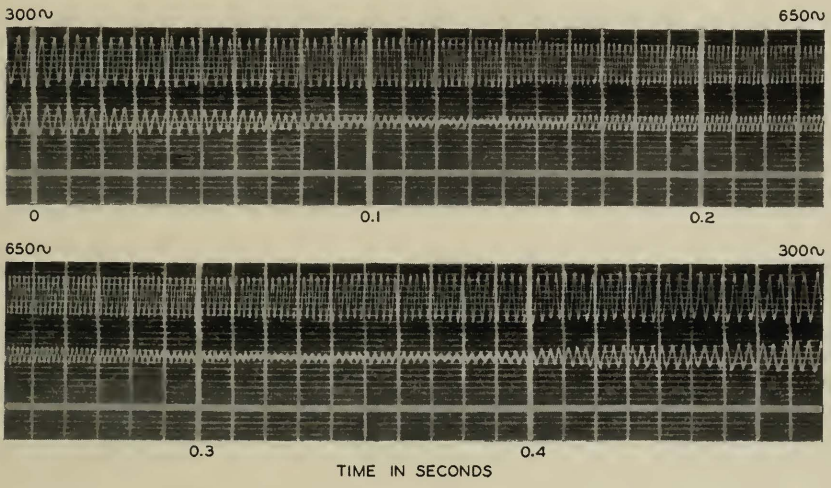


Fig. 13—Oscillographic record of selective fading reduction. The upper trace is proportional to the output of the rhombic antenna, when the angle ϕ equalled 69 degrees, and the center trace is proportional to the output of the half-wave vertical antenna. The lower string was idle. Wobbled carrier from station GBU, April 19, 1933, 4:00 P.M., E. S. T.

fading as recorded by the string oscillograph previously mentioned. The center string was actuated by the signals from the half-wave vertical comparison antenna while the rhombus signal was fed to the upper string. The third string was not utilized. The frequency wobble can be seen on close examination and as each small timing division is 0.01 second, the audio frequency is recorded. The record has been marked at the wobbled frequency extremities.

Figures 14, 15, and 16 are sketches of three interesting series of pulse patterns observed on the rhombic and comparison antennas. The three groups reading from left to right show the effects on the individual pulses of the steering of the rhombus, as indicated by the angle ϕ. The steering achieved at these angles can be seen by referring again

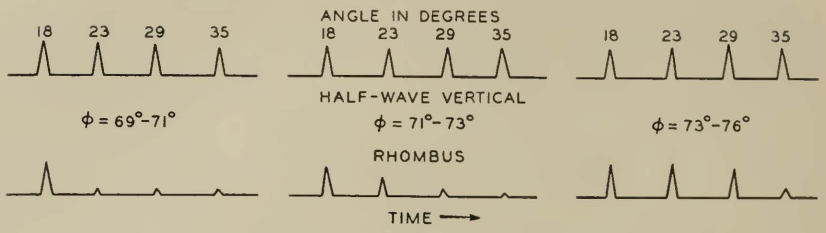


Fig. 14—Pulse pattern changes with steering, March 8 and 9, 1933. Station GCS on 33.26 meters.

to Fig. 4. Marked over the individual pulses are the arrival angles above the horizontal, measured through the cooperation of co-workers.

Figure 14 is of a test, at thirty-three meters, during a period when a wide angular spread of the cluster prevailed. Four narrow pulses of similar magnitude appear on the half-wave antenna. The progressive effect of suppressing the higher angle waves by steering the rhombus is shown. Very appreciable selective fading reductions are possible under such conditions.

Figure 15 is a sketch of twenty-meter observations during a period

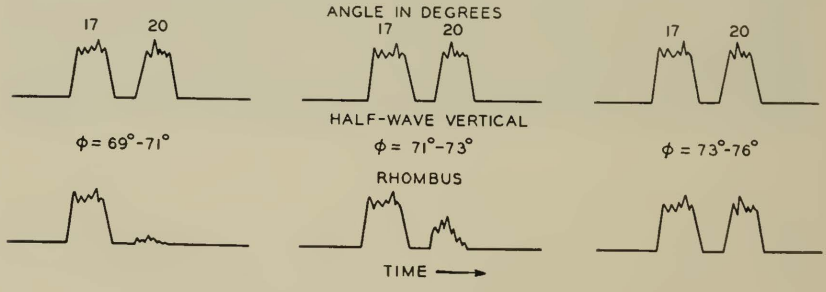


Fig. 15—Pulse pattern changes with steering, April 8, 1933, 2:00 P.M., E. S. T. Station GBW on 20.78 meters.

when selective fading reductions were achieved at the lower angle antenna settings. The broad, flat tops of the pulses are incidentally an interesting contrast to those in Fig. 14. These are possibly due to an increased horizontal spread of wave angles.

Figure 16 is of a case where it was possible deliberately to make the fading on the rhombus worse than that on the comparison antenna. Since the later pulse had a higher amplitude than the earlier one, rhombic steering by equalizing the relative amplitudes, as shown in the left-hand figure, made the selective fading very bad indeed. The opportunities for producing a result of this nature are rather rare, in fact in our previously mentioned wobble studies it was possible to make

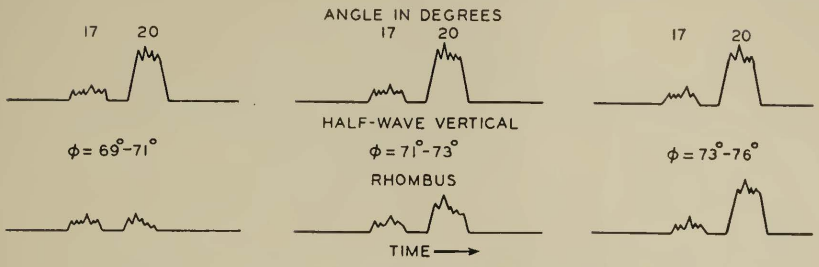


Fig. 16—Pulse pattern changes with steering, April 8, 1933, 2:40 P.M., E. S. T. Station GBW on 20.78 meters.

the selective fading worse only four per cent of the total time of observations.

Occasionally, and in particular on twenty meters, only slight selective fading was observed. When pulse transmissions were available during these times, only one major pulse could be seen. Really bad fading invariably occurs when multiple pulses, which are widely spaced in time, are observed.

It may be evident, from the previous discussions in this paper, that the change in antenna output, with steering, is closely related to the number and spread of the waves arriving and to the selective fading improvements obtainable. Figure 17 shows three cases of results secured by reading relative gain changes, as shown by automatic recorders.

Case 1 is typical of a closely spaced wave cluster arriving at an average angle of about ten degrees above the horizontal. Case 2 can be explained as due to a narrow wave cluster at eleven degrees plus another of less amplitude at eight degrees. We would ordinarily expect annoying selective fading in such an event. Should we deal with many closely spaced waves having a large angular spread, very little gain change would be evident while steering the rhombic antenna, but selective fading improvements over the comparison antenna might still be possible.

Curve 3 is of considerable interest in that it served as one of the experimental checks of the theoretical directive pattern calculations. The change in gain with steering is so well defined that probably only one wave-direction existed. This belief was supported by an absence of noticeable fading. Independent measurements, made by an average angle measuring installation[2] consisting of two horizontal dipoles at different heights which determines the average angle by the ratio of the respective outputs, gave the arrival angle at from eighteen to nineteen degrees above the horizontal. Figure 4 indicates that a ϕ-angle

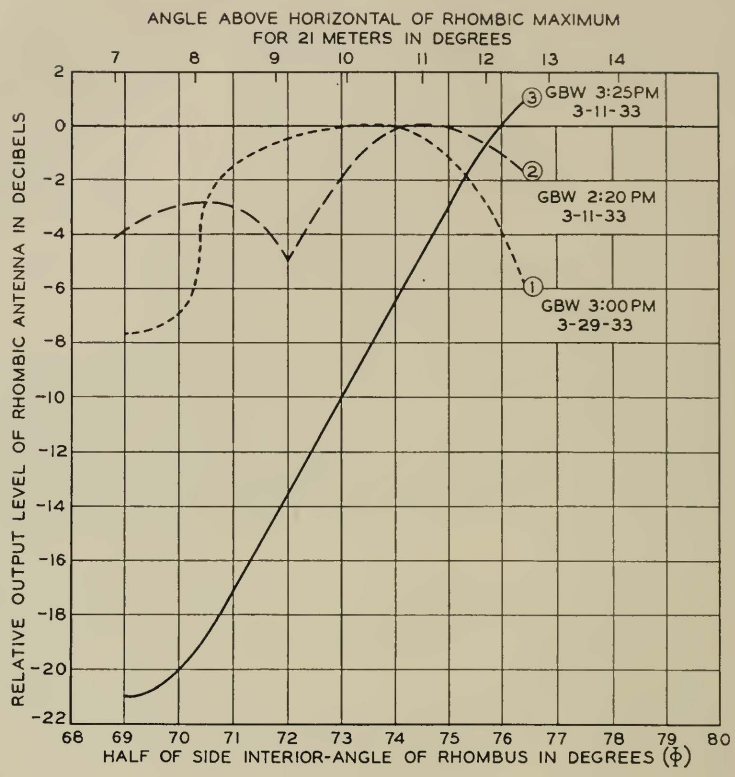


Fig. 17—Horizontal rhombic antenna output changes with steering as shown by automatic field strength recorders. Corrections for changes in signal level with time, as obtained from a half-wave vertical antenna, have been applied.

of about 68 degrees would place a null at this angle. While the range of steering of the rhombic antenna in use did not permit an adjustment to less than about sixty-nine degrees, the trend of the curve leaves little doubt as to the correctness of our null point calculation.

As might have been expected, the previously described fading counters for studying general carrier fading showed that reductions were usually obtained at the directivity positions which also gave the least selective fading. This type of apparatus is incapable of determining whether general fading or selective fading conditions are affecting the amplitude of the fixed carrier frequency.

Conclusion

It is believed that the results, discussed in this paper, demonstrate that sharp angular discrimination is a basically sound method of combating selective fading.

A General Theory of Electric Wave Filters *

By H. W. BODE

THE growth of the field of electric wave filters since their original discovery by Dr. G. A. Campbell shows the filtering action by no means inheres in any particular physical configuration. Filters have been built, for example, as recurrent or non-recurrent ladder structures, as lattices, as bridged-T's, and in a variety of combinations of transformers with ordinary elements. No general theory uniting all these configurations has, however, been developed. Each structure has been treated by methods which are primarily adapted to that configuration alone. In consequence, such questions as the relations between filters of different types and the possibilities of securing improved characteristics by going to a still wider variety of configurations have remained unsettled.

The present paper is an attempt to develop a general filter theory independent of any particular structure, by means of which these questions can be answered. For the sake of a rigorous discussion, the term "filter" has been used to signify a four-terminal network of ordinary lumped elements which, when terminated in its image impedances, transmits freely one continuous band of real frequencies and attenuates all other real frequencies. Since in actual operation the distinction between free transmission and attenuation is always more or less obscured by terminal effects and parasitic dissipation, this definition is necessarily somewhat arbitrary. It agrees, however, with common usage except in its rejection of multiple band-pass filters, which are rarely used in practice.

It follows from the definition that a "filter" can include only reactive elements. Otherwise, however, the structure considered may be an arbitrary four-terminal network and may include transformers as well as ordinary inductances and capacitances. The analysis is based upon a combination of the ordinary image parameter method of analyzing networks and the normal coordinate method familiar in the dynamics of vibrating systems. It is found that the conditions for filtering action can be expressed by means of relations between the set of normal coordinates of the network when it is short circuited at both

* This paper is a summary of a recent article with the same title appearing in the *Journal of Mathematics and Physics* of Massachusetts Institute of Technology, November, 1934. It is included largely for its value in connection with the accompanying paper on "Ideal Wave Filters."

ends and the sets obtained by open circuiting one or both ends. The same normal coordinate solutions also furnish convenient parameters in terms of which general expressions for the image parameters of the structure can be built up. For a band-pass filter, for example, the typical result is

$$\text{Tanh } \theta = k_1 \frac{\sqrt{a_{c_1}} a_{n+1} \cdots a_{p-1} \sqrt{a_{c_2}}}{a_n \cdots a_p},$$

$$Z_I = ik_2 f \frac{a_2 \cdots a_{m-1} \sqrt{a_{c_1}} \sqrt{a_{c_2}} a_{q+1} \cdots a_r}{a_1 \cdots a_m \qquad a_q \cdots a_{r-1}},$$

where a_j symbolizes a frequency factor of the form $1 - \frac{f^2}{f_j^2}$ and

$$0 \leq f_1 \leq f_2 \leq \cdots \leq f_m \leq f_{c_1} \leq f_n \leq f_{n+1} \leq \cdots \leq f_h \leq f_{c_2} \leq f_q \cdots \leq f_r \leqq \infty.$$

The formulæ are almost exactly similar to those familiar in the theory of the lattice, except that the quantities $f_1 \cdots f_n$ which are now natural frequencies of the network as determined under the previousiy described conditions have a different significance. As in the lattice, however, they fall into three groups: $f_1 \cdots f_m$ and $f_q \cdots f_r$, which affect Z_I only; $f_n \cdots f_p$, which affect θ only; and the cut-offs, f_{c_1} and f_{c_2}, which enter into both expressions. The formulæ can be extended to low-pass, high-pass and all-pass structures by allowing the cut-offs to assume the limiting values zero and infinity respectively.

Certain further restrictions upon the image impedance and transfer constant of physically realizable filters may be obtained from the consideration of another system of parameters, $r_1 \cdots r_n$, defined as the roots of the equation $\tanh \theta = 1$. They are usually of either single or double multiplicity. The importance of the roots depends upon the fact that in combination with the cut-offs they are sufficient to determine θ at all frequencies. The restrictions to which they lead may be divided into two sets. The first affects the transfer constant alone and is expressed in terms of limitations on the allowable positions and multiplicities of the roots. The second is concerned with the restrictions which must be placed upon the relation between the transfer constant and the two image impedances. It may be expressed by the statement that when the transfer constant and one image impedance have been chosen as functions of frequency, the second image impedance is determined as a function of frequency to within a constant multiplier. The differences between the two image impedances depend only upon the roots of single multiplicity, so that if only double roots are involved the structure is necessarily symmetrical.

The second half of the paper is devoted to an interpretation of known filter theory in terms of these results and to an attempt to extend this theory until it affords a definite technique for the construction of any filter which the preceding analysis has shown to be physically admissible. The method followed depends upon the fact that when a number of filter structures with matched image impedances are connected in tandem, the roots, $r_1 \cdots r_n$, of the resulting filter will be the aggregate of the roots of all the individual units of which it is composed. This allows us to represent the general filter as a composite structure in which each constituent represents one or at most a few of the total number of roots. The resulting networks are very similar to the familiar Zobel type composite filter, especially when it is noticed that the various required roots can be obtained from simple prototype structures by transformations analogous to the m-derivation, and that the preceding classification into roots of single and double multiplicity corresponds in the composite filter to a classification of the constituent structures into half and full-sections.

In spite of these relations, the usual composite filter theory must be extended in several ways if the general filter is to be adequately represented. The first extension is demanded by the fact that in the general filter we must be able to assign one image impedance characteristic of each of the constituent sections in any form compatible with the preceding general equation. The required image impedances are not obtainable from ordinary ladder structures. When the constituent involved is a double root, or full-section, structure however, the required impedance can readily be realized by resorting to the lattice form. With half-section structures the procedure is more complicated. It is necessary to make use of a combination of Dr. Zobel's multiple m-derivation and a new transformation, described as an h-derivation, which alters the impedances of ladder type half-sections without affecting their transfer constants.

Similar extensions are also needed to provide the requisite variety of transfer constants. Roots falling within certain ranges can be provided by ordinary ladder structures, m-derived, in the usual way, with a real value of m less than one. In order to complete the list, however, it is also necessary to consider single sections derived with real values of m greater than one, which may be realized as lattices or as ladder structures with mutual inductance, and pairs of sections derived with conjugate complex values of m. By including all of these types of sections we can construct physical networks giving any filter characteristics falling within the general limitations discovered in the first part of the paper.

Aside from its purely theoretical interest, the analysis leads to two results of immediate practical value. The first is the introduction of new characteristics by the complex m and h-derived sections. The h-derived structures can be dismissed briefly. They are chiefly of interest for their impedance characteristics, which resemble those found in symmetrical lattices. They allow us, however, to extend these impedances to unsymmetrical structures and they also allow us to extend considerably the range of impedances, even of symmetrical structures, which can be realized in the ladder form. The complex m structures, on the other hand, are chiefly of interest for their novel phase and attenuation characteristics. The novel phase characteristics are particularly important since the elementary constituents of the linear phase shift filters described in the accompanying paper usually turn out to be complex m sections.

The second general result is an increase in our knowledge of the relations existing between filters of different physical configurations. This last point is particularly important because it allows us to convert filters from one type of configuration to an equivalent type which may be better suited for purposes of practical construction. For example, it allows us to convert the general lattice filter, which is very convenient for theoretical purposes but is very difficult to build practically, into a composite of simple lattices and ordinary ladder sections.

Ideal Wave Filters *

By H. W. BODE and R. L. DIETZOLD

The increasing usefulness of wave filters in the telephone plant, together with rising standards of quality, emphasizes the need of a systematic method for approximating ideal characteristics as closely as we please. By an ideal filter is meant a network having the properties of a distortionless transducer over a given frequency range and suppressing all other frequencies. A design method is presented whereby an arbitrarily close approximation to these properties may be realized in a physical network. Examples of actual designs illustrate the engineering features involved in the practical application of the theory.

Introduction

IN the phenomenal advance of telephone practice during the past twenty years, almost every step has further restricted the distortion which individual parts of a transmission system can be allowed to introduce into the signal. The extension of circuits to great distances made it necessary that each link pass on to the next a more faithful copy of the signal so that the accumulated effects of many links might not endanger the intelligibility. The extension of telephone circuits to new uses, such as the transmission of pictures and the distribution of broadcast programs, imposed new demands for accuracy. Each of these has required rising standards of performance for wave filters. More than anything else, however, it has been the introduction of carrier methods, with their comparatively large utilization of selective structures, which has given prominence to the problem of reducing the distortion from wave filters. With the increase in length and complexity of carrier systems, the problem of providing wave filters which will have no harmful effect upon transmission has become one of increasing importance.

What this requires of the filters quickly appears if we recall that a structure which transmits *all* signals without distortion must (1) possess a characteristic impedance which is a pure resistance independent of frequency; (2) attenuate steady sinusoidal signals equally at all values of frequency; and (3) introduce a rotation in phase proportional to the frequency. In filter theory we need consider these requirements over only a limited band, since the signals which filters

* The reader is referred to the preceding paper entitled "A General Theory of Electric Wave Filters."

are meant to transmit, whether voice, telegraph, or television, are of a specified type having energies concentrated in certain portions of the frequency spectrum. We can therefore say that an ideal filter is one which has the ideal phase, impedance and attenuation properties in the frequency range of the desired signal and which totally suppresses all other frequencies.

The conventional ladder type filter structures which have been so extensively studied may be made to yield any desired suppression at the unwanted frequencies. In the range of wanted frequencies, however, they show wide departures from all three ideal properties. The impedance characteristic can be greatly improved by suitable elaboration of the filter structure itself, but to approximate uniformity of loss or linearity of phase shift it has been necessary to make use of supplementary networks of empirical design.[1]

The design of such corrective networks is by no means an easy task, primarily because the filter characteristics for which they are supposed to compensate change very rapidly with frequency in certain intervals. Nevertheless, much has been achieved. Thus it has been found possible to limit reflection coefficients to 2 per cent, in contrast with coefficients of 50 per cent not uncommonly tolerated in the systems of ten years ago. Improvements in the other characteristics have been comparable. In modern systems variations in attenuation of a few hundredths of a decibel, or in phase slope of a few per cent, can be attained if need be. These limits, however, demand the most patient and skillful design, and can seldom be met unless control of a single one of the characteristics is especially important. Since amplitude equalizers introduce non-linear phase, phase correctors non-uniform loss, and so on, the problem becomes increasingly difficult when close requirements must be met in several characteristics simultaneously.

By contrast, the method proposed in this paper gives the various characteristics simultaneously in a single network without recourse to auxiliary corrective structures. The method is a systematic one, requiring comparatively little in the way of cut and try design work. At the same time it preserves a measure of the flexibility of the existing technique, so that when considerable deviation from the ideal is tolerable in one or more characteristics, a corresponding economy of materials may be effected.

[1] The distortion problem has been discussed by several writers in this *Journal*. See, for example, S. P. Mead, "Phase Distortion and Phase Distortion Correction," April, 1928, p. 195; O. J. Zobel, "Distortion Correction in Electrical Circuits . . .," July, 1928, p. 438; C. E. Lane, "Phase Distortion in Telephone Apparatus," July, 1930, p. 493; E. B. Payne, "Impedance Correction of Wave Filters," October, 1930, p. 770.

The discussion which follows has a two-fold objective. The first is purely theoretical: to demonstrate that no matter how close the limits of deviation from the ideal may be set, there is a finite physical network all of whose characteristics meet these limits, except within a certain "transition interval" about each cut-off, which transition interval may also be taken as narrow as we please. This is by no means trivial; for it is known that no network, finite or infinite, can meet the ideal characteristics *exactly*.[2]

The second object is to guide the selection, from among the many networks which would meet the requirements of a given practical problem, of that one which meets them most economically. This part of the paper contains a number of examples, among them some which illustrate the use of slight empirical variations as a means of obtaining the highest measure of economy when wide deviations from the ideal are more tolerable in one respect than in others. The final example, which is segregated as Part III, deals with a situation met in picture transmission circuits, where the selectivity required is frequently small, but the effects of phase distortion may be very serious. Here a modification of the design technique leads to a filter which has comparatively modest selectivity but which exhibits a linear phase characteristic not only in the transmitting band but also in the range of rising attenuation.

Part I—Theoretical Analysis

Since linear phase shift is not available from ladder networks, the analysis will be based upon the more flexible lattice configuration. Although the lattice lends itself particularly well to the theoretical design problem, it is not so satisfactory for purposes of physical construction. After the paper design has been made, therefore, it will usually be desirable to convert it to a more suitable practical configuration. This can be done by methods described elsewhere.[3]

We may greatly simplify the theoretical discussion by ignoring the effects of parasitic dissipation—a simplification warranted by Mayer's Theorem,[4] which states that the attenuation resulting from dissipation

[2] This proposition is due to Dr. T. C. Fry, who showed that in a transducer possessing the steady-state characteristics of an ideal filter, a signal would arrive at the receiving terminals before it began to be impressed on the sending terminals. As this is absurd, we must conclude that no such system exists.

[3] H. W. Bode, "A General Theory of Electric Wave Filters," M.I.T. *Journal of Mathematics and Physics*, November, 1934. A summary of this article appears in this issue of the *Bell System Technical Journal*.

[4] H. F. Mayer, "Über die Dämpfung von Siebketten im Durchlässigkeitsbereich," *E. N. T.*, October, 1925, p. 335. His results were later somewhat extended by Feige and Holtzapfel, "Dämpfung und Winkelmass von Vierpolen mit geringen Verlusten," *T. F. T.*, July, 1932, p. 179. Even these latter results are capable of considerable generalization, so as to include other characteristics of the network besides the transfer constant.

is proportional to the derivative of the phase characteristic. The realization of a linear phase shift in the transmission band therefore automatically carries with it the satisfaction of the requirement of uniform loss in this range.[5] It can also be shown that the other characteristics of the network will not be appreciably affected by slight uniform dissipation.

Moreover, it is well known that the image impedance and transfer constant of a lattice structure are controlled by independent parameters.[6] We can, therefore, dissociate the problem of providing the required constant image impedance in the transmission band from that of providing the required loss and phase characteristics.[7] For the moment we shall fix our attention on the transfer constant.

With these simplifications our problem reduces to that of constructing a filter whose transfer constant on a non-dissipative basis represents a linear phase shift in the transmission band and an infinite loss in the attenuation bands, these being separated by narrow "transition intervals" in the neighborhood of the cut-offs. These transition intervals may be taken small at pleasure, but must be assigned in advance to insure the physical realizability of the network.

Formulation of Requirements—Low-Pass Filters [8]

If the impedances of the arms of a lattice are Z_x and Z_y, Fig. 1, it is well known that the image transfer constant and the image impedance are given by the expressions [9]

$$\tanh \frac{\theta}{2} = \sqrt{\frac{Z_x}{Z_y}}, \tag{1}$$

$$Z_I = \sqrt{Z_x Z_y}. \tag{2}$$

[5] Strictly speaking, a slight qualification should be placed upon this statement. Our process of approximating the ideal characteristics will lead to a phase shift which ripples about the desired linear characteristic, the number of ripples depending upon the number of elements used. As the number of elements is increased indefinitely, the linear characteristic is approximated more and more closely, but it is evidently not a necessary consequence of this that the slope of the ripples should approach constancy. We shall be able to show, however, that with the actual process used, the amplitude of the ripples decreases so rapidly that $dB/d\omega$ approaches constancy as B approaches linearity.

[6] This follows at once from equations (4) and (5), p. 220.

[7] A method of choosing the lattice parameters to give a substantially constant impedance in the transmission band has in fact already been obtained by W. Cauer, "Siebschaltungen," V. D. I. Verlag, Berlin, 1931; or "Ein Interpolationsproblem mit Funktionen mit Positivem Realteil," *Math. Zeit.*, November, 1933, p. 1. An alternative method will eventually be developed as a by-product of the present analysis.

[8] The extension to filters of other types is given on p. 225.

[9] G. A. Campbell, "Physical Theory of the Electric Wave Filter," this *Journal*, Vol. I, No. 2, November, 1922, p. 1.

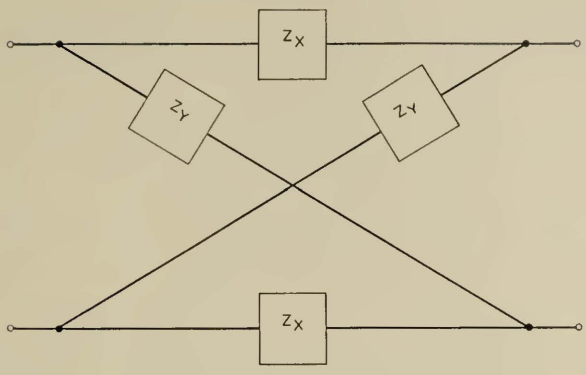


Fig. 1— The symmetrical lattice.

The relation (1) requires for transmission, i.e., for θ imaginary, that Z_x and Z_y differ in sign; for attenuation, i.e., for θ real, that Z_x and Z_y be alike in sign. In the case of the low-pass filter, this amounts to requiring correspondence of zeros (resonances) in one arm to poles (anti-resonances) in the other for $f < f_c$, and of zeros to zeros and poles to poles for $f > f_c$, where f_c, the cut-off, is a critical frequency which appears in one arm only.[10] If we denote these critical frequencies by $f_1, f_2, \cdots, f_r$ in the range below f_c, and by $f_1', f_2', \cdots, f_s'$ in the range above f_c, and if we make use of a well-known theorem [11] we readily find that Z_x and Z_y have forms similar to [12]

$$\left.\begin{aligned} Z_x &= iK_x f \frac{a_2 a_4 \cdots a_{r-1} a_c a_2' \cdots a'_{s-1}}{a_1 a_3 \cdots a_r a_1' \cdots a_s'}, \\ Z_y &= -\frac{iK_y}{f} \frac{a_1 a_3 \cdots a_r a_2' \cdots a'_{s-1}}{a_2 a_4 \cdots a_{r-1} a_1' \cdots a_s'}, \end{aligned}\right\} \tag{3}$$

[10] In the basic theory given by Dr. Campbell, in the paper just referred to, it is shown that in general a lattice having many natural frequencies is a milti-band-pass filter. The extension of the theory in the manner shown above, in which separate parameters for the control of the transfer constant and image impedance are obtained by imposing special conditions on the natural frequencies, thus rendering many bands confluent, was discovered and exploited independently by W. Cauer and one of the present writers (see Cauer, "Siebschaltungen" and later papers; or H. W. Bode, U. S. Patent No. 1828454, also "A General Theory of Electric Wave Filters," loc. cit.). The published work by Dr. Cauer gives a particularly complete discussion. It appears from a recent informal communication from Dr. Campbell to the authors, however, that this extension was also considered by him and was described briefly in the Yale-Harvard Lectures on Wave Filters delivered in 1923. The lectures have unfortunately not been published. Their content, however, is similar to that given in the discussion above.

[11] R. M. Foster, "A Reactance Theorem," this *Journal*, Vol. III, No. 2, April, 1924, p. 259.

[12] The cut-off factor a_c may appear in either the numerator or denominator of either form, and the line- and cross-arms may be interchanged. Otherwise the expression is general.

and hence that [13]

$$\tanh \frac{\theta}{2} = i \sqrt{\frac{K_x}{K_y}} f \frac{a_2 a_4 \cdots a_{r-1}}{a_1 a_3 \cdots a_r} \sqrt{a_c}, \tag{4}$$

$$Z_I = \sqrt{K_x K_y} \sqrt{a_c} \frac{a_2' \cdots a'_{s-1}}{a_1' \cdots a_s'}, \tag{5}$$

where the a's are shorthand notation for

$$a_j = 1 - \frac{f^2}{f_j^2}, \qquad a_j' = 1 - \frac{f^2}{f_j'^2}. \tag{6}$$

We shall have frequent occasion to distinguish between those critical frequencies of (4) which lie in the practical transmission band and those which lie in the transition interval. To this end we shall denote that interval by (f_A, f_B), where obviously $f_A < f_c < f_B$, and shall write P for the group of factors

$$P = i \sqrt{\frac{K_x}{K_y}} f \frac{a_2 a_4 \cdots a_{A-1}}{a_1 a_3 \cdots a_A} \tag{7}$$

which lie in the wanted band, and Q for the remainder

$$Q = \frac{a_{A+1} \cdots a_{r-1}}{a_{A+2} \cdots a_r} \sqrt{a_c} \tag{8}$$

which lie in the transition band. Then, obviously, (4) becomes

$$\tanh \frac{\theta}{2} = PQ. \tag{9}$$

Requirements on Transition Factors

With these formulæ before us, we are now prepared to attack the problem of meeting the double requirement of linear phase shift in the practical transmission band and infinite loss in the practical attenuation band. Expressed analytically, these requirements are simply

$$\tanh \frac{\theta}{2} = \tanh i \frac{\pi f}{2\alpha} = i \tan \frac{\pi f}{2\alpha}, \qquad f < f_A, \tag{10}$$

$$\tanh \frac{\theta}{2} = 1, \qquad f > f_B, \tag{11}$$

[13] It should be noted that, except for the cut-off factor a_c, (4) and (5) are entirely independent. That is, the frequencies $f_1, \ldots f_r$ may be chosen as we desire, in order to control the transfer constant, without in any way affecting the image impedance; and $f_1', \ldots, f_s'$ can be chosen at will without affecting θ. Similarly the constants $\sqrt{K_x/K_y}$ and $\sqrt{K_x K_y}$ may be chosen at will.

where α is a constant which determines the slope of the phase curve.

But it is well known that

$$i \tan \frac{\pi f}{2\alpha} = \frac{i\pi f}{2\alpha} \frac{\left(1 - \frac{f^2}{2^2\alpha^2}\right)\left(1 - \frac{f^2}{4^2\alpha^2}\right)\cdots}{\left(1 - \frac{f^2}{\alpha^2}\right)\left(1 - \frac{f^2}{3^2\alpha^2}\right)\cdots}. \tag{12}$$

If, then, in (4) we choose

$$\sqrt{K_x/K_y} = \pi/2\alpha,$$
$$f_1 = \alpha, \quad f_2 = 2\alpha, \quad \cdots, \quad f_A = A\alpha,$$

so that P becomes identical with the first A terms of (12), and if in addition we choose our unit of frequency so that [14] $(A + 1)\alpha = 1$, we readily see that in the transmitted range Q must equal

$$Q = \frac{(1 - f^2)\left(1 - \frac{f^2}{(1 + 2\alpha)^2}\right)\cdots}{\left(1 - \frac{f^2}{(1 + \alpha)^2}\right)\left(1 - \frac{f^2}{(1 + 3\alpha)^2}\right)\cdots}, \qquad f < f_A, \tag{13}$$

while by (9) and (11) in the attenuated range it must be given by

$$\frac{1}{Q} = P = \frac{i\pi f}{2\alpha} \frac{\left(1 - \frac{f^2}{2^2\alpha^2}\right)\cdots\left(1 - \frac{f^2}{(A - 1)^2\alpha^2}\right)}{\left(1 - \frac{f^2}{\alpha^2}\right)\cdots\left(1 - \frac{f^2}{A^2\alpha^2}\right)}, \qquad f > f_B. \tag{14}$$

Expressed in terms of Gamma functions, (13) and (14) become [15]

$$Q = \frac{\Gamma^4\left(\frac{1}{2\alpha}\right)\Gamma\left(\frac{1 - f}{\alpha}\right)\Gamma\left(\frac{1 + f}{\alpha}\right)}{\Gamma^2\left(\frac{1}{\alpha}\right)\Gamma^2\left(\frac{1 - f}{2\alpha}\right)\Gamma^2\left(\frac{1 + f}{2\alpha}\right)}, \qquad f < f_A; \tag{15}$$

and

$$Q = \frac{1}{4\pi i} \frac{1 - f}{2\alpha} \frac{\Gamma^4\left(\frac{1}{2\alpha}\right)\Gamma^2\left(\frac{f - 1}{2\alpha}\right)\Gamma\left(\frac{f + 1}{\alpha}\right)}{\Gamma^2\left(\frac{1}{\alpha}\right)\Gamma\left(\frac{f - 1}{\alpha}\right)\Gamma^2\left(\frac{f + 1}{2\alpha}\right)}, \qquad f > f_B. \tag{16}$$

[14] This means that we express all frequencies in terms of the first critical frequency of (12) which falls in the transition interval.

[15] The necessary transformations may be found in Whittaker & Watson, "Modern Analysis," 12.13, 12.15, and 12.33.

Asymptotic Series for P and Q

If we now take the logarithm of (15), and apply Stirling's formula, we obtain the asymptotic series

$$\log Q = \log \frac{4^{1/\alpha}\Gamma^4(1/2\alpha)}{8\pi\alpha\Gamma^2(1/\alpha)} + \frac{1}{2}\log (1 - f^2) + \sum_{r=1}^{\infty} \frac{(-1)^r(2^{2r} - 1)B_r\alpha^{2r-1}}{2r(2r-1)} \left[\frac{1}{(1+f)^{2r-1}} + \frac{1}{(1-f)^{2r-1}} \right], \quad (17)$$

where the B's are Bernoulli numbers. Since Stirling's formula holds only for $z > 0$, this expansion is valid, as inspection of equation (15) will show, only for $f < 1$, but as the unit of frequency was so chosen that $f_A < 1 < f_B$, this includes the entire wanted band, and none of the attenuating range.

If we apply a similar process to (16) we are again led to (17), except that now the range of validity is $f > 1$. But this includes the entire attenuating range, and none of the wanted band.

That is, the single formula (17) represents the lacunar [16] function Q in both ranges in which it is well defined.

We shall now determine the transition factors $a_{A+1}, a_{A+2}, \cdots, a_c$ by comparison of (8) with (17). If we adopt the notation

$$f_{A+1} = 1 + c_1, \quad f_{A+2} = 1 + c_2, \quad \cdots, \quad f_c = 1 + c_m, \quad (18)$$

so that the c's measure, not the critical frequencies themselves, but their displacement from unit frequency, each factor of (8) has the characteristic form

$$1 - \frac{f^2}{(1+c_j)^2} \equiv \frac{1 - f^2}{(1+c_j)^2}\left(1 + \frac{c_j}{1-f}\right)\left(1 + \frac{c_j}{1+f}\right);$$

whence (8) becomes

$$Q = K \frac{\left(1 + \frac{c_1}{1-f}\right)\left(1 + \frac{c_1}{1+f}\right)\cdots\sqrt{1 + \frac{c_m}{1-f}}\sqrt{1 + \frac{c_m}{1+f}}}{\left(1 + \frac{c_2}{1-f}\right)\left(1 + \frac{c_2}{1+f}\right)\cdots\left(1 + \frac{c_{m-1}}{1-f}\right)\left(1 + \frac{c_{m-1}}{1+f}\right)} \sqrt{1 - f^2}, \quad (19)$$

where K is a constant multiplier which depends on the c's. We will neglect it in this analysis since it may be readily determined later from the condition that $Q = 1$ when $f = 0$.

[16] A lacunar function is one which is well defined in several regions, but not capable of analytic continuation from one to the other.

The logarithm of Q is, of course, the sum of the logarithms of the individual factors of this expression. Expanding these as series of powers of $1/(1-f)$ or $1/(1+f)$, and collecting terms of like degree in $1/(1-f)$ and $1/(1+f)$, we obtain

$$\begin{aligned}\log Q = {} & \frac{1}{2}\log(1-f^2) \\ & + \left(c_1 - c_2 + c_3 - \cdots \pm \frac{1}{2}c_m\right)\left[\frac{1}{1-f} + \frac{1}{1+f}\right] \\ & - \frac{1}{2}\left(c_1^2 - c_2^2 + c_3^2 - \cdots \pm \frac{1}{2}c_m^2\right)\left[\frac{1}{(1-f)^2} + \frac{1}{(1+f)^2}\right] \\ & + \frac{1}{3}\left(c_1^3 - c_2^3 + c_3^3 - \cdots \pm \frac{1}{2}c_m^3\right)\left[\frac{1}{(1-f)^3} + \frac{1}{(1+f)^3}\right] \\ & - \cdots, \end{aligned} \tag{20}$$

where the sign of c_m is plus or minus according as m is odd or even.[17] As the terms of (20) are similar in form to those of (17),[18] we can make the first m terms identical. This leads to the equations

$$\left.\begin{aligned} c_1 - c_2 + c_3 - \cdots \pm \frac{1}{2}c_m &= -\frac{3}{2}B_1\alpha, \\ c_1^2 - c_2^2 + c_3^2 - \cdots \pm \frac{1}{2}c_m^2 &= 0, \\ c_1^3 - c_2^3 + c_3^3 - \cdots \pm \frac{1}{2}c_m^3 &= +\frac{15}{4}B_2\alpha^3, \\ c_1^4 - c_2^4 + c_3^4 - \cdots \pm \frac{1}{2}c_m^4 &= 0, \\ c_1^5 - c_2^5 + c_3^5 - \cdots \pm \frac{1}{2}c_m^5 &= -\frac{21}{2}B_3\alpha^5, \\ \cdots \quad \cdots \quad \cdots \quad &\cdots, \end{aligned}\right\} \tag{21}$$

whose simultaneous solution gives the desired transition factors.

The number m of transition factors used will depend upon the desired approximation to ideal characteristics in the practical transmitting and

[17] When the c's are evaluated it will appear that these series are all absolutely convergent—so that their termwise sum correctly represents $\log Q$—at all positive frequencies outside the interval $(1 - c_m, 1 + c_m)$. As the complexity of the network is increased, in the approach toward ideal characteristics, the interval of non-convergence closes on the reference frequency 1, and is contained by the given transition interval.

[18] The first term of (17) does not contain f, and may therefore be neglected for the same reason which led us to neglect K in (19).

attenuating ranges and the allowable width of the transition interval. It can best be determined by inspection of results given later.

The result of solving the equations (21) for the ratios c_j/α for values of m between 1 and 5 is given in the following table.

TABLE I
SPACING OF TRANSITION FACTORS

Number of Factors	c_1/α	c_2/α	c_3/α	c_4/α	c_5/α
1	−0.50000				
2	−0.14645	+0.20711			
3	−0.05032	+0.67731	+0.95526		
4	−0.01897	+0.86157	+1.49180	+1.72252	
5	−0.00760	+0.93809	+1.74806	+2.30277	+2.50080

The first of these solutions corresponds to a single frequency, the cut-off, in the transition interval. It follows the uniformly spaced critical frequencies of the practical transmission band at one-half the uniform spacing, α. The other solutions represent networks having, in addition to the cut-off, rational factors which vanish in the transition interval.

When these values of the c's are used in equation (8), with due regard for (6) and (18), the form of Q is completely determined. For example, the frequency pattern corresponding to the case $m = 3$, is illustrated by Fig. 2.

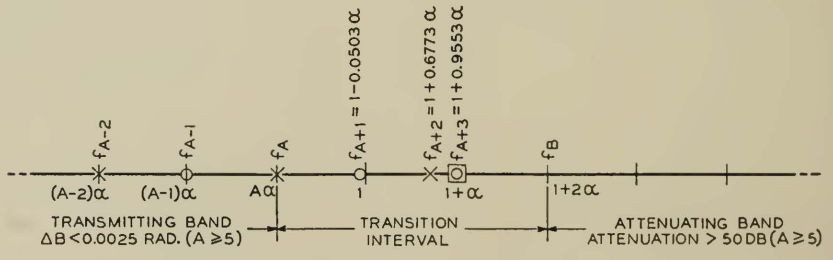


Fig. 2—Location of transition factors with $m = 3$.

Nature of the Approximation

How closely we approach ideal characteristics by this method depends on how nearly $\log Q$ is represented by the first m terms of (17) and (20). In both cases the series of omitted terms can be written in the form

$$A_{m+1}\alpha^{m+1}\left[\frac{1}{(1-f)^{m+1}}+\frac{1}{(1+f)^{m+1}}\right] + A_{m+2}\alpha^{m+2}\left[\frac{1}{(1-f)^{m+2}}+\frac{1}{(1+f)^{m+2}}\right]+\cdots, \quad (22)$$

where the A's are constants. In (20) this series is convergent. In (17) it is merely asymptotic. It is known, however, that the error due to ending (17) at any term is numerically less than the first omitted term. Since we are at present interested in small values of α, therefore, we can estimate the error in the approximation from the first term alone.

Inspection of this term shows that the error is greatest in the vicinity of the transition interval, where the factor $1/(1-f)$ is large. It depends upon all three of the quantities f, α and m; but by choosing them in the proper order there is no difficulty in showing that an indefinitely close approximation can be obtained.

The transition interval must first be selected on the absolute frequency scale. It may be as small as we choose. Next, a value of m must be chosen. What value is used is immaterial for our present purposes, although it is important for later applications. Finally, α must be taken small enough so that all transition factors lie in the prescribed transition interval. Otherwise it may be varied at will. But by choosing it small enough, the error of approximation (22) can obviously be reduced without limit for any value of f outside the interval (f_A, f_B). We may thus conclude that only considerations of expense and of manufacturing precision restrict the accuracy of approach to the ideal filter.

For purposes of future reference approximate formulæ for the attenuation and phase in the limiting condition are given below:

$$e^{-A} \doteq -\frac{1}{2} A_{m+1}\alpha^{m+1}\left[\frac{1}{(1-f)^{m+1}} + \frac{1}{(1+f)^{m+1}}\right], \tag{23}$$

$$B \doteq \frac{\pi f}{\alpha} + A_{m+1}\alpha^{m+1}\left[\frac{1}{(1-f)^{m+1}} + \frac{1}{(1+f)^{m+1}}\right]\sin\frac{\pi f}{\alpha}. \tag{24}$$

It will be seen that the attenuation rises monotonically as we recede from the transmission band while the phase curve ripples about the ideal straight line in a sinusoid of varying envelope. The ripples, of course, increase in frequency as α is diminished but since the exponent $m+1$ is always at least 2 they flatten out so rapidly that $dB/d\omega$ approaches constancy nevertheless. We may also observe that, although the absolute time of delay increases indefinitely as α decreases, it varies only as $1/\alpha$, whereas the precision of approximation can be made indefinitely great by choosing m large.

Filters of Other Types

While the preceding analysis has been restricted formally to low-pass filters, its application to filters of other transmission types is a

simple matter. We need merely repeat, in each transmission or transition interval, the rules for frequency spacing we have already developed.

The method can be understood from the study of a linear phase shift high-pass filter. Since linear phase shift demands arithmetic spacing of critical frequencies in the transmission band, it is clear that the desired characteristic cannot be obtained over the complete transmission band of the high-pass filter with a *finite* network. This difficulty will, however, be ignored for the moment. A method of modifying the analysis to give a finite filter having a linear phase characteristic in a finite interval above the cut-off will be described later.

We begin, then, by assigning to the transmission band of the structure an infinite, evenly spaced chain of critical frequencies, as in (13). The group of transition factors must evidently simulate the reciprocal of this value in the attenuation range, if the condition of high loss is to be realized; while in the transmission band, they must simulate the P of our earlier analysis if we are to obtain a linear phase characteristic. These conditions would be met by using for our transition function the reciprocal of (19), using for the c's the same values as before. Such a group, however, is not physically realizable as part of a high-pass transfer constant, since the rational factors would occur outside the theoretical transmission band. If, however, we transfer the factor $(1 - f^2)$ from (13) to (14), and seek a new Q whose values will take the reciprocals of the old, thus altered, we obtain a series identical with (17) except for a change of sign in every term but $\frac{1}{2} \log (1 - f^2)$. This change, however, reverses the sign of the right-hand members of (21), and therefore changes the sign of each c. The new solution then is the same as the original solution except that the factors occur in reverse order on the frequency scale. They can thus appropriately be combined with the remaining portion of the high-pass transfer constant expression.

A linear phase shift band-pass filter can be constructed similarly. The groups of transition factors associated with the upper and lower cut-offs should follow the arrangements prescribed, respectively, for low-pass and high-pass filters. An illustration will be found in Part II.

The Impedance Property

It will be recalled that the problem of approximating the ideal transmission characteristics for each type of filter was solved only on the assumption that the image impedance could be adjusted to a nearly uniform value in the practical transmitting band. We can

now quickly show how the desired impedance is to be obtained. It is merely necessary to observe that for any filter there exists a complementary structure with the same arrangement of critical frequencies, but having the transmitting and attenuating bands interchanged. The complementary structure is found by replacing the Z_y branch of the original lattice by the inverse impedance $Z_y' = R^2/Z_y$. When these are substituted in (1) and (2), the new transfer constant, θ', is found to be

$$\tanh\frac{\theta'}{2} = \sqrt{\frac{Z_x}{Z_y'}} = \frac{1}{R}\sqrt{Z_x Z_y} = \frac{1}{R}Z_I,$$

and the new image impedance, Z_I',

$$Z_I' = \sqrt{Z_x Z_y'} = R\sqrt{\frac{Z_x}{Z_y}} = R\tanh\frac{\theta}{2}.$$

Thus, for any filter, the problem of adjusting the image impedance to the constant R in its transmitting band is the same as the problem of adjusting $\tanh\theta/2$ to 1 in the attenuating band of the complementary filter. The latter problem, however, is merely a restatement of our original requirement of high loss in attenuating bands and has already been studied for various types of filters.

It follows from this relation that the transfer constant expressions which are appropriate for low-pass and band-pass filters furnish suitable solutions for the impedance problem in high-pass and band-elimination structures. We might also use our high-pass transfer constant expression as a low-pass impedance characteristic except for the difficulty previously mentioned that it requires an infinite number of elements. This difficulty can be avoided, however, by observing that by interchanging coils and condensers we can convert any low-pass filter into a high-pass structure having the same characteristics on a reciprocal frequency scale. We can thus use the finite low-pass solutions to obtain the required finite high-pass filter having high attenuation. For example, if we begin with a low-pass filter having three evenly spaced critical frequencies and a half spaced cut-off the resulting critical frequencies (including the cut-off) are in the ratio $1 : 7/6 : 7/4 : 7/2$.

The device of inverting the frequency scale is, of course, not available to produce a finite high-pass filter having linear phase shift throughout its transmission band since the linear phase property is thereby destroyed. It can be used, however, to produce a finite filter having linear phase shift for a limited region above its cut-off.

To see this, it is merely necessary to observe that the set of rational factors appearing in the low-pass image impedance expression described in the preceding paragraph must approximate the reciprocal of the cut-off factor at lower frequencies. We can therefore use such a set of factors to replace the upper cut-off factor of a band-pass filter, obtaining thereby a high-pass structure which approximates the ideal characteristics over a portion of the transmitting band.

If the cut-off factor of the low-pass filter transfer constant be similarly replaced by rational factors, there results an all-pass "delay network" having a constant impedance and a phase characteristic linear below the original cut-off frequency. This network is of particular interest for its relation to the classic problem of the simulation of a smooth line. As it stands, the network evidently simulates an ideal dissipationless line. To include the effects of dissipation we need merely add resistance and leakance to the coils and condensers in the proportions in which they occur in the actual line.

Part II—Design of Practical Filters

Thus far we have been interested primarily in demonstrating that an indefinitely close approximation to the ideal characteristics could be obtained when all restrictions with respect to economy of elements were removed. In practical designs, on the other hand, we wish to approximate the ideal characteristics only within moderate limits, and our interest centers upon the choice of the most economical network which will prove satisfactory. We must now reappraise the theory from this point of view.

One question which must be examined is that of determining values for m and α which will result in the most economical network meeting a prescribed standard of performance. A second is concerned with the possibility of changing the nature of the approximation with respect either to the frequency, or to the relative emphasis laid upon the phase and attenuation characteristics. In many practical designs such changes can be obtained by slight modifications of the theoretical design parameters and lead to corresponding economies in the use of elements. In investigating both questions we must remember that since α is no longer necessarily small, as it was in the theoretical analysis, the frequency interval actually occupied by the transition factors may be appreciable. Consequently it becomes important to investigate the behavior of the network in this part of the frequency range with more care than was hitherto necessary.

The variety of possible design requirements precludes the possibility of a thorough analytic treatment of these questions. The choice of

the most economical network meeting given requirements consequently cannot always be made without trial. The procedure may, however, be considerably facilitated by a study of the curves and illustrative material given in the sections which follow. The first two sections show the quantitative relations to be expected when the theoretical design parameters are adhered to strictly. The remaining sections indicate modifications obtainable by making slight changes in the theoretical parameters.

Approximate Computation of Network Characteristics

When the frequency in which we are interested is not too close to the transition interval, an approximate determination of the phase and attenuation characteristic is most easily made from (23) and (24). The A's appearing in these expressions are shown in the accompanying table.[19] In addition to A_{m+1} the table also supplies values of A_{m+2}

TABLE II

Coefficients in Series Expansions for Approximation Errors in Phase and Attenuation Characteristics

m	1	2	3	4	5
A_{m+1}.....	−0.063	−0.044	−0.051	−0.084	−0.17
A_{m+2}.....	−0.063	+0.00011	+0.10	+0.41	+1.52
A_{m+3}.....	−0.0078	+0.050	−0.047	−1.07	−7.60

and A_{m+3}, for use if additional terms in the general expression (22) are desired.

A study of equations (23) and (24) shows that, aside from the constant factor A_{m+1}, each expression can be resolved into two factors by means of which the contributions of the various design parameters can be somewhat segregated. The first factor, α^{m+1}, is chiefly important in determining the effect of various choices of α and m on the approximation error, while the factor $\left[\frac{1}{(1-f)^{m+1}} + \frac{1}{(1+f)^{m+1}}\right]$ expresses the variation of the network characteristics with frequency. In order to facilitate design work the quantity

$$-20 \log_{10} \frac{A_{m+1}}{2}\left[\frac{1}{(1-f)^{m+1}} + \frac{1}{(1+f)^{m+1}}\right]$$

[19] In preparing the table, coefficients of corresponding terms in the series expansions for (17) and (20) have been combined, so that the coefficients as given represent the accumulated errors of both approximations.

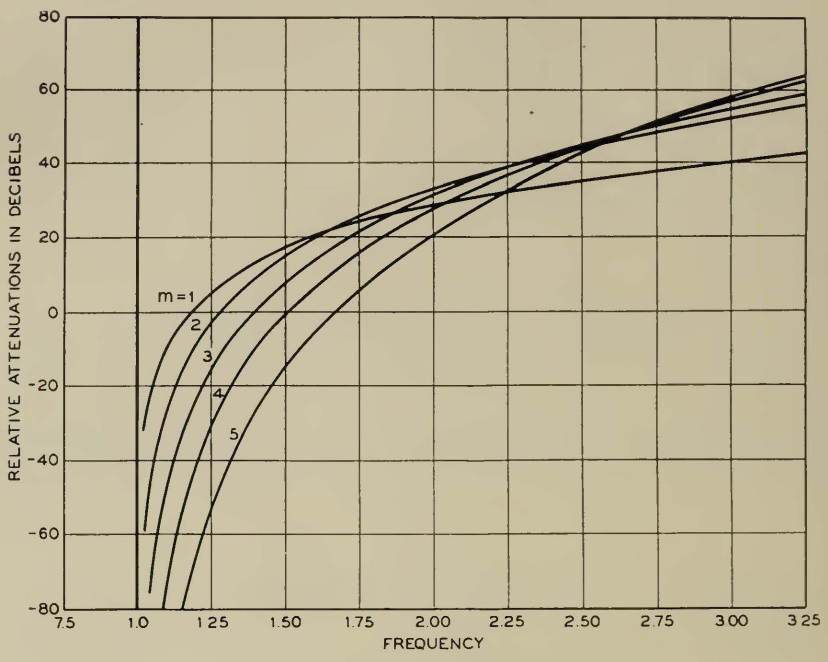


Fig. 3—Chart for loss computations.

has been computed for values of $f > 1$ and is shown plotted for various m's in Fig. 3. The approximate attenuation, in db, for any given values of α and m can be obtained from the chart by adding $20(m + 1) \log_{10} 1/\alpha$ to the appropriate curve.

A similar chart for the phase characteristic is furnished by Fig. 4, which represents the quantity $\frac{180A_{m+1}}{\pi}\left[\frac{1}{(1-f)^{m+1}} + \frac{1}{(1+f)^{m+1}}\right]$. The values given are arithmetic, although the scale is logarithmic. The approximate envelope of the ripple in the phase characteristic about the ideal straight line can therefore be found, in degrees, by multiplying the chart values by α^{m+1}.

In using these charts it should be remembered that they are based upon the approximate formulæ (23) and (24) which fail in the vicinity of the transition interval. The results, therefore, should always be checked by an exact computation.[20] It should also be observed that in complicated filters the numerical departure of $\tanh \theta/2$ from its ideal value in most frequency ranges is very small. The effects of slight errors in calculations or of small deliberate variations in the

[20] See, for example, the comparisons in Figs. 11 and 12.

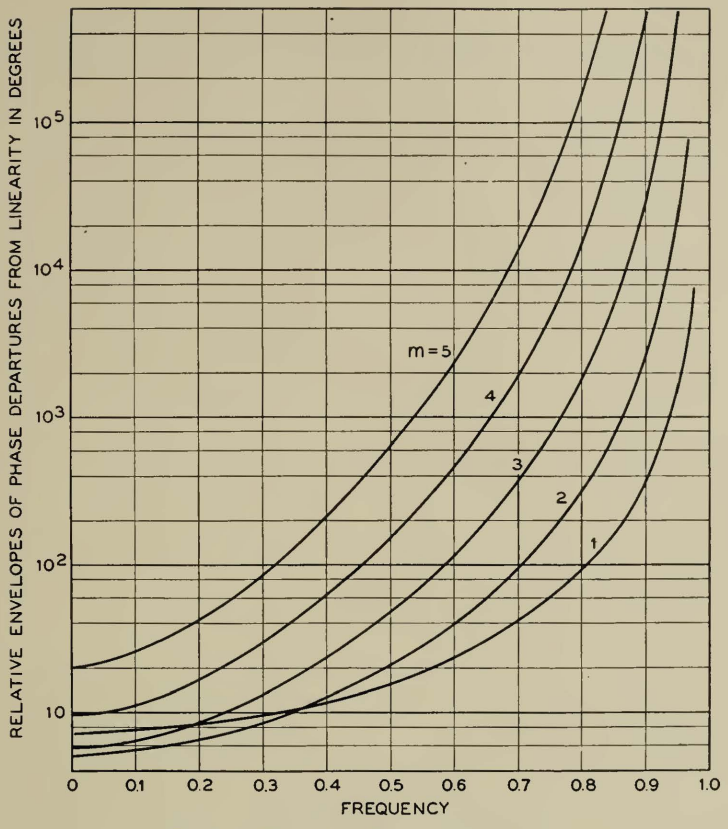


FIG. 4.—Chart for phase computations.

design parameters, may therefore be correspondingly important.[21] Since slight adjustments in the design parameters will normally occur, these charts are chiefly of value in making preliminary estimates.

It is apparent that the approximation error at a given frequency can be diminished either by increasing m or reducing α. Element for element, an increase in m is much the more powerful method. Since the total number of elements in the network is nearly proportional to

[21] A simple example is furnished by the choice of the numerical constant multiplying $\tanh \theta/2$ as a whole. It will be remembered that the constant was left undetermined in the solution for the c's. In the original equation (14) it was chosen to give the best characteristics in the neighborhood of $f = 0$. In preparing Figs. 3 and 4, on the other hand, it was chosen with reference to the characteristics near $f = \infty$, since the error expression used in these figures vanishes at that point. The two conditions are very nearly equivalent; as we can see, for the half-spaced cut-off solution at least, by means of Wallis' theorem. Since they are not identical, however, a change from one to the other may produce a relatively large, though practically unimportant, effect at extreme frequencies.

$m + (1/\alpha)$ it would therefore appear that the most economical structure meeting given requirements will be obtained by using a large m in combination with a large α. This procedure is, however, restricted by two considerations. The first is chiefly theoretical. Since the series we have been using is merely asymptotic, the successive terms obtained by choosing progressively higher m's eventually grow larger. For ordinary values of α, however, the value of m at which the series begins to diverge lies beyond the range of practical interest. A more important limitation is the fact that as we increase the number of transition factors, the width of the transition interval, as measured in terms of α, also increases. Thus, the spread between the last uniformly spaced critical frequency and the cut-off, which is $\alpha/2$ for $m = 1$ and about 2α for $m = 3$, has risen to more than 3.5α for $m = 5$. In each case a certain additional allowance is of course required for the region of rising attenuation beyond the cut-off. When the transition interval is fixed on an absolute frequency scale, therefore, the permissible values of m will depend upon the choice of α. Unless the transition interval is unusually broad only low values of m will be of practical interest.

Illustrative Characteristics

The curves shown in Figs. 3 and 4 are not of use in the neighborhood of the transition interval. To supplement them, therefore, exact computations on a number of typical structures have been made. One set was obtained by choosing $\alpha = 1/12$ and computing the characteristics corresponding to various m's. The resulting phase characteristics are shown by Fig. 5. Since the departures from linearity are too small to be noticeable when the characteristics as a whole are drawn, the figure shows only the departures themselves in terms of an envelope similar to that used for Fig. 4. The curves are drawn approximately as far as the last evenly spaced critical frequency which marks the practical limit of the range within which a high degree of phase linearity is to be expected. Since the curves vary rapidly in this vicinity, however, the fact that they are merely envelopes is important in determining the exact performance of the structure. Curves of the phase characteristics in the transition interval will be given later.

The attenuation characteristics are shown by Fig. 6. As m is increased, the cut-off moves to successively higher frequencies because of the progressively broader intervals consumed by the transition factors. Once past the cut-off, however, the curves for large values of m rise more rapidly and quickly cross the others.

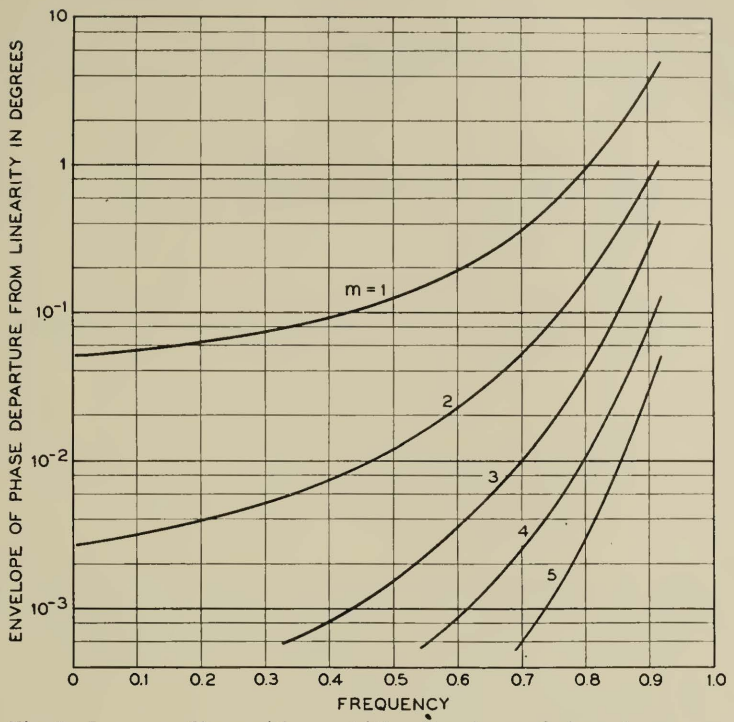


Fig. 5—Low pass filters with $\alpha = 1/12$. Envelopes of phase departures.

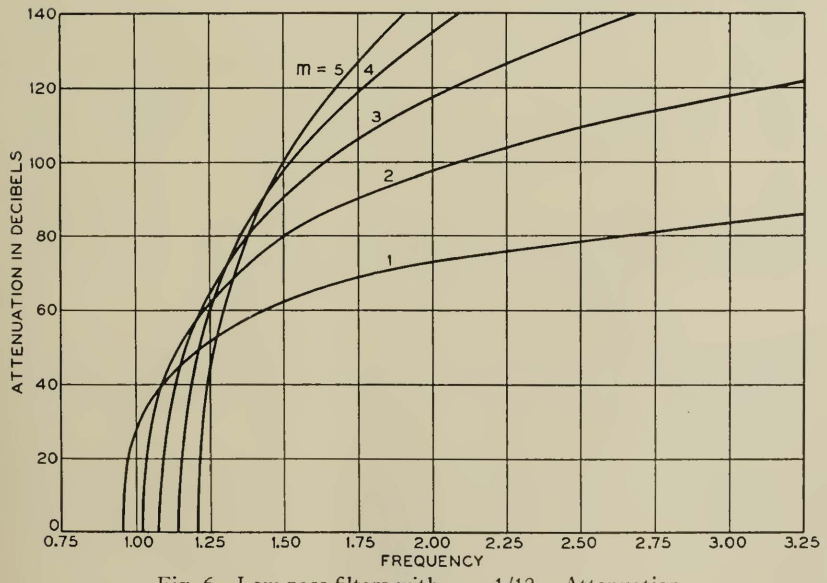


Fig. 6—Low pass filters with $\alpha = 1/12$. Attenuation.

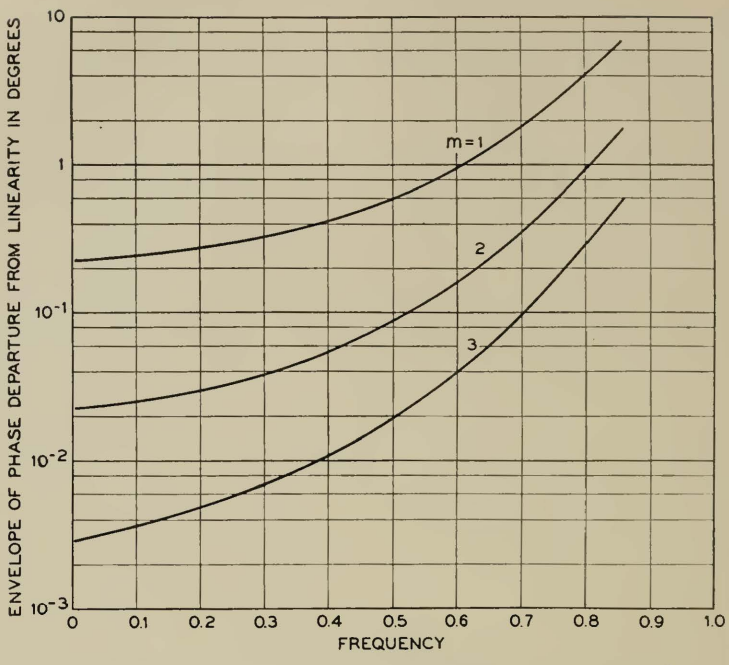


Fig. 7—Low pass filters with $\alpha = 1/6$. Envelopes of phase departures.

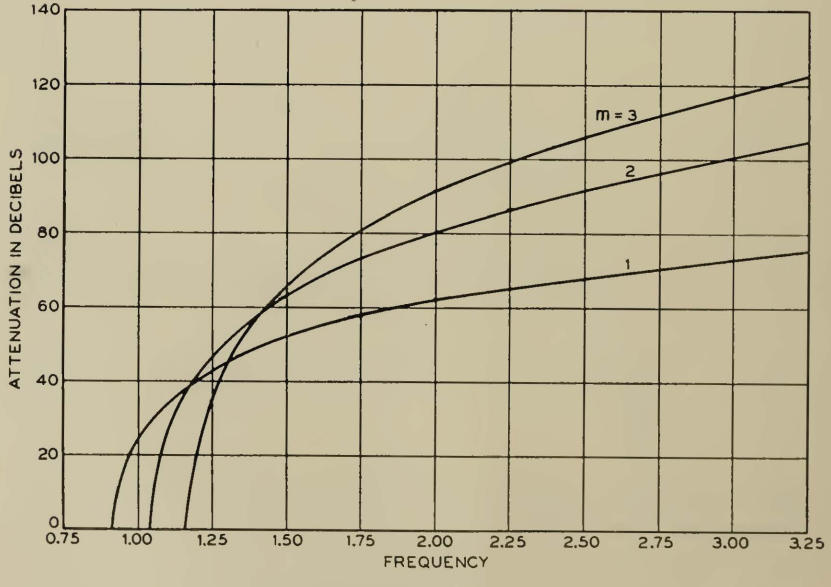


Fig. 8—Low pass filters with $\alpha = 1/6$. Attenuation.

A second set of characteristics was obtained by choosing $\alpha = 1/6$ and adding various groups of transition factors in a similar fashion. The results are shown by Figs. 7 and 8. The characteristics are drawn only for m's between 1 and 3 in this case, since with larger m's the

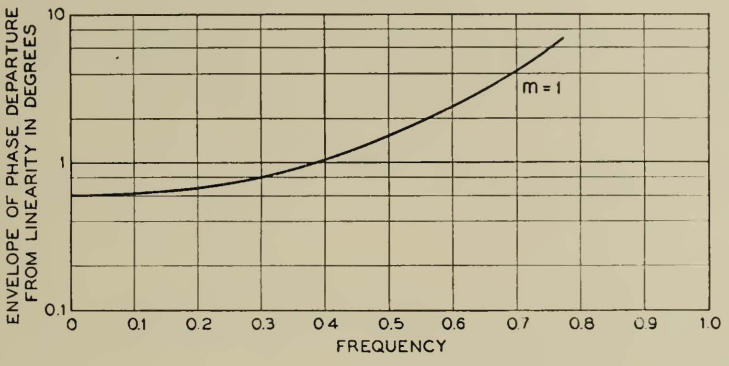


Fig. 9—Low pass filter with $\alpha = 1/4$. Envelope of phase departures.

transition interval becomes disproportionately wide in comparison with the practical transmission range. Still a third set, corresponding to $\alpha = 1/4$ and $m = 1$ is shown by Figs. 9 and 10.

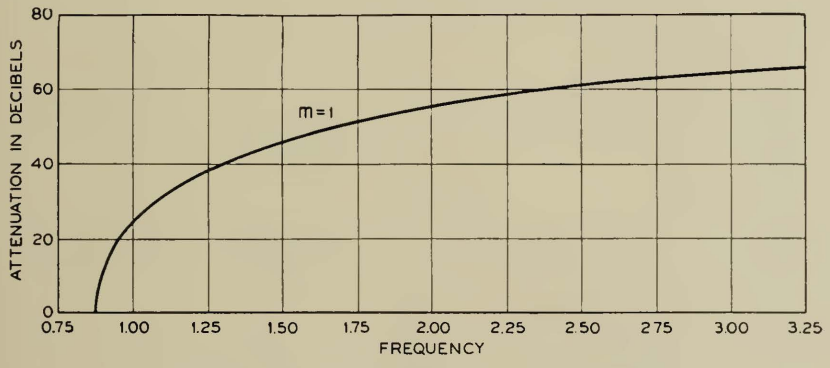


Fig. 10—Low pass filter with $\alpha = 1/4$. Attenuation.

As an illustration of the accuracy to be expected from the approximate method, a comparison between the results obtained by this method and the exact characteristics is shown in Figs. 11 and 12 for the cases $m = 1$ and $m = 2$ of Figs. 5 and 6. On the logarithmic scales used for the figures, the curves appear to be in good agreement almost up to the transition interval. The actual numerical departures in the vicinity of that interval, however, are quite large.

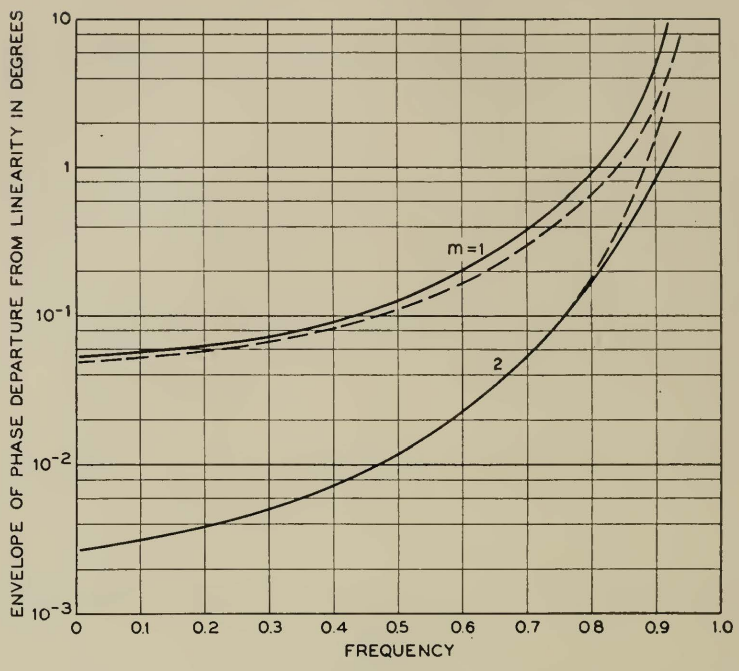


Fig. 11—Approximate and exact envelopes of phase departure for low pass filters with $\alpha = 1/12$.

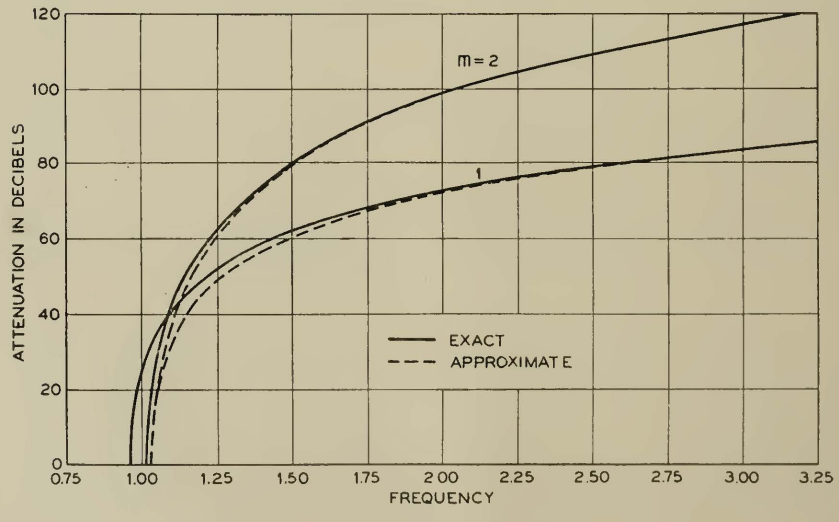


Fig. 12—Approximate and exact attenuation characteristics for low pass filters with $\alpha = 1/12$.

Image Impedance Characteristics

In virtue of the relationship previously developed between the image impedance of a given filter and the transfer constant of its complement, the curves just given might also be used to determine the impedance characteristic. However, the precision required in the approximation of Z_I to R in practical filter design is much less than that required in the approximation of $\tanh \theta/2$ to unity. A satisfactory characteristic can therefore be obtained with a much smaller number of critical frequencies. In a low-pass filter, for example, one or two impedance controlling frequencies is usually sufficient. With such a small number of critical frequencies the analytical machinery we have set up is unnecessarily cumbersome. The problem can be solved more effectively by simple cut and try methods, or by the methods advanced by Cauer [7] and Zobel.[22] For the sake of completeness, however, several illustrative characteristics are given in Fig. 13. They correspond to the choice of impedance controlling frequencies

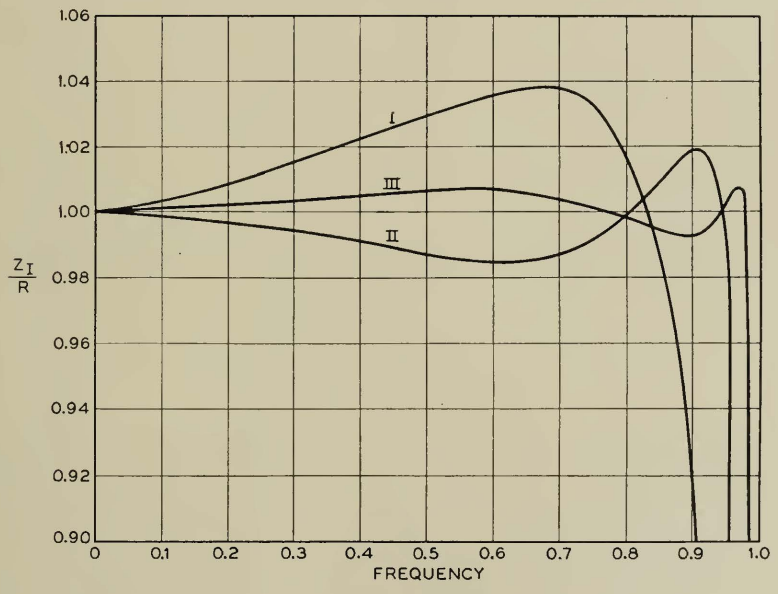


Fig. 13—Typical low pass filter image impedance characteristics.

[7] "Siebschaltungen," loc. cit.

[22] This *Journal*, Apr. 1931, p. 284. Zobel's work is not stated in terms of the lattice parameters. A simple m-type termination (of low-pass or high-pass type) can be identified with a lattice image impedance having one impedance controlling frequency while an mm'-type termination can be identified with a lattice impedance having two such frequencies. The numerical values he gives can therefore readily be adapted to the lattice design problem.

shown in Table III. An illustration of the results obtainable with the present method using a large number of critical frequencies, is furnished by Fig. 14, since the curve can evidently be interpreted as a representation of Z_I/R for a certain high-pass filter.

TABLE III

IMPEDANCE CONTROLLING FREQUENCIES CORRESPONDING TO CHARACTERISTICS OF FIG. 13

I	II	III
$1.250\, f_c$	$1.048\, f_c$ $1.448\, f_c$	$1.013\, f_c$ $1.096\, f_c$ $1.584\, f_c$

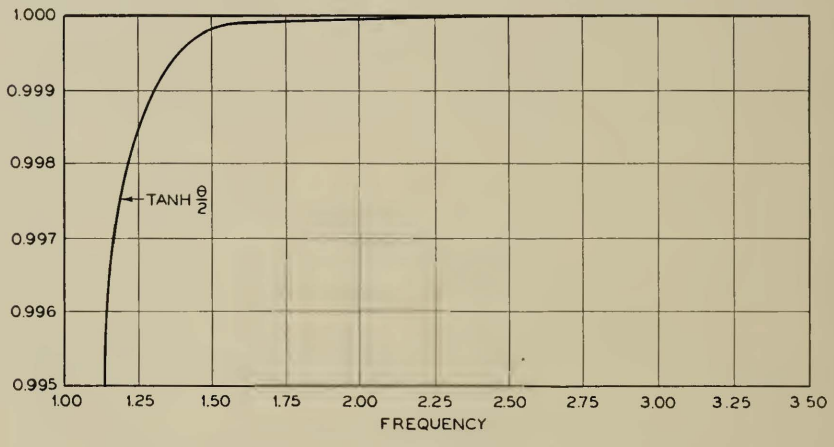


Fig. 14—Tanh $\theta/2$ for low pass filter with $m = 2$, $\alpha = 1/12$.

Weighting the Approximation

In the limiting case in which α is very small while m is fixed, the methods we have followed give the best obtainable results with respect to both attenuation and phase, for the errors in both characteristics depend upon higher order terms and become negligible as α approaches zero. In a practical design, for which α is finite, on the other hand, it will frequently be desirable to make slight adjustments in such parameters as the transition factors or the constant multiplier of the tanh $\theta/2$ expression in order to take some account of the higher order terms. The effect may be either to improve both the phase and attenuation characteristics or, more usually, to improve one at the expense of the other.

The general nature of the problem is illustrated by Fig. 14, which

represents a sketch of $\tanh \theta/2$ corresponding to the $m = 2$ curve of Fig. 6. It will be seen that the curve rises monotonically toward the line unity, at which $\theta = \infty$. What we should evidently like to obtain by slight alterations in the design parameters is a curve which rises more rapidly, or perhaps one which ripples about unity. It is also evident that the curve approximates unity so closely that even slight adjustments may produce a radical effect. To take the simplest possibility, if the constant multiplier of $\tanh \theta/2$ is slightly increased, so that the curve crosses unity at a finite frequency, the appearance of the resulting attenuation characteristic will be greatly altered. The net gain in the general level of attenuation secured, however, will be not more than 6 db. Similar remarks might be made with respect to the phase characteristic.

The relation between the phase and attenuation characteristics where such adjustments are made can be illustrated most easily by reference to the elementary half-spaced cut-off solution for the transition factors. It will be recalled that this solution was obtained by equating the coefficients of the first powers of $1/(1 - f)$ and $1/(1 + f)$ in (17) and (20). The approximation error thus depends chiefly upon the succeeding term involving $1/(1 - f)$ and $1/(1 + f)$ to the second power. A study of the expression shows that the error makes Q too small in both the transmitting and attenuating ranges. If the phase characteristic is the more important this error can be partly compensated by slightly increasing the normal half-space between the cut-off and the preceding critical frequency. On the other hand, the attenuation will be improved if the interval between the cut-off and the preceding critical frequency is decreased. To a more limited extent, both characteristics can be improved by increasing the constant factor which multiplies $\tanh \theta/2$ as a whole.

A similar study might be made of the other groups of transition factors, although the discussion would naturally become more complicated. In general it appears, as with the half-spaced cut-off solution, that the attenuation characteristic will be improved by a slight decrease in the spacings of the transition factors, while the phase characteristic will be improved if they are slightly increased. It should be remarked, however, that as the network becomes more complicated, either by a reduction in α or an increase in m, the desirable modifications in the theoretical spacings are reduced. This becomes evident if it is recalled that the transition factor spacings are proportional to α while the error is roughly proportional to α^{m+1}. It is therefore to be expected that the appropriate modifications in the spacings between transition frequencies will be of the order of magnitude of α^m times their original values.

The relationship between the phase and attenuation characteristics can be seen in another light if we observe that the improvement in attenuation which comes from the use of several transition factors is due essentially to a progressive decrease in the interval between critical frequencies as the cut-off is approached. In the final solution, for example, the intervals between critical frequencies are initially almost equal to the constant interval α. Thus in this solution, the interval between f_A and f_{A+1} is 0.992α and that between f_{A+1} and f_{A+2} is 0.945α. As the cut-off is approached, however, the interval gradually decreases to about 0.2α. In the transition interval, consequently, the phase characteristic is originally almost linear and curves upward sharply near the cut-off. Thus if the phase requirement is not severe we can consider that the first part of the transition region falls within the practical transmitting band, thereby securing a better attenuation characteristic than would be possible if the spacing of critical frequencies in the transmitting band were strictly uniform. A sketch of the phase characteristics through the transition interval

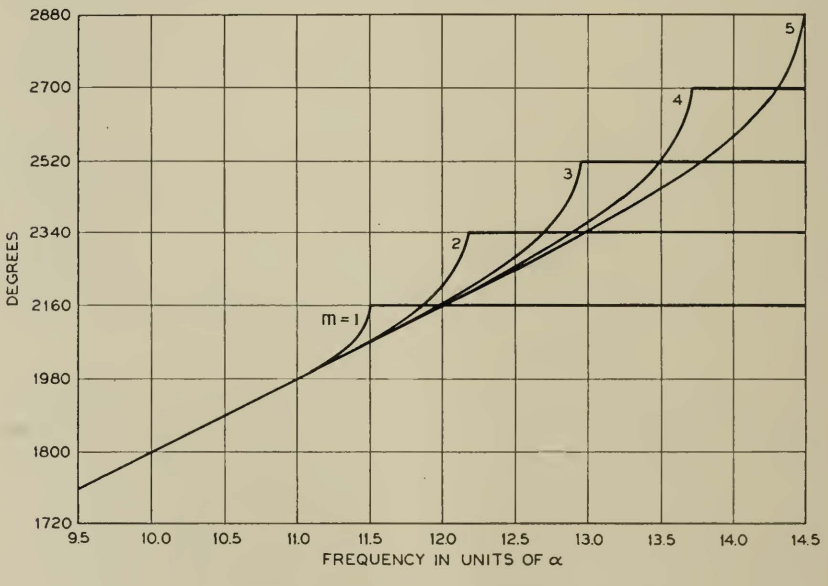


Fig. 15—Transfer constant phase shift in the transition interval; $\alpha = 1/12$.

for the networks corresponding to Figs. 5 and 6 is shown by Fig. 15. The last evenly spaced critical frequency falls at 11α.

In the extreme case when no phase requirement is imposed, it is reasonable to expect that the best attenuation characteristic will be

obtained if the progressive reduction in the spacing between critical frequencies extends over the complete transmission band, so that the phase characteristic should resemble that of familiar ladder type filter structures by becoming continually steeper as the cut-off is approached. The exact arrangement will, however, depend upon the desired type of best approximation to perfect suppression. If the approximation is to be best at frequencies most remote from the cut-off, the critical frequencies must be evenly spaced along an ordinary arc sine curve. In the Tschebycheffian type of approximations studied by Cauer, on the other hand, the spacing must be uniform along the arc of a certain sn function.

Design of a Band-Pass Filter

To illustrate the manner in which small modifications of the theoretical frequency spacings may be employed to control the relative emphasis placed on the phase and attenuation characteristics, we may consider the design of a practical band-pass filter. Suppose that the practical transmitting band is the 2,250-cycle interval between 11,375 and 13,625 c.p.s., in which the approximation of the phase characteristic to linearity is specified by the requirement that $\partial B/\partial\omega$, the so-called "delay," deviate from its average value by less than 0.1 millisecond. The transition intervals are 500 c.p.s. each, beyond which the loss is to be not less than 50 db.

The comparatively liberal tolerances suggest that the approximation furnished by $m = 1$ will be adequate. We notice that we can fit 10 uniform intervals of 250 c.p.s. between 11,250 and 13,750 c.p.s., which locates the half-spaced cut-offs at 11,125 and 13,875 c.p.s. respectively. When the characteristics corresponding to this design are checked, it is found that the phase characteristic is rather better than required, while the loss characteristic is weak.

We then turn to the solution with $m = 2$, making a compensating reduction in the number of uniform intervals. The critical frequency allocation for this case is shown in Table IV. This arrangement meets

TABLE IV

CRITICAL FREQUENCY ALLOCATION FOR LINEAR PHASE SHIFT BAND-PASS FILTER

$m = 1$	$m = 2$	Modified	Limits of Required Linear Region
11,125	11,198	11,174	
11,250	11,289	11,265	11,375
11,500	11,500	11,500	
....	....	....	
13,500	13,500	13,500	13,625
13,750	13,711	13,735	
13,875	13,802	13,826	

the loss requirement with a large margin of safety, but the phase shift curve departs seriously from linearity near the last useful frequencies, which fall in the shortened intervals of the transition factors.

With these two attempts as guides, a compromise frequency pattern which exactly suits the conditions of the problem is readily arrived at. In contrast to the transition factor spacings of 0.853α and 0.353α, as shown by the solution for $m = 2$, those actually adopted are 0.94α and 0.375α so as to make the first transition spacing more nearly uniform with those in the pass-band. The indicated frequency pattern is shown in the third column of the table.

As the values of these transition factors near the band edges are somewhat too large they lead to larger undulations of the phase characteristic in those regions than near the band center. The approximations can be rendered more uniform throughout the band without serious consequence to the loss characteristic by multiplying the tangent expression by a constant slightly smaller than unity. In this case the value chosen was $K_1 = 0.9975$.

The final "delay" and loss characteristics, corrected for the effects of dissipation, are exhibited by Figs. 16 and 17. A noteworthy result

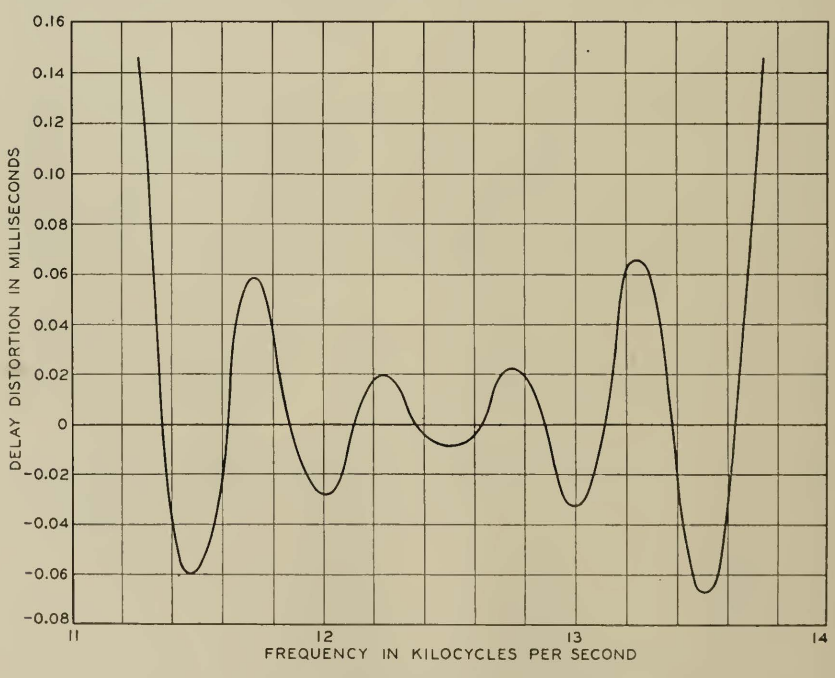


Fig. 16—Variation in the phase slope of a band pass filter.

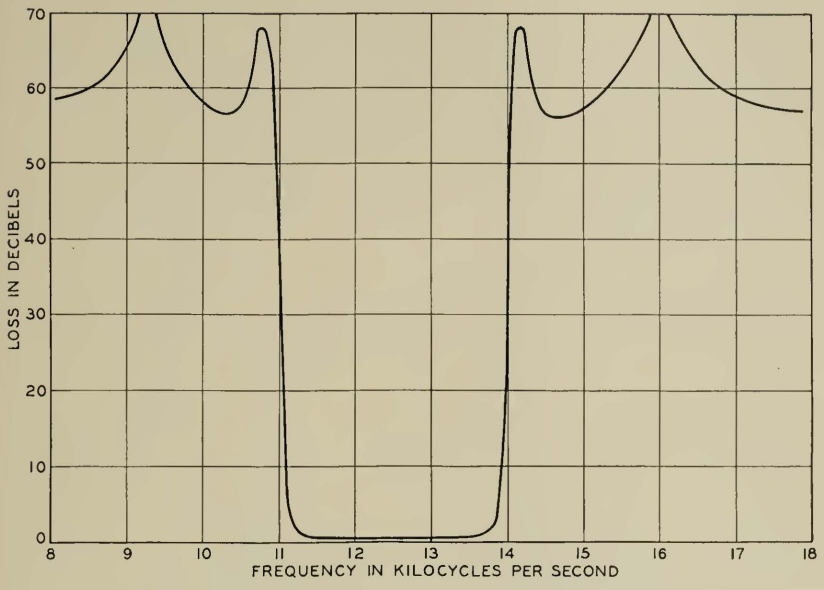


Fig. 17—Attenuation characteristic of a band pass filter.

of modifying the theoretical spacings for the transition factors has been to introduce peaks of loss near the band edges. The shortened intervals adjoining the cut-offs produce tanh $\theta/2$ curves which rise rapidly beyond these points to maxima slightly greater than unity, instead of approaching unity monotonically.

Having thus located the critical frequencies, we may readily complete the design of the filter in lattice form.

The formulation of the transfer-constant expression results in

$$\tanh \frac{\theta}{2} = K_1 \frac{\sqrt{1 - \frac{f^2}{f_a^2}} \left(1 - \frac{f^2}{f_2^2}\right) \cdots \left(1 - \frac{f^2}{f_{10}^2}\right) \sqrt{1 - \frac{f^2}{f_b^2}}}{\left(1 - \frac{f^2}{f_1^2}\right) \left(1 - \frac{f^2}{f_3^2}\right) \cdots \left(1 - \frac{f^2}{f_9^2}\right) \left(1 - \frac{f^2}{f_{11}^2}\right)},$$

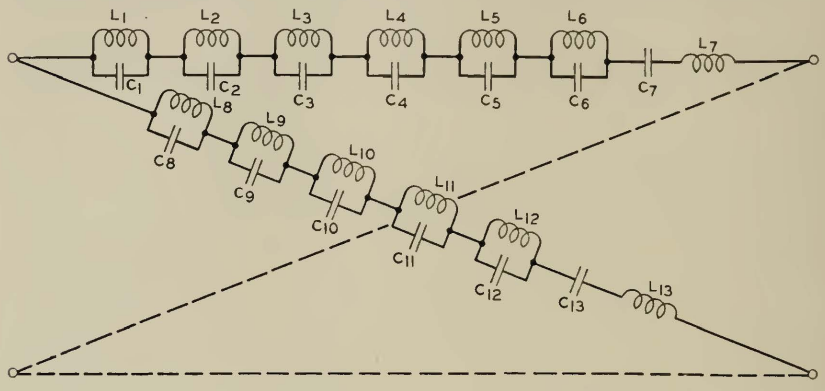
where f_a and f_b represent the cut-offs, and the other f's intervening critical frequencies in order of magnitude, as shown by Table IV.

A suitable form for the image impedance must next be obtained and since it is normally determined by requirements with which we are not now concerned, we will adopt the simplest possible expression, namely

$$Z_I = K_2 \frac{\sqrt{1 - f^2/f_a^2} \sqrt{1 - f^2/f_b^2}}{if},$$

where K_2 is determined from the condition that $Z_I = 600$ ohms when

$f = \sqrt{f_a f_b}$. The impedance functions Z_x and Z_y are now readily found by means of (1) and (2), and with the help of Foster's formula the element values can be obtained. These are shown in Fig. 18.



$L_1 = 0.0675$ mh.	$C_1 = 3.035$ mf.
$L_2 = 0.6529$ mh.	$C_2 = 0.2810$ mf.
$L_3 = 0.1958$ mh.	$C_3 = 0.8622$ mf.
$L_4 = 0.1770$ mh.	$C_4 = 0.8801$ mf.
$L_5 = 0.1330$ mh.	$C_5 = 1.085$ mf.
$L_6 = 0.0404$ mh.	$C_6 = 3.325$ mf.
$L_7 = 35.66$ mh.	$C_7 = 0.0046$ mf.
$L_8 = 0.1620$ mh.	$C_8 = 1.182$ mf.
$L_9 = 0.2021$ mh.	$C_9 = 0.8703$ mf.
$L_{10} = 1.335$ mh.	$C_{10} = 0.1214$ mf.
$L_{11} = 0.1097$ mh.	$C_{11} = 0.9112$ mf.
$L_{12} = 0.1089$ mh.	$C_{12} = 1.276$ mf.
$L_{13} = 35.86$ mh.	$C_{13} = 0.0046$ mf.

Fig. 18—Band pass filter.

This example illustrates the way in which the analysis may be applied to a typical problem in network design. The practical design would not ordinarily be complete at this point, however, since, as was mentioned previously, it is seldom desirable actually to construct the network as a single symmetrical lattice. Improved stability with respect to variations of the elements from their design values is obtained if the lattice is resolved into its components, that is, the elementary lattice sections which when operated in tandem have the same transmission properties. This question is discussed in a recent paper.[3] Furthermore, unbalanced structures equivalent to the symmetrical lattice but employing fewer elements are known,[23] and expense can usually be reduced by resorting to one of these.

[3] H. W. Bode, loc. cit. It may be interesting to observe that in the terminology of that paper the elementary constituents of linear phase shift filters are usually complex m sections.

[23] A linear phase shift lattice filter cannot, of course, be constructed as a sequence of Π or T sections, but equivalences in generalized bridged-T configurations exist. General equivalences in configurations employing ideal transformers are familiar in the literature. See, for example, Cauer, loc. cit., or Jaumann, *E. N. T.*, July, 1932, p. 243.

Part III—Filters with Linear Phase Shift Through the Cut-off

It was the conclusion of the theoretical discussion that any desired approximation to ideal filter characteristics may be obtained from a finite network, so long as a finite transition interval separates transmitting from attenuating bands. The transition interval can be taken small at pleasure, but very small transition intervals are associated with networks of many natural frequencies and numerous elements. We have already seen how considerable economies in meeting a given attenuation requirement could be obtained if the phase requirement were subordinated or removed entirely. We now consider the contrary case, in which major emphasis is placed upon the phase characteristic of the filter. Filters of this type are of practical interest in picture transmission systems since instruments used in the reproduction of images seem to be much more sensitive to the effects of phase distortion than the ear. The selectivity required from filters used in such systems is comparatively modest, but phase linearity is required not only in the practical transmission band but also through the transition interval into the region of rising attenuation.

In one important particular the present problem differs from those previously considered. In the present analysis we can no longer regard the adjustment of Z_I and the adjustment of θ as independent problems. On the contrary, in the attenuating region the contribution of θ to the phase shift is constant and we must therefore rely upon reflection effects to maintain the desired linear characteristic. Moreover, near the cut-off θ must be very carefully adjusted with respect to Z_I in order that the contributions to the phase shift from reflection and interaction effects may preserve the linearity through the transition band also. The added restrictions imposed by the extension of the phase requirement require a revision of the frequency spacings already found, and set limits upon the approximation to ideal characteristics obtainable from reactive networks of reasonable complexity.

Use of Reflection Effects to Produce Linear Phase

In the practical transmission band, Z_I can be adjusted to approximate R sufficiently closely to make reflection and interaction effects negligible. Therefore, in this range the total insertion phase is the same as the transfer constant phase, and, as before, is to be obtained from a chain of uniformly spaced critical frequencies in $\tanh \theta/2$. In the practical attenuating band, on the other hand, we find that the imaginary part of θ is either 0 or π, while interaction effects can be

ignored if we assume the loss to be reasonably high. The variation of the phase shift with frequency must therefore be attributed to reflection effects, which we can write as

$$e^{\theta_r} = \frac{\left(1 + \frac{Z_I}{R}\right)^2}{4\frac{Z_I}{R}},$$

where θ_r is the sum of the reflection effects at the two ends of the structure.

Since Z_I is reactive in the attenuating band, the angle of the denominator in this equation is $\pm \pi/2$, while that of the numerator is 2 arctan Z_I/iR. Thus

$$B_r = \mp \frac{\pi}{2} + 2 \arctan \frac{Z_I}{iR}. \tag{25}$$

Except for the constant term, which we will consider presently, this is a function of precisely the type we have been considering. Hence if the impedance controlling factors are spaced at the same uniform interval that was used in the pass-band, the phase slope will be constant and equal in both bands.

Phase Characteristics in Transition Intervals

The transition factors—or rather, factor, since clearly we have to rule out the solutions for $m > 1$—must be determined so that these linear parts of the phase characteristic are joined by a chord of the same slope. If we suppose the transition interval to be bounded by the last uniformly spaced frequencies of the transfer constant and image impedance chains, and to contain only the cut-off factor, it is easily shown that it must include a net change in phase of $3\pi/2$ radians. The interval must therefore contain 3/2 uniform spaces if the average slope is to be correct. Considerations of symmetry to be described later require that the cut-off be the center of the interval, which thus comprises two three-quarter spaces. The behavior of the several components of the total insertion phase is exhibited by Fig. 19, in which B, B_r, and B_i refer respectively to the phase shifts contributed by the transfer constant, by reflection effects, and by the interaction factor. The mutually annulling discontinuities of $\pi/2$ radians in B, and B_r at the cut-off are noteworthy.

The fact that this choice of parameters is sufficient as well as necessary to obtain the desired linearity of phase shift is not easily shown analytically. It can, however, be verified by direct computa-

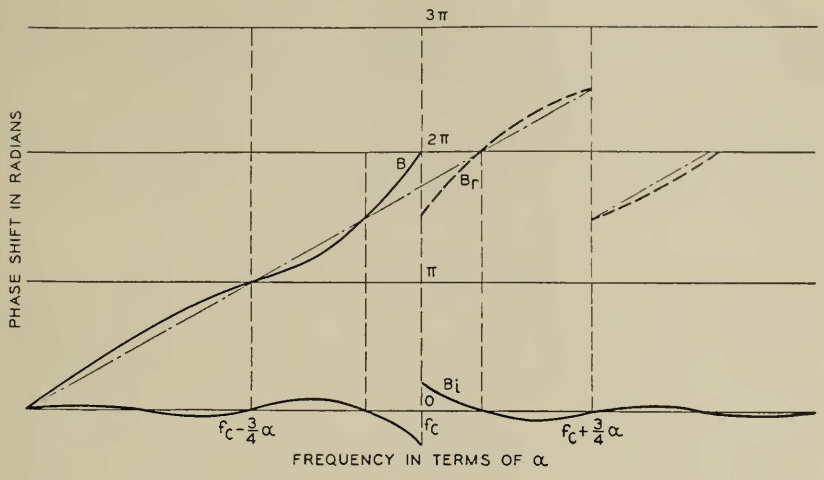


Fig. 19—Transfer, reflection, and interaction phase in the transition interval.

tion. For this purpose the customary resolution of the total insertion loss into transfer constant, reflections, and interaction is not very useful because of the indeterminacies found at the cut-off. This difficulty is avoided by expressing Z_I and θ in terms of the lattice impedances, in which event

$$e^{\gamma} = \frac{1 + \frac{Z_x Z_y}{R^2} + \frac{Z_x}{R} + \frac{Z_y}{R}}{\frac{Z_x}{R} - \frac{Z_y}{R}}, \tag{26}$$

where γ is the total insertion loss.

If iX_x and iX_y be written for Z_x and Z_y, the insertion loss and phase shift are given by

$$\tan B_{\gamma} = \frac{X_x X_y - R^2}{R(X_x + X_y)} \tag{27}$$

and

$$e^{A\gamma} = \frac{\sqrt{(R^2 + X_x^2)(R^2 + X_y^2)}}{R(X_x - X_y)}. \tag{28}$$

Equation (27) can be used to confirm our previous choice of the location of the cut-off. At this frequency one of the two reactances, X_x and X_y, will be either resonant or anti-resonant. It is evident from (27) that if the phase shift is to have the desired value, $(n + 3/4)\pi$, at the assumed cut-off the non-resonant impedance must have the magnitude R. That this value is approximated follows from the symmetrical spacing of transfer constant and image impedance

controlling frequencies with respect to the cut-off. On both sides of the transition interval, in the regions of uniform spacing of poles and zeros, the non-resonant reactance approximates R tan $\pi f/2\alpha$ or R cot $\pi f/2\alpha$ and at the middle of each space, where $\pi f/2\alpha$ is an odd multiple of $\pi/4$, is numerically equal to R. Hence, by symmetry, this must also be the value approximated at the middle of the non-uniform interval between the two chains, i.e., at the cut-off frequency.

Nature of the Approximation

The argument of Part I shows that the three-quarter spacing between the cut-off and the chain of transfer constant controlling factors results in poorer approximations to phase linearity in the transmission band and to complete suppression in the attenuating band than would the half-spaced cut-off solution. The three-quarter spacing between the cut-off and the chain of impedance controlling frequencies also leads to less perfect uniformity of the impedance characteristic. This is the price we pay for the larger range of phase linearity. Nevertheless, the error of approximation for both θ and Z_I if we follow the sense of equation (22) can be shown to be $\frac{1}{8}\alpha\left(\frac{1}{1-f}-\frac{1}{1+f}\right)$, when α is small, and hence can be made as small as we please by a suitable choice of α.[24] So far as the phase and impedance characteristics are concerned, experience shows that satisfactory precision can be obtained with a moderate value of α. The situation with respect to the attenuation characteristic is more serious. As we have already seen, the best approximation in the attenuating band is obtained by a cut-off spacing which is, if anything, slightly less than, rather than slightly greater than, $\alpha/2$. Furthermore, it appears from the above formula that with the three-quarter cut-off spacing, the approximation error at a given frequency in the attenuating band is proportional only to the first power of α. Hence cutting α in two, which substantially doubles the number of elements in the structure, adds but 6 db to the attenuation at this frequency. It is clear that a practical limit is thus set upon the suppression which can be provided.

Since the attenuation of the structure is relatively low, the contribution of reflection effects to the total loss is correspondingly important. A peak of loss occurs at each impedance controlling frequency, where the lattice impedances are zero or infinite together.

[24] It is not true that the error in $\partial\beta/\partial\omega$ vanishes with α. However, in the following example, which may be taken as typical, the variation of $\partial\beta/\partial\omega$ is still only about 1 per cent of its average value.

At these frequencies the image impedance changes sign, and therefore also the constant term of equation (25). Thus, although the phase slope is uniform throughout the attenuating range, the phase characteristic itself suffers discontinuities of π radians at each impedance controlling frequency. Whether the discontinuity is an increase or a decrease of π radians is not distinguishable for a non-dissipative network. When parasitic dissipation is taken into account the peaks are finite and the phase increases or decreases according as the line- or cross-arm of the lattice has the smaller resistance component at the peak frequency. The infinite peak at this frequency, and the associated abrupt change in phase, can evidently be restored by adding additional resistance to the smaller impedance so as to bring the arms into balance.

This observation is of importance in considering the effect of dissipation on the phase shift. A counterpart of Mayer's theorem can be found which relates the change in phase shift resulting from uniform dissipation in the network elements to the slope of the loss curve. The formula is

$$\Delta B \doteq - \omega d \frac{\partial A}{\partial \omega},$$

where d is the dissipation constant, and where A and B are in népers and radians respectively. In the transition interval, where the slope of the loss curve is great, the effect of uniform parasitic dissipation may reduce the phase appreciably. This effect can be compensated by small modifications in the theoretical frequency spacings, or by the introduction of a lumped resistance to balance the bridge at the first impedance controlling frequency, according to the plan suggested above.

Example

To illustrate the performance of this sort of network, we may consider a low-pass filter containing four evenly spaced critical frequencies in the practical transmission band. Subsequent natural frequencies will then occur at 4.75α, 5.5α, 6.5α, etc., according to the rule for three-quarter spacing adjacent to the cut-off. We may suppose that the requirement for linearity of phase shift does not extend above 7.5α, so that the sequence of uniformly spaced impedance controlling frequencies may be terminated after this point according to the scheme proposed in the case of the high-pass filter. In the frequency range of interest, we can replace the omitted chain of uniformly spaced frequencies by a single natural frequency at double spacing. The transfer constant and image impedance expressions can then be written as:

$$\tanh \frac{\theta}{2} = i \frac{\pi f}{2\alpha} \frac{\left(1 - \frac{f^2}{(2\alpha)^2}\right)\left(1 - \frac{f^2}{(4\alpha)^2}\right)}{\left(1 - \frac{f^2}{\alpha^2}\right)\left(1 - \frac{f^2}{(3\alpha)^2}\right)\sqrt{1 - \frac{f^2}{(4.75\alpha)^2}}},$$

and

$$Z_I = R \frac{\sqrt{1 - \frac{f^2}{(4.75\alpha)^2}}\left(1 - \frac{f^2}{(6.5\alpha)^2}\right)\left(1 - \frac{f^2}{(9.5\alpha)^2}\right)}{\left(1 - \frac{f^2}{(5.5\alpha)^2}\right)\left(1 - \frac{f^2}{(7.5\alpha)^2}\right)}.$$

These equations determine Z_x and Z_y, the values of which may be used in equations (27) and (28) to calculate the performance. This is shown by Fig. 20, after dissipative effects have been taken into account.

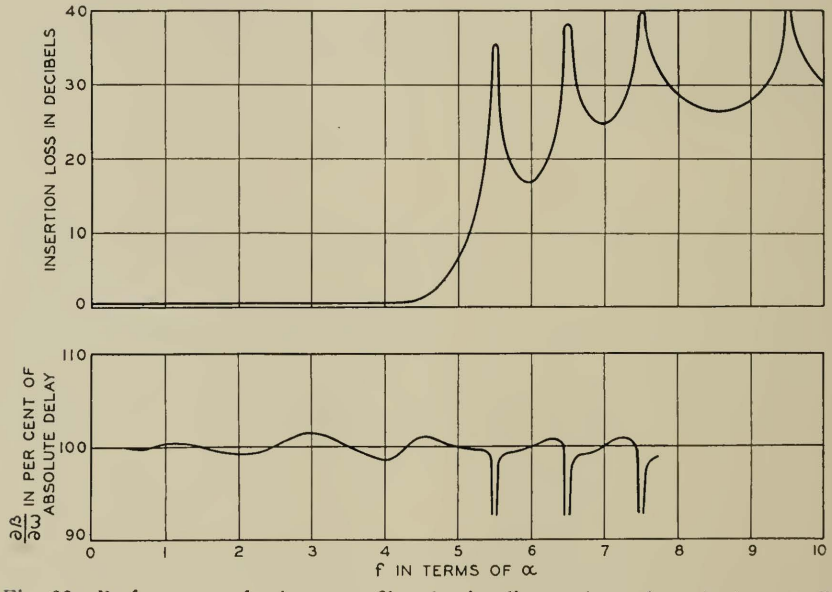


Fig. 20—Performance of a low pass filter having linear phase through the cut-off.

The approximation of the phase characteristic to linearity has again been indicated by exhibiting departures of the slope from the average.

It is observed that the approximation obtained by the three-quarter spacing is as close in the transition interval from 4α to 5.5α as in the practical transmitting band below 4α. In practical design problems, the phase shift is unlikely to be of interest beyond the first or second reflection peak, so that the chain of impedance controlling frequencies might be sooner terminated.

The loss characteristic reveals that no very high degree of suppression is attained. In fact, the loss falls to about 16 db in the trough beyond the first reflection peak. So serious a prejudice in favor of the phase characteristic would render the design unsuitable for certain engineering purposes. There are open, however, several possibilities for increasing the attenuation. Small modifications in the theoretical design parameters of the type which have been described, and in particular, slight separations of the theoretically coincident impedance controlling frequencies in the two arms of the lattice, enable the loss to be somewhat improved without much degradation of the phase characteristic. If very much higher attenuation is demanded, it can be provided by two simple structures of this type, separated by a resistance pad to preserve the reflection effects upon which the phase characteristic depends.

Further possibilities are suggested by combination of two principles already developed. It has been observed that a reduction in the three-quarter spacing of the cut-off would improve the selectivity of the structure but would also unduly increase the slope of the phase characteristic in the transition interval. We have also seen, however, that the result of uniform dissipation in the network elements is to diminish the phase shift in this region. Hence our analysis suggests that we may be able to obtain the desired phase characteristic in conjunction with the shorter cut-off spacing necessary for high selectivity if we deliberately increase the dissipation in the network.

A concomitant result of such procedure is seen to be an increase in the uniform loss in the transmission band, which may not always be desirable. Neither does the attempt to provide the phase property without sacrifice of high loss through the introduction of uniform dissipation represent the most effective attack on the problem. To achieve this end, resistances must be associated with the reactive elements of the lattice impedances in a precisely determined manner, not to be deduced solely from the foregoing theory of reactive networks. The elaboration of the theory to include also resistive impedance elements serves to determine a filter whose attenuation changes continuously from a low, uniform value in the pass-band to an arbitrary value in the attenuation bands with linearity of phase shift and, in addition, the third ideal property of constant impedance. The general theory, however, can more appropriately form the subject of a subsequent paper.

The solution of this problem completes the application of the methods for realizing ideal filter properties. We have seen that if all

the properties are of importance and the desired approximations close, we are led to networks which, while formally simple, involve correspondingly numerous natural frequencies. On the other hand, if the impedance, or the phase, or the loss property be subordinated in respect to the others, suitable modification of the analysis allows the remaining properties to be realized with simplification of the structure.

Ultra-Short-Wave Propagation: Mobile Urban Transmission Characteristics *

By C. R. BURROWS, L. E. HUNT and A. DECINO

This paper, a sequel to one entitled "Ultra-Short-Wave Propagation," deals with transmission in urban areas. The experimental data were obtained in the City of Boston on a frequency of 34.6 megacycles per second by means of a specially equipped motor car for carrying the receiver while the transmitter was fixed or vice versa.

Mass plots of these data show that the mean field strength varies inversely as the square of the path length which is the same variation as would be expected for level terrain in the absence of buildings. The same data are presented in the form of field strength contour maps.

These data are interpreted on the basis of the same physical picture which has been established for open country. The present data preclude interpretation upon the basis employed by earlier investigators of ultra-short-wave propagation through urban areas.

It is concluded that ultra-short-wave transmission in urban territory may be interpreted on the basis of transmission over level land plus the wave interference patterns caused by reflections from the buildings and an additional attenuation which on the average is independent of the length of the transmission path. Also, if the theoretical formula for the propagation of ultra-short waves over level terrain is used to calculate the received field in urban territory, and the height of the fixed antenna is measured from the local roof level instead of from the ground, these data indicate that the field strength so calculated would be near to the mean of the actual received field strengths in urban territory.

Introduction

THIS paper is a sequel to an earlier paper, "Ultra-Short-Wave Propagation," [1] which dealt mainly with transmission across open country. In the present paper the research has been extended to include transmission within a built-up region. Additional problems of transmission within urban areas that result from man-made interferences, such as the noise produced by automobile ignition systems, have been investigated.

A specially equipped motor car was used as a mobile laboratory for most of this work both because of its convenience as a means of obtaining transmission data and because of the importance of mobile communication itself.

This paper describes general characteristics and quantitative measurements of the received signal on 34.6 megacycles. Transmission

* Published in *Electrical Engineering*, January, 1935.

[1] J. C. Schelleng, C. R. Burrows and E. B. Ferrell, "Ultra-Short-Wave Propagation, *Proc. I. R. E.*, Vol. 21, pp. 427–463, March, 1933 and *Bell Sys. Tech. Jour.*, Vol. 12, pp. 125–161, April, 1933.

phenomena were studied in both directions between a fixed location on a building and the mobile laboratory.

Apparatus

Both terminals employed vertical half-wave antennas which were connected to balanced circuits by means of symmetrical two-wire transmission lines. At the fixed locations unloaded antennas were used; at the mobile terminal, in order to limit their heights to eight feet above the ground and maintain the symmetry the antennas were loaded so that their lengths were reduced to about a quarter of a wave-length (Fig. 1).



Fig. 1—Mobile receiving equipment.

The transmitter consisted of an electric oscillator employing two 75-watt tubes operating in push-pull relationship. At the fixed location, where ample power was available, the transmitter was capable of producing one ampere [2] of 100 per cent modulated carrier in its antenna without undue distortion. The mobile transmitter which used a dynamotor for tube plate supply was capable of producing the same current in its antenna. This corresponds to about six decibels less power due to the shorter antenna length.

The measuring set was of the double detection type with balanced high-frequency circuits, push-pull first detector and calibrated inter-

[2] The current was measured by a Weston type 425 thermoammeter at the current maximum.

mediate frequency attenuator.[3] This receiving equipment was calibrated in absolute units by a method described in the appendix. A mechanism for recording the field strength was attached to the measuring set: this consisted of a roll of paper that could be driven either by clock-work or by the rear wheels of the truck. The position of the recording pen was controlled by the setting of a manually operated variable attenuator. Samples of the type of record obtained are shown in Figs. 6 and 7.

Locations

The radiator for the fixed transmitter was supported by a fifty-foot pole above the roof of a seven-story building at the corner of Berkeley and Stuart Streets in the business section of Boston. The building is about 90 feet high, making the center of the antenna about 130 feet above the ground. Thus, the antenna was higher than most of the buildings of the city though it was lower than a few buildings nearby.

The antenna for the fixed receiver was supported by a 20-foot pole from the middle of the highest ridge of a gabled building making the center of the antenna about 80 feet above the street level. This building is located on the side of a slight slope in a fairly heavily wooded territory, on Seaverns Street near Center Street.

Field Strength Measurements

Transmitter at a Fixed Location

With a current of one ampere in the half-wave antenna [4] above the building at Berkeley and Stuart Streets and with the receiver in the truck, field strength measurements were made along various routes throughout Boston. These data have been averaged by one-tenth-mile intervals when the average radial distance was less than two miles and by half-mile intervals for greater average distances. A plot of these data is shown in Fig. 2. The points lie approximately on an inverse-square-of-distance line with deviations ranging up to about $\pm$ 10 db. An effort has been made to separate the points taken in the high building area. These points (shown as open circles) lie somewhat below the others with a few particularly low field strengths. The lowest field strengths of the business district were measured along the shore near Charles River Dam and near State Street on Atlantic Avenue. The field strengths in the business district would be expected to be lower because of the presence of the high buildings. The lower

[3] The set was similar to that described by Friis and Bruce, *Proc. I. R. E.*, Vol. 14, pp. 507–519, August, 1926.

[4] Since the antenna was in free space in so far as radiation resistance is concerned, this corresponds to a radiated power of 73 watts.

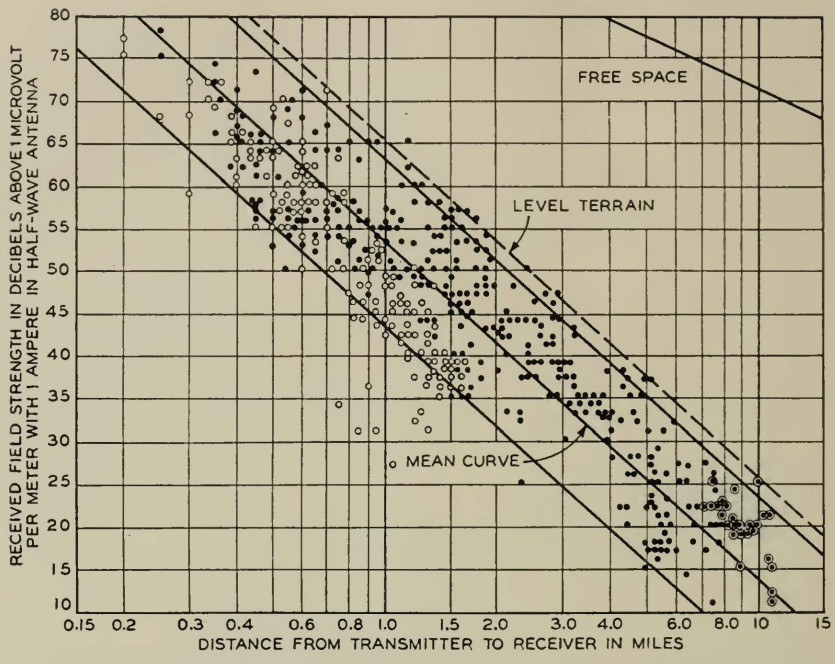


Fig. 2—Mass plot of field intensities measured at various distances from the transmitter at Berkeley and Stuart Streets in Boston. The values corresponding to distance less than two miles represent field strengths averaged over one-tenth mile intervals, while those for greater distances represent averages over one-half mile intervals. The open circles indicate fields in the high building area. Residential points outside the city limits have been enclosed in circles.

residential field strengths correspond to the region beyond Chestnut Hill.

When attempting to interpret the results of the mass plot of Fig. 2 a natural method would be to assume transmission as in free space plus an additional attenuation due to the proximity of the earth and obstacles above the earth's surface. The simpler case of transmission over level terrain in the absence of obstacles will be considered first.

It has been experimentally determined that the propagation of ultra-short waves over unobstructed paths follows the laws of optics [1, 5] so that the resultant field is composed of a well-defined reflected wave superposed upon a direct wave. Consequently for propagation over level terrain, the explanation is as follows (Fig. 3): Energy is propagated from a transmitter at A, at a height of h_1 above the ground, to a receiver at B, at a height of h_2 above the ground, both directly, as

[1] Loc. cit.

[5] C. B. Feldman, "The Optical Behavior of the Ground for Short Radio Waves," *Proc. I. R. E.*, Vol. 21, pp. 764–801, June, 1933.

represented by r_1, and by reflection at G, as represented by r_2, the distance between transmitter and receiver being represented by d. For the practical case where h_1 and h_2 are small compared with d the reflected wave impinges upon the ground at nearly grazing incidence, so that a negative reflection coefficient the magnitude of which is unity [6] for ordinary ground (not water) is obtained. This

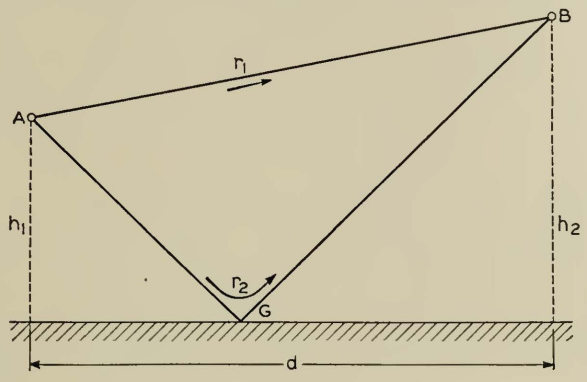


Fig. 3.

results in the field at B being the difference between two vectors of approximately equal magnitude and differing in phase by an amount corresponding to the difference in path lengths, r_2 and r_1. For the case under consideration,

$$r_2 - r_1 = 2h_1h_2/d, \tag{1}$$

and the angle between the vectors is

$$2\pi(r_2 - r_1)/\lambda = 4\pi h_1h_2/\lambda d. \tag{2}$$

[6] The magnitude is more exactly $1 - 2\epsilon(h_1 + h_2)/d\sqrt{\epsilon - 1}$ for vertical polarization and $1 - 2(h_1 + h_2)/d\sqrt{\epsilon - 1}$ for horizontal polarization, making the corresponding values for the received fields,

$$E_0\left(\frac{4\pi h_1h_2}{\lambda d}\right)\sqrt{1 + \frac{\epsilon^2(h_1 + h_2)^2\lambda^2}{(\epsilon - 1)4\pi^2h_1^2h_2^2}} \tag{a}$$

and

$$E_0\left(\frac{4\pi h_1h_2}{\lambda d}\right)\sqrt{1 + \frac{(h_1 + h_2)^2\lambda^2}{(\epsilon - 1)4\pi^2h_1^2h_2^2}} \tag{b}$$

respectively, instead of as in equation (3). When the lower of the two antennas is more than a couple of wave-lengths off the ground, the radicals are substantially unity. For the case under consideration it would be more accurate to refer to expression (*a*) as the theoretical formula but lack of knowledge of the magnitude of the dielectric constant and antenna heights that apply would introduce unnecessary uncertainty if the results were referred to this formula. It might be remarked that neglecting the presence of buildings and referring all heights to the local street level the radical represents an increase of 7–12 db for vertical polarization in the cases under consideration.

Except in the immediate proximity of the transmitter this angle is small and the resultant field is

$$E = E_0(4\pi h_1 h_2/\lambda d), \tag{3}$$

where E_0 is the free space field. Since [7]

$$E_0 = \frac{60\pi HI}{\lambda d}, \tag{4}$$

the resultant field becomes

$$E = 240\pi^2 HIh_1h_2/\lambda^2 d^2. \tag{5}$$

Equation (5) shows that the field over level terrain is inversely proportional to the square of the distance from the source.[8]

Data presented in Fig. 8 of reference 1 show that at a frequency of 69 megacycles per second the field strength variation with distance follows approximately the inverse-square relationship for a range of from two to ninety kilometers.[9] In fact, the best straight line through these data agrees with the numerical values obtained from equation (5) well within the accuracy of the experimental data. Experiments designed to test the validity of this equation now are being conducted at Deal, New Jersey and data obtained to date confirm it both as to absolute value and variation with terminal heights and wave-length for horizontal polarization within the range $2 < h_1 < 25$, $2 < h_2 < 25$, $2 < \lambda < 17$, $d = 9{,}420$ and $26{,}300$, all measured in meters. The experimental confirmation of this formula for these distances indicates that the effect of the earth's curvature is secondary to the negative reflection effect upon which this formula is based. This might be expected in view of the fact that both diffraction and refraction tend to mitigate the additional attenuation that would be caused by reflection from a plane tangent to the earth's surface at the point of geometric reflection.

[7] If I is in amperes, d in meters, and H the effective height of the antenna, and λ the wave-length in the same units, E_0 is given in volts per meter.

[8] Since distance appears in this equation only as a factor and not as an exponent, the reduction with distance of the field strength of ultra-short waves over level terrain is independent of wave-length, polarization, dielectric constant, etc., as all of these quantities cancel in the ratio of the field strength at one point to that at another. The absolute magnitude of the field strength is proportional to the frequency for the same radiated power and antenna heights. If the antenna heights are sufficiently low, the field is also dependent upon the polarization and ground constants as indicated by expressions (*a*) and (*b*) of footnote 6.

[9] While undoubtedly at the greater distances the field suffers additional attenuation above that shown by equation (5) due to the curvature of the earth, such additional attenuation evidently takes place at distances beyond those employed in any of the authors' experiments.

When the propagation is through built-up areas instead of over level terrain the condition is more complicated. Even here, however, for terminals well above the tops of buildings, theoretical considerations [1] indicate that the same explanation of direct and reflected waves is valid. Data presented by Jones [10] may be used as a verification for this explanation even for transmission over buildings. Fig. 11 of his paper shows that for heights between 170 and 1,500 feet the field is proportional to the height in accordance with equation (5). When the terminals are lowered within the building region, the field should decrease more rapidly than proportionally with the height above the ground. In fact, data presented in Fig. 4 indicate that with one

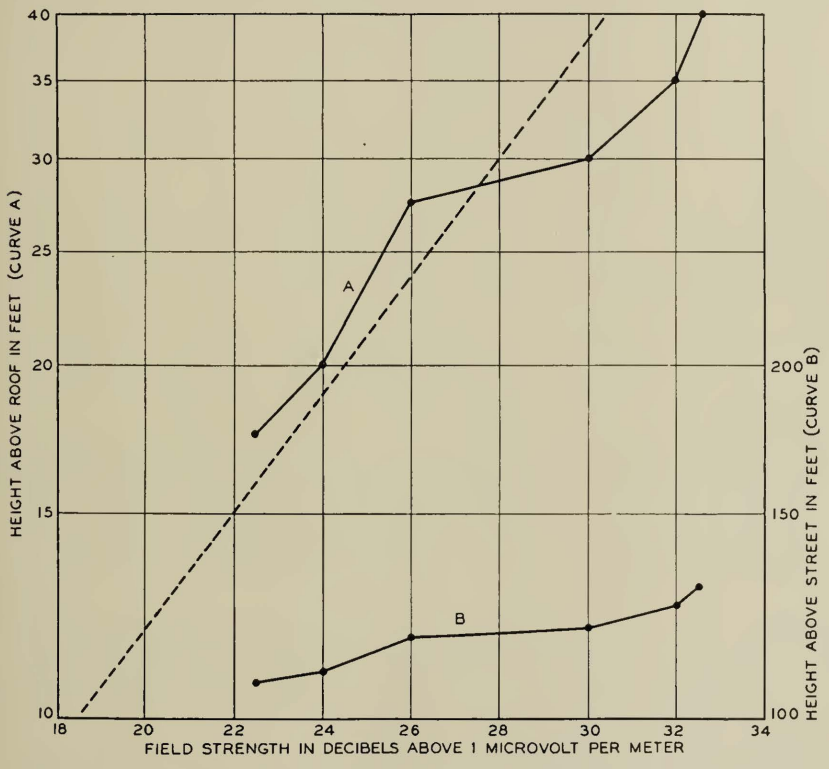


Fig. 4—Variation of field received at Berkeley and Stuart Streets with antenna height. Curve *A* shows the variation with height above the roof while Curve *B* shows the variation with height above the ground. The slope of the broken line indicates a linear relationship between field and height.

[1] Loc. cit.

[10] L. F. Jones, "A Study of the Propagation of Wave-lengths between Three and Eight Meters," *Proc. I. R. E.*, Vol. 21, pp. 439–486, March, 1933.

terminal above a flat roof that was approximately the same height as other nearby flat roofed buildings, the field strength is more nearly proportional to the height above the roof than to the height above the ground. When the terminals are lowered below the average building height additional complications are introduced. While this is somewhat difficult to picture because of the irregularity of the surface bounding the transmitting medium [11] the main outline seems simple enough. Above the building level there is a tendency for the field to follow the simple rules that hold for transmission over level country. In general the field strength actually received in the street would be proportional to the field strength overhead but of smaller amplitude since it is a product of scattering. This does not imply that the street signal comes down vertically; it probably is the result of scattering from points lying in a fairly large zone about the receiver and consists of a multiplicity of signals traveling in inclined directions.

Returning now to the present data on the propagation of ultra-short waves through urban areas, Fig. 2 shows [12] that the field strength is in general inversely proportional to the square of the distance from the transmitter. The mean curve through the data is 12 db below the curve for level terrain free from obstacles, plotted from equation (5) above, indicating the additional attenuation due to man-made structures. An analysis of the individual points shows that the reduction in field due to the obstacles (i.e., in addition to the level terrain attenuation) is independent of the distance so that there is no *absorption* due to the buildings in the usual meaning of the word; otherwise the additional attenuation would increase with the distance.

This method of interpretation is radically different from that of investigators of the propagation of ultra-short waves through urban areas whose papers have come to the attention of the authors.[13, 14, 15] They have assumed that the transmission occurs as in free space except for an additional attenuation *through the absorbing layer of buildings*. Such an assumption that the propagation of ultra-short waves is unaffected by the presence of the ground except in so far as the waves penetrate the absorbing layer of buildings, appears to be

[11] This surface is, of course, that formed by the ground and the walls and tops of buildings.

[12] It will be shown later that the empirical formula assumed by Schröter, Sohnemann, Jones, and Muyskens and Kraus cannot be made to fit these data.

[13] F. Schröter, "Zur Frage des Ultrakurzwellen-Runkfunks," *E. N. T.*, Vol. 8, pp. 431–436, October, 1931.

[14] K. Sohnemann, "Feldstarkemessungen im Ultrakurzwellengebiet," *E. N. T.*, Vol. 8, pp. 462–467, October, 1931.

[15] Henry Muyskens and John D. Kraus, "Some Characteristics of Ultra-High-Frequency Transmission," *Proc. I. R. E.*, Vol. 21, pp. 1302–1316, September, 1933.

inconsistent with the physical picture [1, 5, 16, 17] of ultra-short wave propagation which has been confirmed by basic experimental data.

Since the inverse-square-of-distance relationship (equation 5) which results from this physical picture is so different from the exponential relationship which results from the *absorption* assumption of previous investigators, the question arises as to the possibility of reinterpreting their data on the basis of this physical picture. The data presented by Muyskens and Kraus as Fig. 2 of reference 15 has been replotted in Fig. 5 of the present paper on logarithmic coordinates in order to

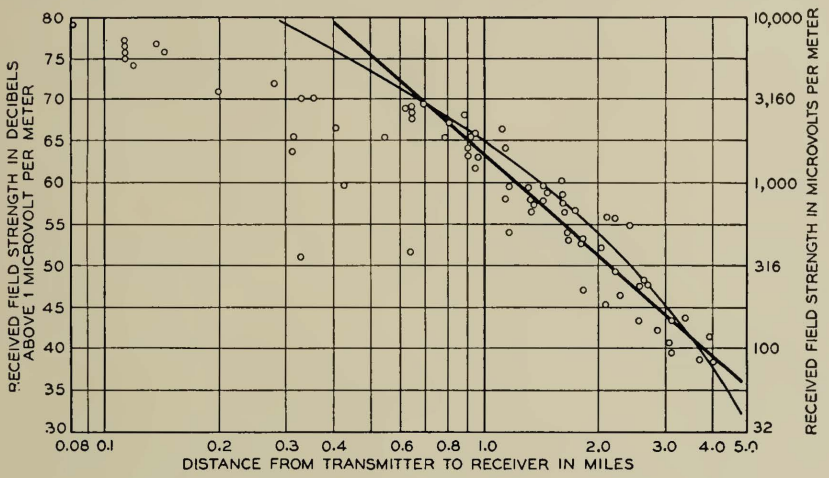


Fig. 5—Attenuation curve for 5 meter transmission as replotted from paper by Muyskens and Kraus, *Proc. I. R. E.*, Vol. 21, p. 1306, Sept., 1933. The straight heavy line shows an inverse-square of distance variation in accordance with the physical picture. The thin curved line is a replot of the curve presented by Muyskens and Kraus as representing these data by a variation according to an inverse-distance times an exponential factor. It is evident that these data may be interpreted equally well on the basis of the physical picture (heavy curve) as on the basis of the empirical equation assumed by Muyskens and Kraus.

facilitate reinterpretation on the basis of this physical picture. Figure 5 shows that it is possible to interpret their data as following an inverse-square-of-distance law equally as well as an inverse-distance law times an exponential factor.[18] In this interpretation little weight

[16] Bertram Trevor and P. S. Carter, "Notes on Propagation of Waves Below Ten Meters in Length," *Proc. I. R. E.*, Vol. 21, pp. 387–426, Mar., 1933.

[17] Carl R. Englund, Arthur B. Crawford and William W. Mumford, "Some Results of a Study of Ultra-Short-Wave Transmission Phenomena," *Proc. I. R. E.*, Vol. 21, pp. 464–492, March, 1933 and *Bell Sys. Tech. Jour.*, Vol. 12, pp. 197–227, April, 1933.

[18] Comparison of Figs. 2 and 5 on an absolute basis is difficult. The lower density of the buildings in Ann Arbor and the fact that the fixed antenna was located in as open a space as possible at the corner of the roof, combine to reduce the effect of the buildings in lowering the mean field strengths. Also the apparent absence of a measurement of the radiated power in the Ann Arbor experiments precludes the determination of the absolute value of the attenuation from the experimental data.

has been given to the points which are well below the curve, in accordance with the view of the experimenters that these points represent particularly unfavorable receiving locations.

Trevor and Carter [16] have made a similar interpretation of the data presented by Jones [10] which shows that for the larger distances the field strength was inversely proportional to the square of the distance. If this inverse-square-of-distance curve were extended to shorter distances it would be found that most of the nearby points would lie somewhat below it. This is presumably because of the lack of favorable receiving locations in the high-building area. While the empirical formula arrived at by Jones may represent his data satisfactorily, the physical picture assumed of a free space field times an absorption factor is untenable since it requires a radiated power approximately 20 db below that measured. Undoubtedly, the power radiated is not in error by this amount, since Trevor and Carter obtained a satisfactory numerical check on the basis of the other picture by using the value of power radiated as given by Jones.

It is possible, of course, to represent any data by an inverse-distance factor times an exponential factor for a limited range of distances. An attempt to do this with the data of Fig. 2 by making the empirical curve agree with the experimental inverse-square-of-distance curve at 1 and 4 miles results in a curve that agrees well with the data between 0.6 and 5.0 miles but is 11 and 22 db low at 0.2 and 12.0 miles, respectively. Even if these rather large discrepancies at the limits of the curve were neglected it would still be impossible to interpret the data in terms of the free space field times an exponential absorption factor, because of the fact that the empirical curve so determined requires a radiated power 35 db below that measured; this is untenable since the over-all uncertainty in the absolute value of the measurements is only a few decibels.

It should be pointed out that each point of Fig. 2 represents the average field over an interval of either a tenth or a half mile depending upon whether the transmission path involved was less or greater than two miles. Within each interval the field varied by five to fifteen decibels because of the local wave interference pattern, as is shown by the samples of the graphs taken with the recorder which are presented in Figs. 6 and 7. Fig. 6 is an example of the record taken in the business district of Boston at a distance of about one and a half miles from the transmitter near the region *A* shown in Fig. 8. The maxima and minima are spaced very closely and differ by ten to fifteen decibels. This was characteristic of the type of record obtained at the shorter distances. At the greater distances the magnitude of the local variations was less, as illustrated by Fig. 7, which is a sample of the record

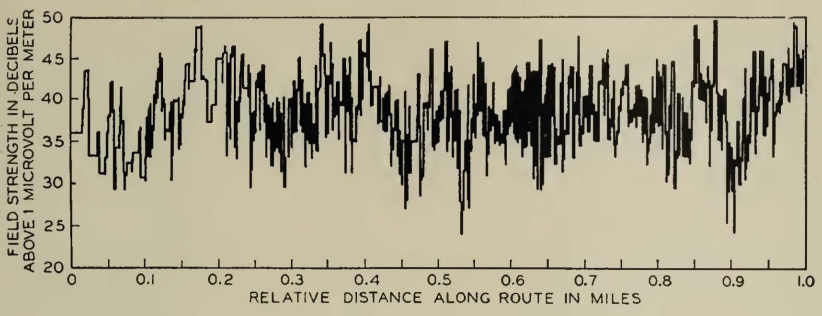


Fig. 6—Portion of record showing the large field strength variations as recorded while driving through the business district of Boston at a distance of about 1.5 miles from the transmitter.

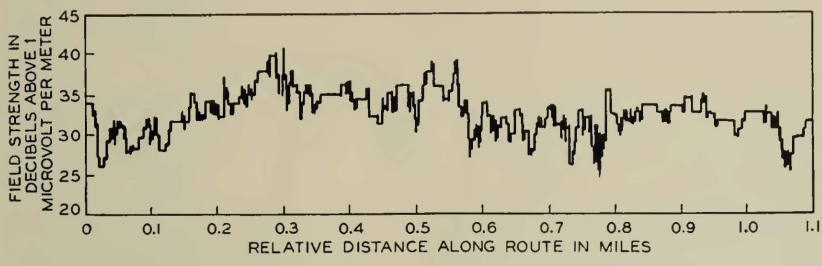


Fig. 7—Portion of record showing the small variations of field strength while driving through the residential section of Boston at a distance of about 5 miles from the transmitter.

taken at a distance of five miles (near *B* in Fig. 8). The change in the magnitude of the variations might have resulted from the fact that all of the data for the greater distances were taken in residential districts with correspondingly lower heights and densities of buildings.

An idea of the variations to be expected from the inverse-square-of-distance relationship is shown by the contour map of Fig. 8. The data already presented will give an idea of the impossibility of showing much detail in such a map. Likewise it would be an almost endless job to make measurements on every street in a city of this size. While data were obtained within reasonable intervals over the area for which solid contours are drawn (continuous field strength records taken over 143 miles of street are represented by this figure), another set of data might result in a somewhat different looking map. The broken contours are not based upon field strength measurements, but are merely a plausible way of joining the solid contours to aid the eye of the reader. The two 20 db contours in the lower part of the figure have not been joined because of insufficient data. The low field strengths along the Neponset River may be a result of local conditions and possibly the 20 db contours should be continued to the right along

an arc of more nearly constant radius. In this connection it may be mentioned that the field is nearly constant within the bulge of this contour to the southwest.

A striking fact brought out by the map is the crowding of the contours in the business district. There are particular directions for which the attenuation is greater presumably due to the combined effects of high buildings, for example to the east-northeast. There are other directions where the field strength is higher than the average. Several such places were noted when salt water extended immediately in front of the measurement location in the direction of the transmitter. The closed contours in the Mystic River to the north illustrate the better reception over salt water as predicted by theory.[1] An example of the records taken over bridges upon which these contours are based is shown in Fig. 9. (The route over which these data were taken is

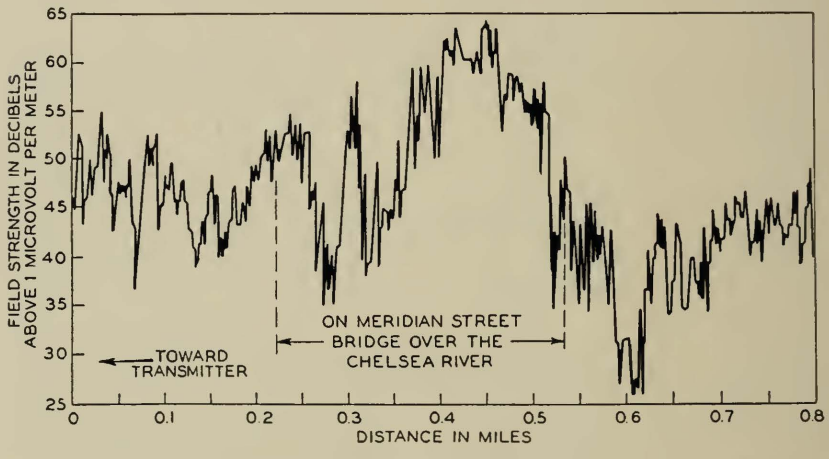


Fig. 9—Portion of records showing the variation of field strength on going over water.

indicated at *C* of Fig. 8.) The average increase in field when going over bridges was 10 db. This may be explained by the better conductivity of the water which results in a receiving directive characteristic that is more favorable to low angle reception. In some cases, the absence of buildings and increase in the height of the antenna due to the elevation of the bridge undoubtedly contributed to the greater field.

Another fact that is illustrated by the map is that the effects of obstacles and type of terrain are in general local. For example, the higher field strengths beyond salt water are soon reduced to normal with the intravention of additional land, illustrated at *D* in Fig. 8.

It was found that the field strengths along Columbus Avenue (indicated by *E* on Fig. 8) were about 10 db higher than those obtained on either side. This increase probably resulted from the fact that a


Fresh Pond
45
BOSTON
RIVER
Creek



Fig. 8—Field strength contour map of Boston, Mass. Transmitter located at Berkeley and Stuart Streets as shown by the solid circle. Frequency = 34.6 mc. Field strength in decibels above one microvolt per meter with one ampere in a half wave antenna.

more or less unobstructed optical path existed along Columbus Avenue.

While a detailed analysis of the attenuation of ultra-short waves over paths as complicated as those considered in this paper is impossible, the known facts indicate several general characteristics that seem worthy of mention for further experimental investigation. The fact that the field was found to be approximately proportional to the height of the antenna above the roof level of the surrounding buildings (in Fig. 4), rather than to the height above the ground, as in the case of transmission over level terrain free from buildings, confirms the expectation that the "ground" conditions in the immediate vicinity of the fixed terminal would play an important part in determining the magnitude of the received field strength.

It is perhaps correct to assume that for the fixed terminal the height to be substituted in equation (5) should be the height above the roof rather than the height above ground. This would reduce the "level terrain" curve of Fig. 2 by 10 db.

Figures 6, 7 and 9 show marked wave interference patterns which indicate reflections from a multiplicity of points in the immediate vicinity of the mobile terminal. Besides these variations in the magnitudes of the fields observed at points in close proximity to each other, which are undoubtedly caused by reflections from irregularities in the immediate vicinity of the terminal, there are the variations represented by the spread of the points about the mean curve (Fig. 2). That these variations may be attributable to conditions local to the terminal is indicated by the fact that the increase in field on the far side of salt water and the decrease in the field on the far side of hills, etc., do not persist at further distances. Even if the irregularities of the contours of Fig. 8 were removed the contours would not be concentric circles about the transmitter. At this stage in the development it would be unwise to attempt to say how much of the deviation of the contours from circles may be attributable to directional characteristics at the fixed terminal and how much may be due to the intervening terrain. Statistically speaking, however, it is safe to say that the additional attenuation attributable to deviations from level terrain such as is produced by the presence of buildings, is independent of the length of the transmission path, so that there is no *absorption* in the usual meaning of the word. Another experimental result that points to the possibility that the effect of the buildings is local may be deduced from the data presented by Jones [10] and by Trevor and Carter.[16] The latter showed that the more distant points lie on the inverse-square-of-distance curve expected for transmission over level terrain free from buildings. The nearby points, however, lie below this curve, indicating that the major effect of the buildings is a local one.

more or less unobstructed optical path existed along Columbus Avenue.

While a detailed analysis of the attenuation of ultra-short waves over paths as complicated as those considered in this paper is impossible, the known facts indicate several general characteristics that seem worthy of mention for further experimental investigation. The fact that the field was found to be approximately proportional to the height of the antenna above the roof level of the surrounding buildings (in Fig. 4), rather than to the height above the ground, as in the case of transmission over level terrain free from buildings, confirms the expectation that the "ground" conditions in the immediate vicinity of the fixed terminal would play an important part in determining the magnitude of the received field strength.

It is perhaps correct to assume that for the fixed terminal the height to be substituted in equation (5) should be the height above the roof rather than the height above ground. This would reduce the "level terrain" curve of Fig. 2 by 10 db.

Figures 6, 7 and 9 show marked wave interference patterns which indicate reflections from a multiplicity of points in the immediate vicinity of the mobile terminal. Besides these variations in the magnitudes of the fields observed at points in close proximity to each other, which are undoubtedly caused by reflections from irregularities in the immediate vicinity of the terminal, there are the variations represented by the spread of the points about the mean curve (Fig. 2). That these variations may be attributable to conditions local to the terminal is indicated by the fact that the increase in field on the far side of salt water and the decrease in the field on the far side of hills, etc., do not persist at further distances. Even if the irregularities of the contours of Fig. 8 were removed the contours would not be concentric circles about the transmitter. At this stage in the development it would be unwise to attempt to say how much of the deviation of the contours from circles may be attributable to directional characteristics at the fixed terminal and how much may be due to the intervening terrain. Statistically speaking, however, it is safe to say that the additional attenuation attributable to deviations from level terrain such as is produced by the presence of buildings, is independent of the length of the transmission path, so that there is no *absorption* in the usual meaning of the word. Another experimental result that points to the possibility that the effect of the buildings is local may be deduced from the data presented by Jones [10] and by Trevor and Carter.[16] The latter showed that the more distant points lie on the inverse-square-of-distance curve expected for transmission over level terrain free from buildings. The nearby points, however, lie below this curve, indicating that the major effect of the buildings is a local one.

The greatest difficulty encountered in attempting to apply the results of these experiments to the pre-determination of the field to be expected for transmission in other cities would be the interpretation of the method of assigning values to the heights in equation (5). While sufficient data are not available to establish an empirical relation, for a first estimation it seems reasonable that if the height of the fixed antenna were measured from the general roof level of the surrounding buildings and the height of the mobile antenna were measured from the street level, the resulting field strengths would lie within the range of expected values.[19]

Field Strength Measurements

Receiver at a Fixed Location

The field strength data obtained with the receiver at Seaverns Street and the transmitter in the truck are shown in Fig. 10. These

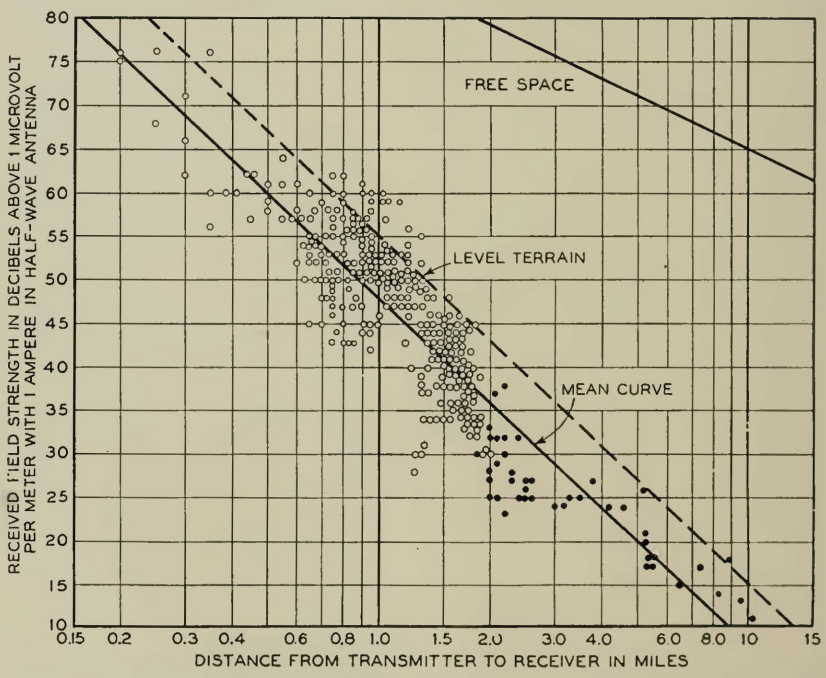


Fig. 10—Mass plot of field intensities measured with the transmitter at various distances about the Seaverns Street location. The open circles represent fields averaged over one-tenth mile intervals while the solid circles represent averages over one-half mile intervals.

[19] This should not be expected to be true for antennas close to the roof, since the field-strength obviously would not go to zero when the antenna height goes through the roof level. Extreme caution should be used in applying these results to cities in which the density and uniformity of buildings differ from those of Boston. When the topography departs greatly from that of level terrain, it would be difficult to infer the transmission conditions from the data here presented.

data may be represented also by an inverse-square-of-distance curve. By comparison with Fig. 2 no differences that can be attributed to the change in position of the fixed terminal nor the direction of transmission is evident. There is a small difference in the separation between the mean curves and the "level terrain" curves, but in both instances the mean curve lies very close to the level terrain curve (not shown) that would result from measuring the antenna height of the fixed terminal from the average roof level instead of from the ground.

Most of the nearby points were taken in the park system. They indicate that it is not more difficult to transmit through wooded areas than through built-up sections.

A field strength contour map illustrating the results obtained with the receiver fixed at Seaverns Street is shown in Fig. 11. The disturbing effect of hills is illustrated at several points on the map. There was a reduction of field when the transmitter was behind either Bussey Hill at *A* on the map or Green Hill at *B*, while the field was higher when the transmitter was between them. There was a rather deep minimum when the transmitter was immediately behind Parker Hill at *C* to the north, but half a mile farther away there was no noticeable effect.

Effect of Obstacles

In the course of the measurements some qualitative observations were made which will be summarized in this section. It was noticed in particular that when the receiver passed underneath an intersection of overhead trolley wires, the field was somewhat reduced. An example of this effect was observed at the intersection of Massachusetts and Huntington Avenues where the reduction was 15 db. At this point the maze of overhead trolley and support wires apparently constituted a fairly efficient screen for these waves. Observations made during the tests showed that sometimes the field was considerably reduced on the far side of hills as has been brought out by the contour maps. A large reduction in field strength resulted upon going behind a low hill on Saratoga Street on Breeds Island. The most striking example of this effect occurred with the measuring set at Seaverns Street and the mobile transmitter being driven from Huntington Avenue onto South Huntington Avenue (*C* of Fig. 11). Soon after rounding the corner at the foot of Parker Hill, which is 200 feet high, the average field dropped about 15 db.

The field was 20 db lower under Funeral Bridge than on either side of it. This is a stone and earth bridge appearing as a short tunnel to the road beneath it. The field was usually reduced upon passing underneath bridges of this general type of construction.

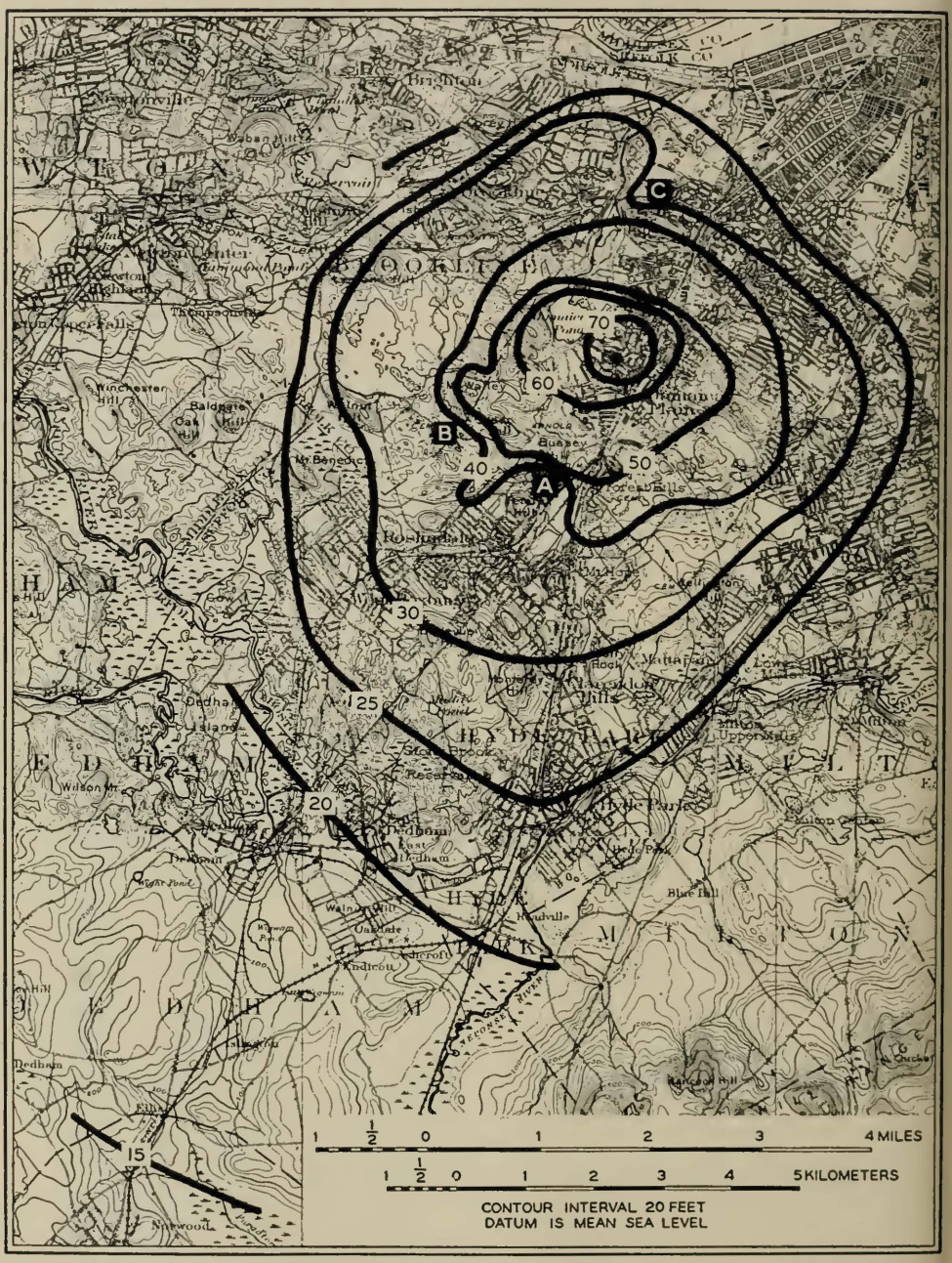


Fig. 11—Field strength contour map of Boston, Mass. Receiver located on Seaverns Street nea Center Street as shown by the solid circle. Frequency = 34.6 mc. Field strength in decibels abov one microvolt per meter.

A separation of the field strengths into those obtaining in the high-building area (indicated by open circles in Fig. 2) from those in the lower-building area (indicated by solid dots) shows that the attenuation is somewhat greater in the former.

No effect of the elevated railway structures on the average field was observed.

Noise Measurements

For reception in the car, by far the greatest interference is that caused by the electrical systems of passing automobiles. Special tests to determine whether or not street cars produced any noise gave a negative result. That is, under conditions where automobile noise was a limitation to reception, trolley noise was unaudible. While no special tests were made of the noise from elevated trains, at no time was it found objectionable.

With the receiver on top of the building at Berkeley and Stuart Streets, the predominating interfering noise was caused by an electrical substation next door. When the antenna was lowered approximately to the roof level the noise from the elevator motors and switching equipment in the pent house near by was well above any other noise. Upon raising the antenna to its proper position the elevator noise was reduced to a negligible amount compared with the power station noise, because of the combined effect of the directivity of the antenna and increased distance. The resulting noise was of approximately the same magnitude, indicating that the elevator switching noise was reduced by a fairly large factor in raising the antenna. This fact has an important bearing on reception of signals on the roofs of office buildings, since elevator switching noise is in general the limiting factor. Occasionally an automobile, started in the street below, would produce measurable interference. At Seaverns Street, however, the most objectionable noise was caused by the ignition systems of automobiles which were accelerating in low gear in the street below.

Acknowledgment

The authors wish to acknowledge the coöperation of the New England Telephone and Telegraph Company and the Graybar Electric Company, and also Commissioner E. A. Hultman and Signal Director T. A. J. Hayes, both of the Boston Police Department.

APPENDIX

Methods of Calibrating

In order to obtain an absolute calibration of the measuring equipment, the field strength at a distant point in space, at the same height

above the ground as the center of the receiving antenna, was obtained by a standard method of field strength measurement.[20] This equipment then was removed and the truck placed in such a way that its antenna was always at the point where the field strength had been determined. By pivoting the truck about this point the horizontal polar diagram of the mobile receiving equipment was obtained. It is shown in Fig. 12. As the field was known at this point in the absence

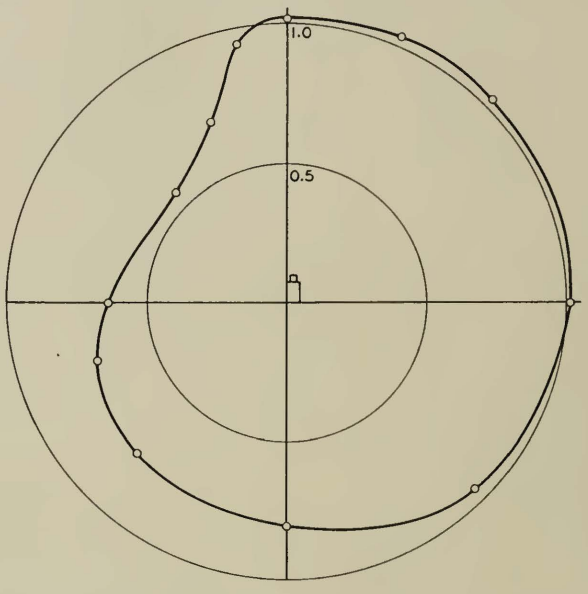


Fig. 12—Directional characteristic in the horizontal plane of the mobile receiver.

of the truck the attenuator settings gave a calibration of the receiving equipment for all directions. With this calibration as a standard, a comparison field generator which employed a loop radiator attached to the opposite side of the truck was calibrated also. The comparison field generator then was used throughout the test to check the calibration of the receiver. Since the polar diagram was fairly constant on the right side of the truck, care always was taken to orient the truck in such a way that the direction of the transmitter was at the right side rather than the left.

[20] This method consists of comparing the unknown field with a known field produced by a "Standard Field Generator," which is a small compact self-contained oscillator. It is very carefully shielded except for a small balanced loop extending in a vertical plane above the shield. A thermomilliammeter is located in the loop at the point of low potential with respect to the shield. From the reading of the meter and the dimensions of the loop, the field at nearby points may be computed. See *Proc. I. R. E.*, Vol. 21, pp. 430–431, March, 1933.

When the receiver was at Berkeley and Stuart Streets it was possible to employ the usual method of calibration [18] because the roof was flat. With the receiver at the Seaverns Street location, however, the gabled roof made it impractical to support the standard field generator opposite the midpoint of the antenna. In the latter case, accordingly, the constancy of the gain of the receiving equipment was depended upon in reducing the measurements to field strengths in absolute units. This lack of calibration did not introduce a large uncertainty, since the receiving equipment has been used over a period of years during which time its gain has remained constant within a few decibels.

Two-Way Tests

At the conclusion of this survey, actual two-way tests were made between a cruising car and fixed locations. For this purpose a car was equipped by E. B. Ferrell and R. C. Shaw [21] with an ultra-short-

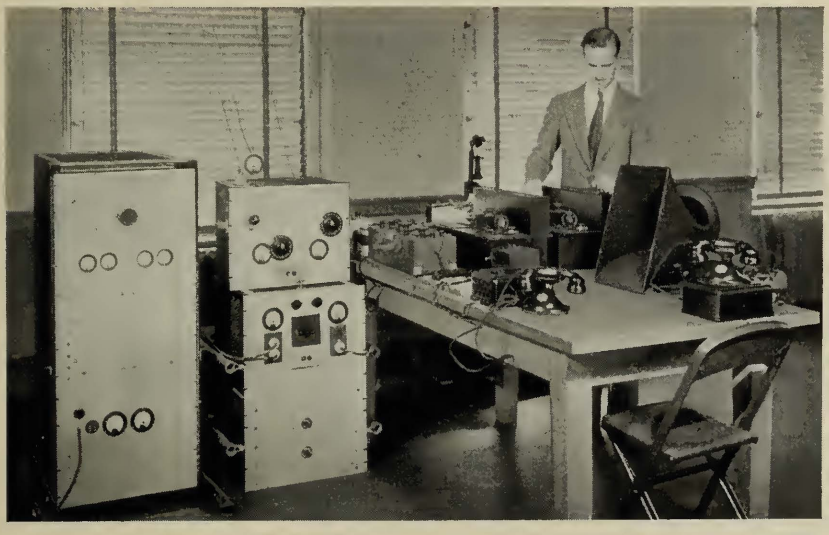


Fig. 13—Fixed terminal transmitting equipment located at Berkeley Street.

wave transmitter and receiver arranged for simultaneous two-way communication. A distinctive feature of this equipment was the use of a single antenna for simultaneously transmitting and receiving. This was made possible by the use of a suppressor circuit in the receiver to prevent overloading of the first detector by the outgoing signal. With this suppressor circuit, which consisted of only a half

[21] Both of Bell Telephone Laboratories, Deal, New Jersey.

section of a simple band-elimination filter, it was possible to transmit and receive simultaneously on frequencies differing by only five percent.

The equipment at the fixed transmitting location is shown in Fig. 13.

The car was used for communication at distances up to about 5 miles from the fixed transmitter and up to a little over 3 miles from the fixed receiver. The circuit from the moving car to the fixed receiver was consistently good for distances up to about two miles.

An Application of Number Theory to the Splicing of Telephone Cables *

By H. P. LAWTHER, JR.

The consideration of a simple and practical splicing scheme for minimizing the recurrence of same-layer adjacencies among telephone circuits in long cables leads to a problem in Number Theory whose solution calls for some extension of the previous work in this field. The solutions for numbers not greater than 139 have been computed, and a table of these is included.

SOME time ago in connection with the placing of a long telephone cable the writer had occasion to attempt the specification of a splicing scheme designed to minimize the recurrence of same-layer adjacencies among the telephone circuits as they threaded their way through successive lengths of the completed cable. The task, superficially so simple, proved to be one of most intriguing difficulty, and the pursuit of the solution led a confused investigator stumbling into the province of number theory. That speculation upon an art so mundane as that of telephone cable splicing should have led to a proposition in the oldest and most neglected branch of mathematics seemed to be especially worthy of note, for few applications so practical have been found. In the course of the investigation certain small ground apparently was covered for the first time. It was felt, therefore, that the story would be of passing interest alike to the mathematician and to the engineer.

The present standard cables for long distance telephone service are manufactured as a series of concentric layers of conductor units contained within a cylindrical sheath. The conductor units are either pairs of quads of wires. The layers are one unit in thickness, and successive layers either spiral in opposite directions of rotation, or in the same direction but with different pitches. The feature of importance to this discussion is that in an unbroken length of cable any one conductor unit will experience shoulder-to-shoulder adjacency throughout this distance with the two conductor units lying on either side in the same layer, and its experience with these two conductor units will be unique. Cables usually are manufactured in uniform lengths of from 750 to 1000 feet, and a longer cable is made up from a succession of such

* Published in *Amer. Math. Monthly*, February, 1935.

lengths spliced end-to-end. At each splice point a large number of different splices is possible among conductor units. In general, wire-to-wire splices are not made, and considerable mixing up is achieved. For reasons which need not be given here it is considered desirable from the standpoint of crosstalk control that each telephone circuit experience the minimum amount of same-layer adjacency with every other telephone circuit.

For the purposes of this discussion it will suffice at present to consider the cross-section of a cable as a simple closed sequence of N consecutively adjacent units. As an example, the array presented by a circular picket fence would be of this character. Each conductor unit in a cable is identifiable, and it will be assumed that each has been "tagged" with one of the numbers 1, 2, 3, 4, $\cdots$, N in such sequence that units bearing consecutive numbers lie adjacent—remembering that unit No. 1 and unit No. N also lie adjacent. While this simple picture of the cable cross-section is representative truly of only a single layer structure, still the results of a study of it may be fitted to apply to practical cases. Schemes for accomplishing this will suggest themselved to the practical worker, and their discussion here would burden this presentation unduly.

Consider now two consecutive lengths in a completed cable and focus attention upon a conductor unit in one of these. At the splice point this conductor unit may connect to any one of the conductor units in the second length, and the two conductor units which lie alongside the latter in the same layer in the second length may connect to any two of the $N - 1$ remaining conductor units in the first length. As an extended conductor unit traverses the completed cable, then, it may experience same-layer adjacency successively with any possible combinations two at a time of the other extended conductor units, and in any order, sequence, or repetition of these as determined by the splicing scheme that is used. Since there can be but $[(N - 1)/2]$* totally different combinations two at a time of $N - 1$ different objects it is evident that $[(N - 1)/2]$ successive cable lengths is the maximum possible number for an extended conductor unit to traverse without incurring repetition of at least one of the same-layer adjacencies that occurred in the first of these lengths.

Any splicing scheme that is devised for practical use must embody the utmost in simplicity. For this reason it is considered highly desirable (1) that the required results be achieved through repetition of the same splicing instruction at consecutive splice points, and (2)

* The symbol $[x/y]$ means the greatest integer not greater than x/y.

that this instruction follow the simplest possible system—e.g., any two adjacent conductor units in one length of cable shall connect to two conductor units having a constant separation in count in the next length. The exposition which follows makes no attempt to solve the general problem, and seeks only to establish the results which can be realized when the above two simplifying restrictions are imposed. At the conclusion is added a description of a minor and acceptable deviation from the second restriction which will enable the practical worker to supplement these results and achieve the maximum possibilities in a number of cases sufficient for his needs. The problem now will be formulated.

1 → 1	1 → 1	1 → 1
2 → 2	2 → 3	2 → 4
3 → 3	3 → 5	3 → 7
4 → 4	4 → 7	4 → 10
5 → 5	5 → 9	5 → 2
6 → 6	6 → 11	6 → 5
7 → 7	7 → 2	7 → 8
8 → 8	8 → 4	8 → 11
9 → 9	9 → 6	9 → 3
10 → 10	10 → 8	10 → 6
11 → 11	11 → 10	11 → 9
Fig. 1	Fig. 2	Fig. 3

The three tabulations exhibited in Figs. 1, 2, and 3 show possible ways of splicing two pieces of eleven-unit cable together in systematic fashion. The left-hand columns indicate the consecutively adjacent conductor units in the first or reference piece of cable (remembering that No. 1 and No. 11 are adjacent), and the numbers opposite in the right hand columns indicate the conductor units in the second piece of cable to which splice is made. No importance attaches to the splicing of unit No. 1 to unit No. 1 in each instance. This is simply one of eleven possible "starts," and from the point of view of this discussion there is no preference among these. Note that with Fig. 1 two conductor units which lie adjacent in the first piece of cable connect to conductor units separated by a count of one (adjacent) in the second piece. With Fig. 2 conductor units which lie adjacent in the first piece connect to conductor units separated by a count of two in the second piece. With Fig. 3 conductor units which lie adjacent in the first piece connect to conductor units separated by a count of three in the second piece. Splices made in accordance with the schemes of Figs. 1, 2, or 3

will be described as made with a "spread of one," a "spread of two," of a "spread of three," respectively. It is readily shown that for a spread number s to be applicable to cable of N units it is necessary and sufficient that s be prime relative to N.

Figure 4 shows the splicing of six pieces of eleven-unit cable through

$$
\begin{array}{rrrrrr}
1 \rightarrow & 1 \rightarrow & 1 \rightarrow & 1 \rightarrow & 1 \rightarrow & 1 \\
2 \rightarrow & 3 \rightarrow & 5 \rightarrow & 9 \rightarrow & 6 \rightarrow & 11 \\
3 \rightarrow & 5 \rightarrow & 9 \rightarrow & 6 \rightarrow & 11 \rightarrow & 10 \\
4 \rightarrow & 7 \rightarrow & 2 \rightarrow & 3 \rightarrow & 5 \rightarrow & 9 \\
5 \rightarrow & 9 \rightarrow & 6 \rightarrow & 11 \rightarrow & 10 \rightarrow & 8 \\
6 \rightarrow & 11 \rightarrow & 10 \rightarrow & 8 \rightarrow & 4 \rightarrow & 7 \\
7 \rightarrow & 2 \rightarrow & 3 \rightarrow & 5 \rightarrow & 9 \rightarrow & 6 \\
8 \rightarrow & 4 \rightarrow & 7 \rightarrow & 2 \rightarrow & 3 \rightarrow & 5 \\
9 \rightarrow & 6 \rightarrow & 11 \rightarrow & 10 \rightarrow & 8 \rightarrow & 4 \\
10 \rightarrow & 8 \rightarrow & 4 \rightarrow & 7 \rightarrow & 2 \rightarrow & 3 \\
11 \rightarrow & 10 \rightarrow & 8 \rightarrow & 4 \rightarrow & 7 \rightarrow & 2
\end{array}
$$

Fig. 4

the successive application of five consecutive identical splices, each with a spread of two. Following the "key" of the first and second columns, the succeeding columns are written down immediately. Scrutiny of the sequences of numbers appearing in the several columns reveals at once the fundamental properties of the spread. For a cable of N units these are:

1. Successive applications of a spread of s for n times result in a spread of s^n.
2. A spread of minus s is equivalent effectively to a spread of plus s.
3. A spread of $KN + s$ (K is an integer: positive, negative, or zero) is the same effectively as a spread of s.

The problem of achieving the minimum possible recurrence of same-layer adjacencies among conductor units through the application of successive similar splices in accordance with a simple spread now may be stated formally in the terminology and symbols of number theory. If N, an integer, is the number of conductor units in the cable, and if s, an integer prime to N, is the spread number used, then it is required to find a value for s for which the companion relations

$$
\begin{aligned}
s^d &\equiv \pm 1 \pmod{N}, \\
s^b &\not\equiv \pm 1 \pmod{N}, \quad b < d
\end{aligned}
$$

determine the largest possible integer d.

From the foregoing introductory discussion it should be noted that

values for N less than 5 are of no significance to this problem. In the analysis which follows, therefore, no particular effort has been made to render the general conclusions capable of extension to these extreme and trivial cases.

It is necessary at this point to recall and introduce certain working material. First, there is the established theorem that every positive integer N greater than unity can be represented in one and only one way in the form

$$N = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_t^{\alpha_t},$$

where $p_1, p_2, \cdots, p_t$ are different primes and $\alpha_1, \alpha_2, \cdots, \alpha_t$ are positive integers. Then there is the familiar number theory function $\phi(N)$ which indicates the number of positive integers not greater than N and prime to N.* If p is a prime number and α is a positive integer, then

$$\phi(p^\alpha) = p^{\alpha-1}(p - 1);$$

also

$$\phi(p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_t^{\alpha_t}) = \phi(p_1^{\alpha_1}) \cdot \phi(p_2^{\alpha_2}) \cdot \cdots \cdot \phi(p_t^{\alpha_t}),$$

where $p_1, p_2, \cdots, p_t$ are different primes.

Then there is the λ-function defined in terms of the ϕ-function as follows:

$$\lambda(2^\alpha) = \phi(2^\alpha) \text{ for } \alpha = 0, 1, 2,$$

$$\lambda(2^\alpha) = \frac{\phi(2^\alpha)}{2} \text{ for } \alpha > 2,$$

$$\lambda(p^\alpha) = \phi(p^\alpha) \text{ for } p \text{ an odd prime},$$

$$\lambda(2^{\alpha_1} p_2^{\alpha_2} p_3^{\alpha_3} \cdots p_t^{\alpha_t}) = M,$$

where M is the least common multiple of

$$\lambda(2^{\alpha_1}), \lambda(p_2^{\alpha_2}), \lambda(p_3^{\alpha_3}), \cdots, \lambda(p_t^{\alpha_t}),$$

$2, p_2, p_3, \cdots, p_t$ being different primes.† Finally, it is established that for two relatively prime integers s and N the value $\lambda(N)$ is the largest possible for the exponent m for which the relations

$$s^m = 1 \pmod{N},$$

$$s^n \not\equiv 1 \pmod{N}, \quad n < m,$$

* Euler, "Novi Comm. Ac. Petrop.," 1760–61, p. 74. Carmichael, "The Theory of Numbers," John Wiley & Sons, Inc., 1914, pp. 30–32. Dickson, "Introduction to the Theory of Numbers," Univ. of Chicago Press, 1929, Chap. I.

† Cauchy, *Comptes Rendus*, Paris, 1841, pp. 824–845. Carmichael, p. 53.

will hold, and that a value for s belonging to this exponent does exist.*

Here it is convenient to consider separately numbers of the two classes—those for which $\lambda(N) = \phi(N)$ and those for which $\lambda(N) < \phi(N)$. For numbers of the first class established theorems may be drawn upon to furnish a complete analysis. For numbers of the second class, however, it will be necessary to extend a bit beyond the ground covered by previous workers, and the steps will be given in considerable detail. This procedure coupled with the inherent complexity will render the treatment for the latter class much less compact and elegant than that for the former.

Case I. $\lambda(N) = \phi(N)$.

From the defined relation between the ϕ-function and the λ-function it follows that numbers of the class such that $\lambda(N) = \phi(N)$ are confined to the values

$$1, 2, 4, p^\alpha, \quad \text{and} \quad 2p^\alpha,$$

where p is an odd prime and α is a positive integer.† For a number N of this class it is established that there exists a set of $\phi(\phi(N))$ numbers r, such that

$$(1) \qquad r^{\lambda(N)} \equiv 1 \pmod{N} \quad \text{and}$$

$$(2) \qquad r^n \not\equiv 1 \pmod{N}, \quad n < \lambda(N), \quad \lambda(N) = \phi(N).$$

Such a number is known as a "primitive root" of N.‡ From the properties of the primitive root r of the number N as defined by relations (1), (2) it follows readily that

$$(3) \quad r^{\lambda(N)/2} \equiv -1 \pmod{N},$$

$$(4) \qquad r^n \not\equiv \pm 1 \pmod{N}, \quad 0 < n < \frac{\lambda(N)}{2}, \quad \frac{\lambda(N)}{2} < n < \lambda(N).$$

First there will be considered the companion relations

$$(5) \qquad s^d \equiv -1 \pmod{N},$$

$$(6) \qquad s^b \not\equiv \pm 1 \pmod{N}, \quad b < d,$$

and, from comparison with relations (3) and (4), these clearly are satisfied for s a primitive root of N and for $d = \lambda(N)/2$. That no

* Carmichael, p. 54 and pp. 61–63.
† Carmichael, p. 71.
‡ Gauss, "Disquisitiones Arithmeticae," Art. 52–55. Carmichael, pp. 65–71. Dickson, Sec. 17.

value for d greater than $\lambda(N)/2$ is possible is evident immediately. For let a be any integer prime to N. Then for some exponent k

$$r^k \equiv a \pmod{N} \quad \text{and}$$

$$a^{\lambda(N)/2} \equiv r^{k\lambda(N)/2} \equiv \pm 1 \pmod{N}.$$

Next there will be considered the companion relations

$$s^d \equiv 1 \pmod{N}, \tag{7}$$

$$s^b \not\equiv \pm 1 \pmod{N}, \quad b < d. \tag{8}$$

The reasoning just above shows that d cannot be greater than $\lambda(N)/2$. Suppose for the moment that d has this greatest possible value $\lambda(N)/2$. Relations (7), (8) then become

$$s^{\lambda(N)/2} \equiv 1 \pmod{N},$$

$$s^b \not\equiv \pm 1 \pmod{N}, \quad b = 1, 2, 3, \cdots, \lambda(N)/2 - 1.$$

These relations may be written

$$(s^{1/2})^{\lambda(N)} \equiv 1 \pmod{N}, \tag{9}$$

$$(s^{1/2})^{2b} \not\equiv \pm 1 \pmod{N}, \quad 2b = 2, 4, 6, \cdots, \lambda(N) - 2. \tag{10}$$

Now relations (9), (10), will be compatible with relations (1), (2), (3), (4) only if $\lambda(N)/2$ is an odd number, for otherwise the restrictions of relation (10) applying to the even numbered exponents from 2 to $\lambda(N) - 2$ inclusive would be in conflict with relation (3). For $\lambda(N)/2$ an odd number, then, relations (9), (10), are satisfield for $s^{1/2}$ a primitive root of N. Consequently, with relations (7), (8), $\lambda(N)/2$ is the largest possible value for the exponent d, and a value for s equal to the square of a primitive root of N permits this to be attained.

Case II. $\lambda(N) < \phi(N)$.

The inquiry for this case will be divided into four parts. In general $N = p_1^{\alpha_1} p_2^{\alpha_2} p_3^{\alpha_3} \cdots p_t^{\alpha_t}$ where $p_1, p_2, p_3, \cdots, p_t$ are different primes.

(a) First will be considered the case where $p_1, p_2, p_3, \cdots, p_t$ are all odd primes. Then $\lambda(N)$ is the least common multiple of $\lambda(p_1^{\alpha_1})$, $\lambda(p_2^{\alpha_2})$, $\lambda(p_3^{\alpha_3})$, $\cdots$, $\lambda(p_t^{\alpha_t})$. Suppose now that the highest power of 2 dividing any of the λ's divides $\lambda(p_i^{\alpha_i})$. If this same power of 2 divides more than one of the λ's, arbitrarily select $\lambda(p_i^{\alpha_i})$ as one of them. Then this power of 2 will be exactly that occurring in $\lambda(N)$. Now arbitrarily select p_j as any one of the odd primes other than p_i. Then clearly

$\lambda(N)$ will also be the least common multiple of

$$\lambda(p_1^{\alpha_1}),\ \lambda(p_2^{\alpha_2}),\ \cdots,\ \lambda(p_{j-1}^{\alpha_{j-1}}),\ \frac{\lambda(p_j^{\alpha_j})}{2},\ \lambda(p_{j+1}^{\alpha_{j+1}}),\ \cdots,\ \lambda(p_t^{\alpha_t}).$$

Now take

$$\begin{aligned} r &\equiv g_j^2 \pmod{p_j^{\alpha_j}}, \quad g_j \text{ a primitive root of } p_j^{\alpha_j}, \\ &\equiv g_k \pmod{p_k^{\alpha_k}}, \quad g_k \text{ a primitive root of } p_k^{\alpha_k}, \\ &\qquad\qquad k = 1, 2, 3, \cdots, j-1, j+1, \cdots, t. \end{aligned}$$

The r thus chosen must be prime to each of the prime factors of N, and hence must be prime to N. Consequently it is known that

$$r^{\lambda(N)} \equiv 1 \pmod{N}.$$

Suppose that m is the smallest exponent for which the congruence

$$r^m \equiv 1 \pmod{N}$$

is true. Then it is noted that the chosen r is such that m must be a multiple of

$$\lambda(p_1^{\alpha_1}),\ \lambda(p_2^{\alpha_2}),\ \cdots,\ \lambda(p_{j-1}^{\alpha_{j-1}}),\ \frac{\lambda(p_{1j}^{\alpha_j})}{2},\ \lambda(p_{j+1}^{\alpha_{j+1}}),\ \cdots,\ \lambda(p_t^{\alpha_t}),$$

and the least multiple common to these is, of course, $\lambda(N)$. Therefore it can be written that

$$\begin{aligned} r^{\lambda(N)} &\equiv 1 \pmod{N}, \\ r^b &\not\equiv 1 \pmod{N}, \quad b < \lambda(N). \end{aligned}$$

Now suppose that for some exponent n less than $\lambda(N)$

$$r^n \equiv -1 \pmod{N}.$$

Then

$$r^{2n} \equiv 1 \pmod{N},$$

and if n is less than $\lambda(N)$, $2n$ is less than $2\lambda(N)$ and can only be equal to $\lambda(N)$. It would necessarily follow then that

$$r^{\lambda(N)/2} \equiv -1 \pmod{N},$$

and it would follow in turn that

$$r^{\lambda(N)/2} \equiv -1 \pmod{p_j^{\alpha_j}}.$$

However, r has been chosen such that

$$r^{\lambda(N)/2} \equiv (g_j^2)^{\lambda(N)/2} \equiv g_j^{\lambda(N)} \equiv 1 \pmod{p_j^{\alpha_j}}.$$

This last relation is incompatible with the one immediately above, and it must be concluded that the assumption

$$r^n \equiv -1 \pmod{N}, \quad n < \lambda(N)$$

is false, and that for the r that has been chosen

$$r^{\lambda(N)} \equiv 1 \pmod{N},$$
$$r^b \not\equiv \pm 1 \pmod{N}, \quad b < \lambda(N)$$

and no exponent greater than $\lambda(N)$ is possible.

(b) Next will be considered the case where $p_1 = 2$, $\alpha_1 = 1$, and $p_2, p_3, \cdots, p_t$ are all odd primes. Select p_i as above and take p_j different from 2. Then take

$$\begin{aligned} r &\equiv 1 \pmod{2}, \\ &\equiv g_j^2 \pmod{p_j^{\alpha_j}}, \quad g_k \text{ a primitive root of } p_j^{\alpha_j}, \\ &\equiv g_k \pmod{p_k^{\alpha_k}}, \quad g_k \text{ a primitive root of } p_k^{\alpha_k}, \\ &\qquad k = 2, 3, 4, \cdots, j-1, j+1, \cdots, t, \end{aligned}$$

and the same line of reasoning may be repeated and the same conclusions reached as under part (a) above.

(c) Next will be considered the case where $p_1 = 2$, $\alpha_1 = 2$ and $p_2, p_3, \cdots, p_t$ are all odd primes. Since $\lambda(2^2) = 2$ take p_i different from 2, and for simplicity take p_j as 2. Then take

$$\begin{aligned} r &\equiv 1 \pmod{4}, \\ &\equiv g_k \pmod{p_k^{\alpha_k}}, \quad g_k \text{ a primitive root of } p_k^{\alpha_k}, \\ &\qquad k = 2, 3, 4, \cdots, t \end{aligned}$$

and the same line of reasoning may be repeated and the same conclusions reached as under part (a) above.

(d) Finally will be considered the case where $p_1 = 2$, $\alpha_1 > 2$, and $p_2, p_3, \cdots, p_t$ are all odd primes. Now 5 has the property that

$$5^{\lambda(2^{\alpha_1})} \equiv 1 \pmod{2^{\alpha_1}},$$
$$5^b \not\equiv \pm 1 \pmod{2^{\alpha_1}}, \quad \alpha_1 > 2, \quad b < \lambda(2^{\alpha_1}).$$

So by taking

$$r \equiv 5 \pmod{2^{\alpha_1}},$$
$$\equiv g_k \pmod{p_k^{\alpha k}}, \quad g_k \text{ a primitive root of } p_k^{\alpha k},$$
$$k = 2, 3, 4, \cdots, t$$

it is concluded immediately that

$$r^{\lambda(N)} \equiv 1 \pmod{N},$$
$$r^b \not\equiv \pm 1 \pmod{N}, \quad b < \lambda(N).$$

The preceding formal analysis for Case I and Case II may be summed up as having established the following general theorem:

If N is a given positive integer and if s is an integer prime to N, then the largest possible exponent d for which the companion congruencial relations

$$s^d \equiv \pm 1 \pmod{N},$$
$$s^b \not\equiv \pm 1 \pmod{N}, \quad b < d$$

will be true is $\lambda(N)/2$ for numbers such that $\lambda(N) = \phi(N)$ and is $\lambda(N)$ for numbers such that $\lambda(N) < \phi(N)$, and a value for s belonging to this exponent in each instance does exist.

In order to apply the foregoing results to a practical case Table I has been prepared. In the left-hand column appear the numbers 5 to 139, inclusive. In the next column is listed for each number the value of $\lambda(N)/2$ or of $\lambda(N)$, depending upon whether $\lambda(N) = \phi(N)$ or $\lambda(N) < \phi(N)$. In the final column there is listed for each number a suitable value for the spread. There appears to be no advantage of one spread figure over another, and the listing of additional acceptable values is omitted in the interest of economy of space. For the numbers for which $\lambda(N) = \phi(N)$ and for which $\lambda(N)/2$ is odd care has been taken that the listed spread figures are primitive roots, and not the squares of primitive roots which were shown to be equally acceptable. This fact will be recalled later.

It was shown earlier that $[(N-1)/2]$ successive cable lengths would be the maximum possible number for an extended conductor unit to traverse without incurring repetition of at least one of the same-layer adjacencies which occurred in the first of these lengths. On referring to Table I it is seen that only for the prime numbers is this maximum attainable. The prime numbers are distinguished by the fact that for them $\lambda(N)/2 = (N-1)/2$, and each has been indicated by an asterisk. The composite numbers are seen to yield quite inferior results in general.

TABLE I

For each N there is listed the value d and a value s for which the companion relations $s^d \equiv \pm 1 \pmod{N}$ $s^b \not\equiv \pm 1 \pmod{N}$, $b < d$ determine the largest possible integer d.

N	d	s	N	d	s	N	d	s
5*	2	2	50	10	3	95	36	2
6	1	5	51	16	5	96	8	5
7*	3	3	52	12	7	97*	48	5
8	2	3	53*	26	2	98	21	3
9	3	2	54	9	5	99	30	5
10	2	3	55	20	2	100	20	3
11*	5	2	56	6	3	101*	50	2
12	2	5	57	18	5	102	16	5
13*	6	2	58	14	3	103*	51	5
14	3	3	59*	29	2	104	12	7
15	4	2	60	4	7	105	12	2
16	4	3	61*	30	2	106	26	3
17*	8	3	62	15	3	107*	53	2
18	3	5	63	6	2	108	18	5
19*	9	2	64	16	3	109*	54	6
20	4	3	65	12	3	110	20	3
21	6	2	66	10	5	111	36	2
22	5	7	67*	33	2	112	12	3
23*	11	5	68	16	3	113*	56	3
24	2	5	69	22	2	114	18	5
25	10	2	70	12	3	115	44	2
26	6	7	71*	35	7	116	28	3
27	9	2	72	6	5	117	12	2
28	6	5	73*	36	11	118	29	11
29*	14	2	74	18	5	119	48	3
30	4	7	75	20	2	120	4	7
31*	15	3	76	18	21	121	55	2
32	8	3	77	30	2	122	30	7
33	10	5	78	12	7	123	40	7
34	8	3	79*	39	3	124	30	7
35	12	2	80	4	3	125	50	2
36	6	5	81	27	2	126	6	11
37*	18	2	82	20	7	127*	63	3
38	9	3	83*	41	2	128	32	3
39	12	2	84	6	5	129	42	14
40	4	3	85	16	3	130	12	3
41*	20	6	86	21	3	131*	65	2
42	6	11	87	28	2	132	10	5
43*	21	3	88	10	3	133	18	2
44	10	3	89*	44	3	134	33	7
45	12	2	90	12	7	135	36	2
46	11	5	91	12	2	136	16	3
47*	23	6	92	22	3	137*	68	3
48	4	5	93	30	13	138	22	7
49	21	5	94	23	5	139*	69	2

* The asterisk indicates a prime number.

For the benefit of the practical worker there must be described a slight deviation from the second simplifying restriction imposed at the beginning which will permit the maximum possibility to be realized if N is one plus a prime number. This artifice is based upon the fact

that for r a primitive root of a number N for which $\lambda(N) = \phi(N)$ and in particular for a prime number N

$$r^{\lambda(N)/2} \equiv -1 \pmod{N}.$$

This means that $\lambda(N)/2$ consecutive splices with a spread r result in a spread of minus one. It is readily shown that this in turn means that there will be two conductor units No. b and No. $b + 1$ in the first length of cable which ultimately will be extended to connect respectively to units No. $b + 1$ and No. b. In Fig. 4 two units No. 6 and No. 7 meet this requirement. To illustrate the use of this artifice it will be supposed that a cable of 12 units is to be spliced. Referring to Fig. 4. for guidance, the arrangement shown in Fig. 5 is set up readily. The first two columns indicate the splicing assignment, and the succeeding columns are then derived from these. The eleven units 1, 2, $\cdots$, 5, 6, 8, 9, $\cdots$, 11, 12 are assigned exactly in conformity with the scheme of Fig. 4, ignoring the break in sequence between No. 6 and No. 8. Unit No. 7 is then simply spliced to itself throughout.

$$\begin{array}{cccccc}
1 \to & 1 \to & 1 \to & 1 \to & 1 \to & 1 \\
2 \to & 3 \to & 5 \to & 10 \to & 6 \to & 12 \\
3 \to & 5 \to & 10 \to & 6 \to & 12 \to & 11 \\
4 \to & 8 \to & 2 \to & 3 \to & 5 \to & 10 \\
5 \to & 10 \to & 6 \to & 12 \to & 11 \to & 9 \\
6 \to & 12 \to & 11 \to & 9 \to & 4 \to & 8 \\
7 \to & 7 \to & 7 \to & 7 \to & 7 \to & 7 \\
8 \to & 2 \to & 3 \to & 5 \to & 10 \to & 6 \\
9 \to & 4 \to & 8 \to & 2 \to & 3 \to & 5 \\
10 \to & 6 \to & 12 \to & 11 \to & 9 \to & 4 \\
11 \to & 9 \to & 4 \to & 8 \to & 2 \to & 3 \\
12 \to & 11 \to & 9 \to & 4 \to & 8 \to & 2
\end{array}$$

Fig. 5

Undoubtedly there are other equally acceptable artifices for extending further the practical scope of the simple results. The prime numbers and the prime numbers plus one constitute nearly fifty percent of all numbers in the range in which the practical worker is likely to be interested, however, and when it is borne in mind that normally he has latitude in his choice of N it is seen that the material here presented is adequate for his needs.

The writer is indebted to Dr. D. H. Lehmer for pertinent suggestions. The entire treatment for the case of numbers for which $\lambda(N) < \phi(N)$ follows a line of attack suggested by Mr. Marshall Hall, and but for his helpful interest this presentation would have been lacking in formal completeness.

Contemporary Advances in Physics, XXIX
The Nucleus, Fourth Part

By KARL K. DARROW

The earlier parts of this series have dealt with the charge, the mass, the stability or instability, and the liability to transmutation, of the atom-nucleus; this one deals with the two remaining properties which are ascribed to nuclei, to wit, magnetic moment and angular momentum. Since these exhibit themselves chiefly by influencing the spectra of the atoms to which the nuclei belong, the bulk of the article is concerned with various laws of atomic spectra; advantage is taken of the opportunity to describe some features of the electron-systems surrounding the nuclei, and to explain how the concept of the spinning electron enters into atomic physics. There follows an account of various experiments in which streams of atoms are deflected by inhomogeneous magnetic fields, and the laws of the deflections indicate the magnetic moment or the angular momentum of the nucleus or both. Finally there is a summary and tabulation of existing knowledge of these quantities.

The Nucleus as a Quantized Vector

UNDER this somewhat forbidding title I propose to discuss some phenomena—mostly spectroscopic, but in certain cases magnetic or even chemical—which are interpreted by supposing that the nuclei of atoms are endowed with two vectorial qualities, *magnetic moment* and *angular momentum*. One may say that these nuclei are to be visualized, no longer simply as particles possessed of mass and charge alone, but as bodies—usually, as congeries of particles both charged and uncharged—which are in incessant rotation: the spinning of the mass constitutes their angular momentum, the spinning of the charge a perpetual circular current-flow which is equivalent to a magnet. Why, then, should I not have entitled this section "The Spins and Magnetic Moments of Nuclei"? Chiefly because it might have suggested, at the very outset, that nuclear spins and magnetic moments are observed as clearly and measured as directly, as are nuclear masses and charges: which in the main is not so. With only a couple of exceptions (for magnetic moment) they are deduced from phenomena which certainly carry no obvious sign of their character. Indeed, what is called the nuclear angular momentum or the *spin* is distinguished by a feature, which is altogether strange and foreign to angular momentum of ordinary wheels and gyroscopes and the other spinning things of daily life; and it is by virtue of this foreign and paradoxical feature—and not any of the familiar qualities of spinning things—

that the concept of "spin" is a valuable addition to atom-models. This feature it is which is responsible for calling the nucleus a "quantized" vector. But I need not further justify at this point my suggestion that spin is a quality both strange and remote from immediate experience, for this will be only too evident in what is to follow.

It is desirable to make a longish excursion through some of those early-known and more-or-less familiar features of spectra, which are explained solely by calling into account the orbital or circum-nuclear electrons of the atom, without taking notice of any property of the nucleus excepting its charge which holds those electrons together. This excursion will have the incidental advantage of making us aware of the magnetic moment and the spin of the *electron*, though we cannot pause long enough to get an adequate appreciation of the scope and value of this concept of the "spinning electron." It is best to simplify these preliminary steps as much as is conveniently possible, and therefore I will speak of a single spectrum only; yet not the spectrum of hydrogen, for this would carry simplicity rather too far. By a queer paradox of which we are destined to see altogether too much, it often happens in quantum mechanics that the seemingly-simplest of all cases—the ones to which one would go by choice for a start—are liable to a singularly-confusing complexity quaintly known as a "degeneracy." Let us therefore take sodium for our example.

The sodium atom has a nuclear charge $+ 11e$, and eleven orbital electrons, of which the eleventh or "valence electron" is sharply contrasted with the other ten. Employing the original atom-model of Bohr, one visualizes the ten as describing a network of interlacing close-packed orbits enclosing the nucleus as in a sort of cage, while the eleventh darts about in some sort of a far-flung orbit which at one end may enter the cage, while the other end reaches far out into space ("penetrating orbits") or alternatively an entire orbit may encircle the cage completely, not entering it at all ("non-penetrating orbit").

We now consider the major feature of the sodium spectrum—which indeed is the only feature of the absorption-spectrum of tranquil sodium vapor—a beautiful converging series of lines. All of the lines correspond to transitions between one and the same state, *viz.* the normal state of the sodium atom, and the various members of a sequence of abnormal or excited states. These are distinguished by an index n to which consecutive integer values are attached; and all together they are known as the P sequence, to distinguish them from other sequences of states which a fuller study of the sodium spectrum (not confined to the absorption-spectrum of the tranquil vapor) discloses. The terminology, of course, is a detail. What is essential is,

that they are all discriminated from each other by something physical, and yet they all have something in common which distinguishes them as a whole from states of other sequences. As to the nature of these "somethings," the original theory of Bohr is perfectly explicit. All the states correspond to orbits of the valence electron; the different P states correspond to orbits of different but definite sizes and shapes; what all the P orbits have in common is, *a common value of the angular momentum of the electron revolving in its orbit.*

To repeat Bohr's argument here would be an unjustifiable use of time and space. Let me merely recall that when it was applied to the hydrogen atom, it led to several extraordinary agreements with the data of experiment, to which others have since been added. To some degree these are transferable to the sodium atom, especially since the electric field in which the valence-electron of the sodium atom mostly revolves is extremely like that in which the electron of the hydrogen atom always revolves (*i.e.*, beyond the "cage" to which I referred, the ten electrons of the cage effectively cancel the influence of the portion $+ 10e$ of the nuclear charge, leaving uncancelled only the field of a nuclear charge $+ e$ which is the same as the charge of a hydrogen nucleus). These agreements were the primary cause of the enormous role which angular momentum has ever since been playing in all atom-models. They were responsible also for the numerical values now assigned to angular momenta which figure in atoms; for according to the original theory, the orbital angular momentum of the P states is $2(h/2\pi)$, and all the other values which it may take for states of other sequences are other small-integer-multiples of $(h/2\pi)$; and while these values have since been somewhat altered without impairing the numerical agreements on which they rested and on which now the new ones rest, it remains true that *all angular momenta occurring in atom-models are expressed as multiples of* $\frac{1}{2}(h/2\pi)$, the multiplying factors being integers usually smaller than ten.

As my words have already implied, there are various types of angular momentum nowadays fitted into atom-models, the one already described—hereafter to be called the "orbital angular momentum"—being only one and the first. We turn now back to the principal series of sodium to discover why another type is required.

Examined with a sufficiently good spectroscope, each "line" of that series is found to be actually a close doublet. These imply that each of what I have been calling the P states is actually a pair of states.[1] (The confusion introduced into language by referring to one and the

[1] That it is not the normal state which is resolved into a pair is proved by various facts which it is not necessary to mention here.

same spectroscopic object sometimes as a line and sometimes as a pair or group of lines, and the corresponding confusion about states, are very bad impediments to clarity of exposition, but there is simply no way of getting around them.) Such is in fact the case. The two members of a pair have a common value of the index n and a common value of the orbital angular momentum of the valence-electron, and yet there must be something physical which distinguishes them. To leap at once to the conclusion: in the atom-model, this something is *orientation.*

But, orientation of what with respect to what? We have as yet introduced only one outstanding direction, only one vector, into the atom-model. The outstanding direction is that of the normal to the orbital plane of the valence-electron; the vector is the orbital angular momentum. To these it is necessary to supply a second vector,—as it turns out, a second angular momentum.

Since it may occur to some reader that the natural place to seek this angular momentum is among the electrons of the cage—that we should begin by assigning a net or resultant angular momentum to the ten electrons which we have hitherto so much neglected—I will recall that such was actually the first suggestion. It prevailed during the early twenties, and was generally accepted; but it suffered from certain disadvantages, which now there is no particular reason for retelling at length. Yet it was of the greatest assistance in preparing the ground and the technique for the suggestion which superseded it in the middle twenties, that the second angular momentum is to be ascribed to the electron itself; the electron is to possess, like the earth, not only a motion of revolution but also a motion of rotation. (As for the ten electrons of the cage, their angular momenta both of rotation and of revolution are so oriented as to balance one another out, and we have made no error in neglecting them.)

Now there are two vectors and two directions in the model of the sodium atom: that of the axis of the rotating electron, and that of the normal to the orbit—the spin momentum and the orbital momentum. To speak of different orientations of the one with respect to the other is now sensible. But different orientations must correspond to different energies if they are to explain the data, since the two lines of a principal-series doublet are separate and distinguishable because and only because the two members of a pair of P-states differ in energy. Why should they?

This happens to be the easiest question of the lot, or at any rate the one which can be answered from classical physics. The rotating electron is a magnet, by virtue of its whirling charge. Also it behaves

as though it were moving through a magnetic field, and this is not so easy to grasp, since we have postulated merely that it is revolving in the electrostatic field surrounding the nucleus and the cage of inner electrons. Yet if we were to shift our frame of reference and imagine, not the nucleus standing still and the valence-electron revolving around it, but the electron standing still and the nucleus revolving around *it*—then we should have no difficulty in realizing that the revolving nucleus, being a charge describing a closed path, and being therefore equivalent to a current, would produce a magnetic field. Reverting now to our original frame of reference, we may bring the magnetic field back with us, and say that the *spinning* electron revolves both in the electrostatic field aforesaid and in the magnetic field, and its energy is influenced by both. This is not a very sophisticated way of looking at the matter, and there is a tricky little relativistic detail in the shifting of the frame of reference, which produces an error of a factor 2 if disregarded; but it serves to bring out the idea. The energy of the valence-electron in its orbit is affected by this quasi-magnetic interaction due ultimately to the fact that it is a magnet moving through an electrostatic field; and *the value of the energy depends on the orientation*—on the angle θ between the axis of the magnet-electron and the normal to the orbital plane. One now sees readily that there will be a minimum energy occurring when these two directions are parallel, and a maximum energy occurring when they are anti-parallel. But why then do we not find the individual P-state spread out into a continuous band of states corresponding to all the energy-values between these two extremes, and all the infinity of different orientations between the value 0° and the value 180° of the angle θ?

This is no question which classical physics can solve. The fact that the individual P-state, or what would otherwise be the individual P-state, is split into two instead of into an infinity—this fact implies that only two orientations occur in nature, are "permitted," as the phrase is; and this instance of *quantization of direction*, like all the other instances of quantization, is a consequence of the quantum-mechanical constitution of the world. More lucid instances occur when the atom with its electrons is immersed in an applied magnetic field of known intensity; that is, when sodium vapor is exposed to the measurable field of a large-sized magnet, and its spectrum is observed. We will take up some of these instances before beginning with the assignment of spin to the nucleus.

When a magnetic field of moderate strength is applied to sodium vapor, each of the doublets of the principal series is split up into a pattern of several lines or "components." I can no longer say that

the P-states themselves are split up into as many components as are the lines: that was a happy coincidence while it lasted, but it does not repeat itself henceforward (nor usually). From the subdivision of the lines it is necessary to deduce the subdivision of the initial and the final states of the corresponding transitions: that is a classical task of spectroscopy, which we may assume to have been achieved. It is found that the normal state of the sodium atom (not belonging to the P sequence) is resolved by the magnetic field into two components, while of each pair of states belonging to the P sequence, one is resolved into two and the other into four components. Incidentally, the separations of these components are proportional to the strength of the magnetic field. It appears, therefore, that the sodium atom possesses properties of a magnet, quantized in direction; or rather, that in different states it is equivalent to different magnets, since in certain states it has two permitted orientations in the field, while in others it has four.

As might be guessed, the magnet to which the atom-as-a-whole is equivalent is a sort of *resultant* of the two magnets, spinning electron and "orbital magnet," which have already been separately inserted into the model. This quantum-mechanical resultant, however, possesses a couple of peculiarities, into which we shall have to look rather carefully. To lead up to them, it is desirable to look at the two special cases in which (a) there is no spin and (b) there is no orbital angular momentum, so that the resultant reduces to a magnet of one of the two types. Strictly speaking, case (a) never occurs in sodium, but to work it out is useful, nevertheless.

I have spoken of the electron revolving in its orbit as being equivalent to an "orbital magnet": now is the time to fortify that statement by giving a cardinal relation between the magnetic moment of that orbital magnet (not now the magnetic moment due to the electron-spin!) and the angular momentum of that orbital motion. It is sufficient to work out the relation for a circular orbit. Let r stand for the radius of the circle, v for the speed of the electron, $n(= v/2\pi r)$ for the number of times per second that the electron runs around the circle; e for the charge, m for the mass, p for the angular momentum of the electron, M for the magnetic moment of the system. The revolving electron is equivalent to a current[2] ne/c flowing in a circuit enclosing the area πr^2; the magnetic moment of such an affair is equal to the current-strength times the enclosed area:

$$M = (ne/c)\pi r^2, \tag{1}$$

[2] The factor c enters in because e is commonly expressed in electrostatic units, whereas in equation (1) the current must be expressed in electromagnetic units.

while for the angular momentum, we have:

$$p = mvr, \tag{2}$$

and eliminating vr between (1) and (2), we get:

$$M/p = e/2mc. \tag{3}$$

Here we have on the left the *ratio of magnetic moment to angular momentum*, a very important quantity for all these categories of atomic and subatomic magnets. It is equated to $e/2mc$, a value which is correct whenever we are dealing with the magnet constituted by the orbital motion (not the spin!) of an electron; this equation in fact is valid for any sort of an orbit described in a central field, and is one of the few that have survived unamended all of the stages in the evolution of Bohr's original theory into quantum mechanics. I will rewrite the equation thus:

$$M/p = g(e/2mc), \quad g = 1, \tag{4}$$

and this is meant to imply that for other categories of atomic and subatomic magnets, the ratio of the moments is not always equal to $e/2mc$, which is true. In general it is the custom to characterize any one of these magnets by giving its value of g. *Orbital magnets, then, are characterized by the value unity for the g-factor.*

We next ascertain the energy which the orbital magnet possesses by virtue of being in the applied field H. Letting α stand for the angle between the direction of the field and the axis of the magnet, we find for the torque exerted on the magnet by the field,

$$T = -MH \sin \alpha \tag{5}$$

and integrating to obtain the energy in question,

$$U = \int T d\alpha = MH \cos \alpha. \tag{6}$$

This we now write as follows:

$$U = (MH/p)p \cos \alpha = (geH/2mc)\ p \cos \alpha, \tag{7}$$

and here is as good an opportunity as any to recall a well-known theorem of classical mechanics, fundamental in the theory of the gyroscope. When a torque is acting upon a rotating body, the body precesses around the direction of the field responsible for the torque; and if the torque be equal to a constant T_0 times the sine of the angle between the field-direction and the axis of the rotating body, then the angular velocity ω of the precession is equal to the ratio between T_0

and the angular momentum of the rotation.[3] This ratio is the one enclosed in parentheses in equation (7): we may therefore write:

$$U = \omega p \cos \alpha. \tag{8}$$

The precession of which ω is the angular velocity is known as the *Larmor precession.* To keep this precession in mind as a feature of the atom-model is usually desirable, though not constantly necessary. It must be supposed to occur not only when an atom is immersed in an extraneous magnetic field, but also when two of the subatomic magnets within a single atom are influencing one another.

We have treated the (non-existent) extreme case in which there is nothing but the angular momentum and the magnetic moment of the orbital motion of a single electron to be taken account of; we turn now to the other extreme in which there is nothing but the spin of a single electron to be taken into account. Were we still confined to the original atom-model of Bohr, this case would be equally non-existent; for no electron-orbit could have a vanishing angular momentum unless it were a straight line passing to and fro through the nucleus, and this was formerly excluded as unthinkable. Quantum mechanics, however, assigns the value zero to the angular momentum of a valence-electron in a state of the S sequence (to which the normal state of the sodium atom belongs). Whether the student prefers to visualize a straight-line "orbit" for such a case, or a spherical cloud of charge or of "probability-of-charge," is to some extent a matter of taste, though usually the latter is the better policy. For such a state, there is no angular momentum and there is no magnetic moment save those of the electron-spin itself.

To this spin of the electron—whether isolated as in this extreme example, or compounded with an orbital motion into a resultant—we are compelled by various reasons to assign the value $2(e/2mc)$ for that important ratio of magnetic moment to angular momentum. Otherwise expressed: *the spin of the electron is characterized by the value 2 for the g-factor.*

There is a classical argument for this assertion, based on an evaluation of the ratio in question for a sphere of homogeneous charge rotating about an axis passing through its centre. There is a more powerful quantum-mechanical argument, based on the fact that when Schroedinger's fundamental equation of wave-mechanics was amended by Dirac to be conformable with relativity, there appeared in it a term attributable to a whirling charge with a g-factor of 2. Apart from

[3] Slater and Frank, "Introduction to Theoretical Physics," Chapter X.

this last, the strongest argument is furnished by the validity of the verifiable formula for the g-factors of various atoms in various states, which we shall presently be deriving. For, inasmuch as in nearly every state of an atom there are both electron-spins and electron-orbits, and the net magnetic moment and net angular momentum of the atom are sorts of resultants of these, the g-factor of the atom-as-a-whole varies from state to state and from one kind of atom to another in a remarkable fashion, which imposes a stringent test on the contemporary atom-model.

In explaining how this test is satisfied, I can no longer postpone specific numerical statements about the angular momenta of electron-orbits and electron-spins and the atoms into which these enter. These statements will be made confusing by the fact that to each one of these angular momenta it will be necessary to assign two different numbers. This is one of the impediments which are unavoidable in fitting visualizable atom-models to the results of quantum-mechanical theory, and which to avoid, some theorists would be willing to forego models altogether.

Were it not for this impediment, I could say quite simply that in every P-state of the sodium atom, the valence-electron has a spin, with angular momentum $\frac{1}{2}(h/2\pi)$, and is moving in an orbit with angular momentum $(h/2\pi)$; that as regards the two members of each of the aforesaid pairs of P-states, these two angular momenta are oriented parallel for one member and anti-parallel for the other, so that the net angular momentum of the atom-as-a-whole is $\frac{1}{2}(h/2\pi)$ in one case and $\frac{3}{2}(h/2\pi)$ in the other; that when an atom with a net angular momentum of $\frac{1}{2}(h/2\pi)$ is exposed to an applied magnetic field, it orients itself either parallel or anti-parallel to the field, so that the projection of its angular momentum upon the field-direction is either $+\frac{1}{2}(h/2\pi)$ or $-\frac{1}{2}(h/2\pi)$; that when an atom with a net angular momentum of $\frac{3}{2}(h/2\pi)$ is exposed to an applied magnetic field, it orients itself in one or another of four permitted ways so that the projection of its angular momentum upon the field-direction is either $+ 3/2$ or $+ 1/2$ or $- 1/2$ or $- 3/2$ times $h/2\pi$.

This sort of thing *is* frequently said in the literature, and one must realize its limitations. The trouble is, that quantum mechanics prescribes for these angular momenta (but not for their projections on the field-direction!) magnitudes which differ from those which I have been giving. For the spin s of the electron, it substitutes $\sqrt{1/2 \cdot 3/2}(h/2\pi)$ for $\frac{1}{2}(h/2\pi)$; for the orbital motion l in the present case, it substitues $\sqrt{1 \cdot 2}$ for the factor unity whereby $(h/2\pi)$ was multiplied; in the two values j' and j'' of angular momentum of the atom-as-a-

whole, it substitutes $\sqrt{1/2 \cdot 3/2}$ and $\sqrt{3/2 \cdot 5/2}$ for the factors 1/2 and 3/2. In order that these new values of j' and j'' may be resultants of s and l, it is necessary that the angular momenta of orbital motion and of spin should be neither quite parallel nor quite anti-parallel to one another—the two permitted values of the angle between them must differ from 0° and 180°; and this also is admitted by quantum mechanics. The last two clauses of the foregoing paragraph remain, however, unchanged.

Then why not say from the outset that the spin of the electron has the magnitude $\sqrt{1/2 \cdot 3/2}(h/2\pi)$, and that all the other angular momenta occurring inside an atom have the magnitudes assigned to them by quantum mechanics? Inertia of habit, sanctified by years of earlier theories, is itself a mighty obstacle: after so long a time of saying that the electron-spin is 1/2, the world of physics could scarcely get accustomed to saying that really it is $\sqrt{1/2 \cdot 3/2}$ (in terms of $h/2\pi$ as unit). There is also a difficulty of printing to be considered: these numbers have often to be used as subscripts; it is bad enough to print 1/2 or 3/2 as a subscript, without resorting to their quantum-mechanical substitutes. The most serious reason, however, is, that the original 1/2 and 3/2 (and, of course, their analogues in the many other kinds of atomic states) lend themselves uniquely well to stating how many permitted orientations there are. Thus in the present example, I quoted 2 and 4 respectively as the number of permitted orientations in a magnetic field, of certain P states for which the angular momenta were designated as $j(h/2\pi)$ and j had the values 1/2 and 3/2 respectively. I was reminded of those numbers by the rule that they are equal to $(2j + 1)$. Had I kept in mind only the magnitudes $\sqrt{3/4}$ and $\sqrt{15/4}$ assigned by quantum mechanics to these angular momenta, the rule would not have been available.

One should therefore keep in mind both the "marker" or "quantum-number" of an angular momentum, of which the foregoing 1/2 and 1 and 3/2 are examples; and the numerical value assigned by quantum-mechanics to the magnitude of that angular momentum. Fortunately this is rendered easy by the fact that there is a general formula for the latter in terms of the former, with which we can now make acquaintance in the course of taking a deeper plunge into the lush notation of spectroscopy; as follows:

The angular momentum of electron-spin has the quantum-number s and the magnitude $\sqrt{s(s + 1)}$, and s is always equal to 1/2.

The angular momentum of the orbital motion of an electron has the quantum-number l and the magnitude $\sqrt{l(l + 1)}$, and l may have the various values 0, 1, 2, 3 $\cdots$.

The angular momentum of the atom-as-a-whole has the quantum-

number j and the magnitude $\sqrt{j(j+1)}$; for the sodium atom j may have the various values $1/2, 3/2, 5/2 \cdots$.

I next give the rule for compounding the last of these three angular momenta out of the first two, it being as I earlier said a "sort of resultant" thereof.

Rule for compounding l and s into j. If l is greater than s, start with the numerical value of $(l+s)$ and write down all the sequence of numbers spaced at unit intervals from $(l+s)$ to $(l-s)$ inclusive: to wit $(l+s)$, $(l+s-1)$, $(l+s-2)$, $\cdots (l-s)$. These $(2s+1)$ numbers are the permitted values of the quantum-number j. If on the other hand s is greater than l, start with the numerical value of $(s+l)$ and write down all the sequence of numbers spaced at unit intervals from $(s+l)$ to $(s-l)$ inclusive: to wit, $(s+l)$, $(s+l-1)$, $(s+l-2) \cdots (s-l)$. These $(2l-1)$ numbers are the permitted values of the quantum-number j.

(It will be noticed that this rule is much more generally phrased than is required for the case of sodium, where $s = 1/2$, and it suffices to say that $j = 1/2$ for $l = 0$ and $j = l \pm 1/2$ for $l > 0$. If, however, we were dealing with an atom having more than one valence-electron, s might be replaced by a quantum-number different from $1/2$—not because the individual electrons would have new values of spin, but because the spins of two or more of them would be compounded—and the general phrasing of the rule would then be required). And now, to close (temporarily) the sequence of quantum-numbers and of rules:

Suppose the atom immersed in a magnetic field of strength H, parallel to the z-direction. It may then take any of several distinct permitted orientations, these being denoted by various values of a quantum-number m_j. In any such orientation the projection of the angular momentum of the atom-as-a-whole upon the z-direction or field-direction is equal to $m_j(h/2\pi)$. To ascertain how many permitted orientations there are and what are the corresponding projections, start with the numerical value of j and write down all the sequence of numbers spaced at unit intervals from $+j$ to $-j$, inclusive: to wit, $j, j-1, j-2, \cdots -j$. These $(2j+1)$ numbers are the permitted values of the quantum-number m_j.

We now start out upon a train of reasoning which leads to the remarkable verifiable formula already once alluded to, the successfulness of which speaks more powerfully than any other single test for the rightness of this elaborate hypothetical structure which has been devised for the atom.

Note, in the first place, that the rule given above for the permitted orientations of the atom in the applied magnetic field is in accordance

with a major fact of experience: the energy-values of the atom in its various "magnetic levels," as I shall hereafter call them, are known from spectroscopy to follow upon one another in a uniform evenly-spaced sequence. This is just what the rule requires: for if M denotes the magnetic moment of the atom and θ_j the angle which it makes with the z-direction, the energy due to the field is equal to $MH \cos \theta_j$, which is $MH(m_j/\sqrt{j(j+1)})$, and this changes by uniform steps as m is changed from each of its permitted values to the next.

Now recalling the importance of the ratio of magnetic moment to angular momentum, and noticing that $M/\sqrt{j(j+1)}$ is none other than this ratio, and introducing the g-factor of equation (7), and the precession ω of equation (8), we may write for ΔU the energy-difference between one magnetic level and the next:

$$\Delta U = g(e/2mc)H = \omega, \tag{9}$$

so that a measurement of the energy-difference or separation of two magnetic levels of an atom gives immediately the value of g for that atom. One sees at once how to make a special test of the value 2 assigned to the g-factor of the electron-spin; for when the sodium atom is in any state for which $l = 0$, it is the spinning valence-electron which contributes the whole of the magnetic moment and the angular momentum of the atom; and when from the spectrum of sodium vapor in the magnetic field the value of ΔU is determined for these states, it is precisely the value $2(e/2mc)H$ which is found.

To get a notion of what actually happens in the general case, it is best to take a sheet of paper and make a graphic composition of the angular momenta. These three—s, l, and j, to denote them by their quantum-numbers—are to be laid down as a triangle having sides of the lengths $\sqrt{s(s+1)}$, $\sqrt{l(l+1)}$, and $\sqrt{j(j+1)}$; for convenience I drop out the common factor $h/2\pi$ for the next few lines. The cosines of the three angles are obtained in terms of the sides by applying the well-known trigonometric formula and getting three equations of which here is one,

$$s(s+1) = l(l+1) + j(j+1) - 2\sqrt{l(l+1)}\sqrt{j(j+1)} \cos \theta_{l,\,j}. \tag{10}$$

The magnetic moment of the orbital electron-motion is a vector parallel to l and of the length $g(e/2mc)\sqrt{l(l+1)}$, with *unity* put as the value of g. The magnetic moment of the electron-spin is a vector parallel to s and of the length $g(e/2mc)\sqrt{s(s+1)}$, with *two* put as the value of g. Owing to the inequality of these g factors, the resultant of the magnetic moments is not parallel to j. We could easily calculate its magnitude and direction, but they are not relevant. It is

to be presumed that s and l are constantly revolving or precessing around the direction of j,—the triangle aforesaid is constantly revolving around its side j as a fixed axis, while retaining its size and shape unchanged. If we resolve the resultant angular momentum into a component parallel to j and another component perpendicular to j, the latter will be forever changing in direction and its average will vanish, leaving only the former as a perpetual constant. But this former component we can evaluate by projecting separately upon the j-direction the magnetic moments associated with orbital motion and spin, and adding the two projections. Thus for the magnetic moment of the atom-as-a-whole we get an effective average which is a vector parallel to the angular momentum of the atom-as-a-whole, and it is of the magnitude:

$$M_a = [\sqrt{l(l+1)}\cos\theta_{l,j} + 2\sqrt{s(s+1)}\cos\theta_{s,j}]\frac{e}{2mc}\frac{h}{2\pi} \qquad (11)$$

(where I have restored the factor $h/2\pi$). If we work this out with the aid of (10) and the similar equation for $\cos\theta_{s,j}$, and then divide it by the angular momentum $(h/2\pi)\sqrt{j(j+1)}$ and by $e/2mc$, we get *the g-factor for the atom-as-a-whole*—commonly denoted by g_j—in terms of the quantum-numbers s, l and j:

$$g_j = 1 + \frac{j(j+1) + s(s+1) - l(l+1)}{2j(j+1)}, \qquad (12)$$

and this is the celebrated *g-formula*, which is tested by applying magnetic fields H to atoms, splitting their stationary states into clusters of levels, measuring the separation between successive levels of a cluster, equating it to $\omega = g(e/2mc)H$, evaluating g and comparing it with the value which the right-hand member of (12) assumes when in it s, l and j are given the values appropriate to the state from which the cluster of levels was formed. So great is the variety of atomic states, so great the number of different triads of values of s, l, j represented among them, that the study of even a single element like sodium produces many different checks of the validity of (12); and since many different elements have been studied, the total of the available verifications of the g-formula, and therefore of the intricate network of its underlying ideas, is considerably impressive.

The temptation of going onward and onward into the details of these properties of the extranuclear electrons and their orbits is difficult to resist, but it must be overcome, for the field is practically endless. I must add only, that when an atom possesses two or more valence-

electrons, their orbital momenta $l_1, l_2 \cdots$ are likely[4] so to orient themselves as to form a fixed resultant L, and their spin-momenta s_1, $s_2 \cdots$ are likely to orient themselves so as to form a fixed resultant S; then L and S are likely so to orient themselves as to form a fixed resultant J, following in so doing the rule of page 295. Considered as quantum-numbers, J and S may have half-integer values only (1/2, 3/2 $\cdots$) or full-integer values only (0, 1, 2 $\cdots$) according as the number of valence-electrons is odd or even, while L for any atom may have full-integer values only. The three resultants are vectors of magnitudes $\sqrt{J(J+1)}$, $\sqrt{S(S+1)}$ and $\sqrt{L(L+1)}$. The g-factor associated with L is unity and the g-factor associated with S is two, and the value of g for the atom-as-a-whole is given by (12) with capital letters replacing the small ones; so that the g-formula is verifiable with atoms of all kinds, as I intimated before. A final point: one might expect the conplexity to go on increasing tremendously from one end to the other of the Periodic Table, but there is a counteraction. In all atoms excepting the lightest, most of the electrons have oriented their orbits and their spins in such a way that they have interlocked themselves into groups or "closed shells" for which L is zero and S is zero and J is zero and the magnetic moment is zero, as have the ten electrons of the "cage" of the sodium atom to which I alluded. The so-called valence-electrons are those few which have not been locked into any such a cage. It is this quality which makes the Periodic Table periodic; but this must be left for some other place.

We arrive at last at the nuclear moment.

Suppose that even with these spins and these orbital motions of all the extra-nuclear electrons, we have not yet exhausted the internal angular momenta of the atom, and that the nucleus itself possesses one. Suppose, to be specific, that the nucleus has an angular momentum with a quantum-number I and a magnitude $\sqrt{I(I+1)}(h/2\pi)$, and a propensity for orienting itself in distinct permitted directions with respect to the other angular momenta of the atom. How shall we detect this, and how shall we determine I?

It is practically necessary to be yet more specific. One could probably not tell *a priori* whether the nuclear angular momentum would tend to orient itself with special respect to individual electron-momenta, or with special respect to some resultant or in particular to that grand resultant of all electron-momenta which we have denoted by J. However, in the cases which have been successfully analyzed, it

[4] This is the description of what is known as "Russell-Saunders coupling" or "LS coupling"; in certain states of certain atoms, the mutual orientations of the vectors conform to different schemes.

turns out that the last condition is the one which prevails. Suppose then that I compounds itself with J according to the following rule, repeated word-for-word with appropriate changes of symbol from the rule for compounding s and l into j:

Rule for compounding I and J into F. If I is greater than J, start with the numerical value of $(I + J)$ and write down the sequence of numbers spaced at unit intervals from $(I + J)$ to $(I - J)$ inclusive: to wit, $(I + J)$, $(I + J - 1)$, $(I + J - 2)$, $\cdots$ $(I - J)$. These $(2J + 1)$ numbers are the permitted values of the quantum-number F. If J is greater than I, start with the numerical value of $(J + I)$ and write down the sequence of numbers spaced at unit intervals from $(J + I)$ to $(J - I)$ inclusive. These $(2I + 1)$ numbers are the permitted values of the quantum-number F.

The quantum-number F refers to a vector of magnitude $\sqrt{F(F + 1)}$ $(h/2\pi)$, which has taken over from J the role of the angular momentum of the atom-as-a-whole, being the resultant of J and of the nuclear angular momentum I.

Now if all these suppositions are correct, we may expect to find not individual states, but whole serried clusters of states, corresponding to individual values of J. If out of the manifold term-system of an atom we select a state for which $J = 1/2$, one for which $J = 3/2$, one for which $J = 5/2$ and so on as far upward as our knowledge extends, we may expect on close scrutiny to find that these apparent states are actually clusters, each cluster comprising a number of states which for one or two or more of the lowest values of J may be equal to $(2J + 1)$, but for higher values reaches and remains at a limit which we identify as $(2I + 1)$.

What is observed with the spectroscope, though, is not the individual state, but the line which reveals a transition between two different states. What in a feeble spectroscope appears as a single line, and is attributed to a transition between two states with resultant electronic angular momenta (I fear no shorter term will serve henceforth) j' and j'', should in an excellent spectroscope appear as a cluster of lines due to transitions between the several members of two clusters of states.

This again is exemplified by the principal series of sodium. As I said earlier, this appears in a feeble spectroscope as a series of single lines, each of which is resolved by a good spectroscope into a doublet. This structure, by the way, is called the "fine structure" of the lines; and this it is which indicates that the P-states of sodium are close pairs, and which thus invites and requires the introduction of the quantum-numbers j and s and the spin-momentum of the electron.

With a very good spectroscope indeed, each member of each principal-series doublet is in its turn resolved into a pair. This structure is called the "hyperfine structure" of the lines; and this it is which indicates that a still further subdivision of the states is necessary, and invites and requires the introduction of the quantum-numbers F and I and the angular momentum of the nucleus. Indeed, the concept of the nucleus as a quantized vector was invented or discovered (whichever word the reader may prefer) during the interpretation of hyperfine structure of spectrum lines.

Hyperfine structure of lines or states (for the name is applied to both) is usually more crowded and compact than fine structure, and yet there are exceptions: the fine structure of hydrogen is much harder to resolve than the hyperfine of (say) the familiar mercury lines 2537 and 5461, which itself was called fine before the theory was developed. These structures, however, are generally near and often, it is to be suspected, beyond the utmost capacities of the most refined of optical instruments; whence, in many cases, extraordinary difficulties in measuring or even estimating the separations, the relative intensities, actually the mere number of the distinct lines forming a hyperfine pattern; observers of great skill will often disagree with one another, and judgment will often depend on a photograph taken with a spectroscopic instrument such as an echelon or an etalon, which looks totally different from the pictures obtained with gratings or prisms. Perhaps this last is an advantage after all, as it discourages attempts by the inexpert to interpret published photographs. Often several different isotopes of an element produce different patterns which signify different values of I, and are so nearly superposed on one another as to make analysis superlatively hard. Hyperfine structure is for the present, and quite probably will be forever, the "last frontier" of spectroscopy.

The task of deriving, from the hyperfine line-pattern connecting two states or (better) state-clusters, the hyperfine subdivision of the state-clusters or "hyperfine multiplets" themselves, is again an example of the classical function of spectroscopy, which we shall take as having been achieved. Actually it involves, of course, the use of selection-principles, themselves connected with the atom-model, but omitted from this article in order not to complicate it still more. Some confusion may be prevented if I state that in our favorite case of sodium, where the fine-structure splitting of the principal-series lines implies a splitting of the P-states only, the hyperfine splitting implies something more complex: it is due jointly to hyperfine structures of both the P-states and the normal S-state, the latter being predominant.

We now consider briefly the methods of determining the quantum-number I of the nuclear angular momentum.

(*a*) The *ideal method* is the one already described: investigate line-clusters connecting state-clusters of as many different values of J as there are; ascertain thus the number N of states per cluster; verify that N is equal to $(2J + 1)$ whenever J is less than or equal to some particular value J_m (say), and that it is equal to $(2J_m + 1)$ whenever J is equal to or greater than J_m. All this being verified, the value of I must be J_m.

One seldom if ever finds such a programme as this worked out very fully. The difficulties seem to be that, at best, it is a lot of work to analyze the hyperfine-structure of even one line, let alone a great number; while at worst, lines connected with states of certain J-values, high ones especially, may be quite unobservable; also there is the striking fact that in a given spectrum a very few lines or even one alone may have their hyperfine-structures spread out so much more broadly than all the rest, that research is practically concentrated on them alone. (One hears so much about 4722 of bismuth as almost to have it blotted from mind that bismuth has other lines!) But of course if one is willing to accept without test the rule for compounding the vectors I and J, then it suffices to discover and analyze a state-cluster for which N is less than $(2J + 1)$.

(*b*) *The intervals between the members of a state-cluster* may give a clue to the value of I. These intervals are of course energy-differences, and the fact that they exist shows that there are forces between the spinning nucleus and the system of revolving and spinning electrons which surrounds it. If these forces are magnetic, then they may reasonably be expected to vary as the sine of the angle $\theta_{I,J}$ between the angular momenta of nucleus and extranuclear electron system; for either the magnetic moments of these two will be parallel to the respective angular momenta, or else (by the reasoning of page 297) their non-parallel components will presumably change so rapidly as to be ineffective, leaving only the parallel components to be detectable. Comparing the different orientations of I and J which correspond to the several values of F and thus to the states of the cluster, one sees that their energies—or rather, the parts W_F thereof which are due to the interaction—should then vary as $\cos \theta_{I,J}$: putting for which the formula based on (10)

$$W_F = \text{const.} \frac{F(F+1) - I(I+1) - J(J+1)}{2\sqrt{I(I+1)}\sqrt{J(J+1)}}. \qquad (13)$$

Insert in (13) the permitted values of F, which are $(2I + 1)$ or $(2J + 1)$ in number according as J is or is not greater than I, and are spaced at unit intervals from $(I + J)$ downwards (page 299); call them, in order of descending magnitude, F_m, F_{m-1}, F_{m-2} $\cdots$. Form the $2I$ or $2J$ consecutive first differences between the so-computed permitted values of W_F; call them ΔW_m, ΔW_{m-1}, ΔW_{m-2}, $\cdots$ Then as is readily worked out,

$$\Delta W_m : \Delta W_{m-1} : \Delta W_{m-2} \cdots : : F_m : F_{m-1} : F_{m-2} \cdots$$
$$: : (I + J) : (I + J - 1) : (I+J - 2) \cdots.$$

The successive energy-differences or *intervals* should stand to one another as the successive members of the chain of integers (or half-integers, as the case may be) stepped off at unit intervals and stretching from $(I + J)$ downwards.[5]

This is an *interval-rule* based on a specific notion of the intra-atomic forces (the sine-law aforesaid), and having analogues in the parts of atomic theory having to do with the interactions between electrons the extra-nuclear electrons only. If verified, it enables one to determine $(I + J)$ and therefore I from the analysis of a single cluster of states with a single value of J, even when J is smaller than I and the preceding method would fail. Much use has been made of this method, and there are a few cases in which a fairly accurate measurement of a chain of intervals has shown that it closely agrees with a chain of consecutive integers or half-integers, though more usually the intervals are small and the measurements rough and it is merely assumed that there is perfect agreement with that particular succession of half-integers or integers with which there is the nearest apparent agreement.

(*c*) *The relative intensities of the members of a line-cluster* are capable of giving information about the quantum-numbers of the states which they connect, provided one adapts quantum-mechanical formulae developed for transitions into which the nuclear angular momentum does not enter. The formulae are of appalling complexity, while intensity-measurements, especially when one is working so near the limits of the possible as when hyperfine-structure is being measured, are notoriously liable to error. This method is probably to be classified as by far the least reliable, for the present at any rate.

[5] The biggest interval may be that between the highest and the next-to-highest energy-value, or that between the lowest and the next-to-lowest; whichever case is realized gives a clue to the "sign" (page 318) of the magnetic moment; usually the former corresponds to a positive, the latter to a negative moment, but features of the extra-nuclear electron-system may cause this statement to be reversed. Incidentally, relative intensities of lines also have a bearing on the sign of the moment.

(*d*) *The phenomenon of alternating intensities in band-spectra* serves to reveal the spins of a few kinds of nuclei, and in a very interesting and reliable way, but must be left for another occasion.

There remain the methods which involve the use of a magnetic field in one way or another, and some of which incidentally tell most of what we know about the magnetic moments of nuclei, though nowhere near so amply or so exactly as we should like.

First it must be said that the analogy between the vectors I and J on the one hand, L and S on the other, which thus far has been so full and helpful, breaks down completely when the atom is exposed to a magnetic field of ordinary strength. Were the analogy perfect, an atom in a state distinguished by the quantum-number F for its total angular momentum would act as a rigid spinning body and would be able to assume $(2F + 1)$ discrete orientations in the magnetic field, corresponding to $(2F + 1)$ magnetic levels. This would be true of each of the $(2I + 1)$ or $(2J + 1)$ states comprised in what I have been calling a "cluster" with a common value of J, though the value of F and hence of $(2F + 1)$ would differ from one state to the next. The magnetic levels would be distributed in groups, each corresponding to a different value of F. The numbers in the different groups would be unequal. The total number for all the groups or states of the cluster would amount, as the reader can figure out, to the product $(2I + 1) \times (2J + 1)$.

It is altogether probable that this is precisely what does happen in magnetic fields so weak as not to separate the magnetic levels perceptibly (their separation being then, it will be recalled, proportional to the field-strength). Yet in fields strong enough to produce a measurable effect, the disposition of the magnetic levels has only one thing in common with this hypothetical distribution. Their total number *is* precisely $(2I + 1)(2J + 1)$. They are, however, distributed in $(2J + 1)$ groups, each consisting of $(2I + 1)$ levels; as though first of all the atoms were to forget their nuclear angular momentum and remember only their electronic angular momentum, and were to orient themselves in the field in the $(2J + 1)$ different ways which were prescribed for them (page 295) while the nucleus was still being neglected; and as though then they were to remember the nuclear angular momentum, and were to allow for it by adopting, in place of each separate one of the $(2J + 1)$ very different orientations, a group of $(2I + 1)$ orientations differing only a little from it and from each other.

This rather animistic idea is not very far from the model commonly conceived. It is supposed that in the strong magnetic field the nucleus

is somehow broken away from its interlocking with the system of extra-nuclear electrons; not in the sense that it is torn out of the system or that its electrostatic attraction for the electrons is suspended, but in the sense that somehow or other the rule for the compounding of I and J into various values of a resultant vector F is done away with. The system of electrons with its angular momentum J chooses among $(2J + 1)$ orientations with respect to the field, just as if the nucleus were not there; and the nucleus chooses among $(2I + 1)$ orientations with respect to the field, just as if the electronic system were not there. Nucleus and electrons, I and J, are said to be "decoupled" from each other; it is supposed that their angular momenta precess each at its own rate separately around the direction of the field.[6]

This account suggests that the energy-values of the magnetic levels would be given by the various values of the expression

$$M_e H \cos \theta_{J,\,H} + M_n H \cos \theta_{I,\,H}, \tag{14}$$

where $\theta_{J,\,H}$ and $\theta_{I,\,H}$ stand for the inclinations of the angular momenta J and I with respect to the field-direction, while M_e and M_n signify the magnetic moments of the extra-nuclear electron-system and of the nucleus, or, if these magnetic moments be not parallel to J and I respectively, then their projections upon the directions of J and I. The different groups of levels would then correspond to different permitted values of $\cos \theta_{J,\,H}$, the different levels of any one group to different permitted values of $\cos \theta_{I,\,H}$. The first term, one might say, would determine the $(2J + 1)$ separate energy-values which would occur if there were no angular momentum of the nucleus, while the second term would subdivide each of these into a group of $(2I + 1)$ levels.

It is found, however (as we shall later see) that the magnetic moments M_n of nuclei are always so very small by comparison with those of extranuclear electron-systems, that the second term of (14) is quite negligible. We have therefore to look for some other cause for the observable subdivision. This cause is thought to be the force between the moments of the nucleus and the electron-system. We assumed it to be overborne by the strong field in so far as its ability to control the quantized directions of the angular momenta is concerned,

[6] The analogy of I and J with L and S is restored when the impressed field is very strong, for then L and S are similarly decoupled from one another—"Paschen-Back effect" as distinguished from the "Zeeman effect" which we have hitherto been considering. Thus it is roughly correct to say that hyperfine structure reacts to a weak magnetic field as fine structure does to a strong one, though this statement should be carefully qualified if use were to be made of it.

but now must suppose it still to be potent enough to affect the energy of the atom. It adds a third term to (14) which is supposed to be of the form *const.* $M_e M_n \cos \theta_{I,\ H} \cos \theta_{J,\ H}$, but in any event will have $(2I + 1)$ distinct values for every value of J, and is not necessarily (or anyhow is not known to be) too small to account for the observed separations between the members of each group. It therefore allows us to estimate I by counting the members of such groups and equating their number to $(2I + 1)$, and this (when available) is one of the most acceptable ways of determining the angular momentum of the nucleus.

With a spectroscope one counts, as always, not the number of levels but the number of lines connecting them with some other family of levels, and expects that the two numbers will not be the same. By a rare if not unique coincidence, however, they *are* the same: the selection-principle which is involved is such, that each group of levels produces a group of an equal number of lines, or (in other words) if the influence of the nucleus resolves every state of the atom in a magnetic field into a group of $(2I + 1)$ different levels, then it also resolves every line connecting two such states into a group of $(2I + 1)$ different lines. There are in the literature magnificent photographs of the spectrum lines of bismuth exposed to a magnetic field, each line under high resolution exhibiting ten components and proving the value 9/2 for I.

It is, however, sometimes possible to count the levels directly, by sending a beam of fast-moving atoms through an inhomogeneous magnetic field which spreads it out into a diverging fan of smaller beams or pencils, each consisting exclusively of atoms having a certain distinctive value for the projection of the magnetic moment upon the field-direction. This requires a great refinement of the celebrated method of Gerlach and Stern, a refinement which has been achieved by Rabi and his school.

We take, as usual, sodium for our example. Consider a narrow beam of sodium atoms, moving with uniform speed along the x-direction into a region pervaded by a magnetic field which is parallel to the z-axis, and of which the magnitude H varies as rapidly as possible with z. Were it not for this variation of H with z, nothing would happen to the beam, for (to make the crudest possible picture) each atomic magnet would have both its north and its south pole exposed to the same field-strength, and one would be pushed as hard as the other was pulled, resulting in no net force upon the magnet and no deflection. But when the field varies with z and the atomic magnet is oriented otherwise than at right angles to the z-axis, the north and the south pole will be exposed to different field strengths, there will be a resulting

force and a resulting deflection of the flying atom. Force and deflection will increase with the z-component of the magnetic moment of the atom, and atoms with different values of this component will go in different directions.

First we disregard the influence of the sodium nucleus. The sodium atoms in experiments of this type are always in their normal state, for which I recall that $l = 0$ and $j = s$: the angular momentum and the magnetic moment are exclusively those of the spin of the valence-electron, the former having quantum-number 1/2 and magnitude $\sqrt{\frac{1}{2}(\frac{1}{2}+1)}\ (h/2\pi)$, the latter being parallel to the former and having magnitude $g(e/2mc)$ times the magnitude of the angular momentum, with 2 for the value of g. Having this value, the angular momentum of the atom may orient itself in either of the two ways which are crudely (page 292) called "parallel" and "antiparallel" to the field, though it is better (page 293) to think of the two permitted inclinations to the field-direction as being $arc\ cos\ \frac{1}{2}/\sqrt{\frac{1}{2}(\frac{1}{2}+1)}$ and $arc\ cos\ (-\frac{1}{2}/\sqrt{\frac{1}{2}(\frac{1}{2}+1)})$. Half of the atoms are so oriented that the z-component of their magnetic moment is $\frac{1}{2}g(e/2mc)(h/2\pi)$, the other half so that the z-component has the negative of this value: the beam is split into two, diverging oppositely and symmetrically from the axis of x. The detection of this splitting is the Gerlach-Stern experiment.

Now we suppose that the sodium nucleus has an angular momentum of quantum-number I, as a result of which the two orientations aforesaid are not really two, but actually are two groups of $(2I+1)$ not-very-different orientations apiece. The problem is, to refine the method sufficiently to bring out the fact (if it is a fact) that each of the two apparent beams aforesaid is actually a close group of several, and to count the several.

It is necessary to lengthen out the path of the atoms in the inhomogeneous magnetic field, since the longer their exposure to the deflecting agent lasts, the farther the separate beams are drawn apart; this means magnifying the scale of the apparatus and the volume which has to be kept evacuated, and carrying to a very high pitch the geometrical accuracy of its design, since the initial not-yet-separated beam must be exceedingly narrow and must be shot forth in a very-exactly-adjusted direction from its source into the field. It is essential also to reduce the broad distribution-in-velocity which the atoms owe to the fact that they come out of a furnace (in which sodium is being vaporized) with the random velocities of thermal agitation appropriate to the temperature of the furnace, and which would more than suffice to merge the beams which it is now desired to separate. One gathers that even at present it would not be possible to make the wished-for separation,

were it not for one more feature of the laws of the behavior of atoms and their internal angular momenta in the magnetic field.

I have described at length how, in magnetic fields of the field-strengths customary in spectroscopy, where the angular momenta I and J of nucleus and electron-system are "decoupled" and orient themselves independently in the field, each state with a given value of J is converted into $(2J + 1)$ groups of $(2I + 1)$ levels apiece. But in no field at all, I and J are "coupled" into a resultant F, or rather into one or another of several such resultants; and I mentioned that there is reason to suppose that, in fields very much weaker than the customary ones, this coupling subsists and the atom orients itself as a single entity in the field. I emphasized then (page 303) that different as are these two extremes, they have one feature in common: the total number of the levels corresponding to a single value of J, which at both extremes is $(2I + 1)(2J + 1)$. The practical usefulness of this theorem is diminished by the fact that some of these levels may have identical values of energy, but in the atom-model they are nevertheless distinct.

One naturally gl[illegible]hat as the field-strength is increased from "very weak" to "c[illegible]ary," each level of the one extreme passes over into a level of [illegible]her extreme, so that for any field-strength low, intermediate or [illegible] there are always just $(2I + 1)(2J + 1)$ of them. There arise t[illegible]he lesser problem of ascertaining the "correlation," i.e. which le[illegible]f the one extreme goes over into which of the other; and the grea[illegible]roblem of ascertaining just how, for each of these continuously-d[illegible]ite levels, the energy-value and the component of the magnetic mo[illegible]ent along the field-direction—which latter determines the deflection, and which let us call M_z—vary with the field-strength H. Formidable theoretical articles have been written on both of these problems, culminating in rules for the former and formulae for the latter. They were worked out originally for the behavior of the vectors L and S in applied magnetic fields, but are translated into rules and formulae available for our present interests by simply replacing these vectors with I and J and making corresponding changes in the g-factors.[7] For such an atom as hydrogen or sodium in its normal state, for which $J = \frac{1}{2}$, I will quote the formula from Breit and Rabi.

[7] Strictly one should take into account the influence of the magnetic field on the interrelations between L and S and on those between I and J simultaneously, but it usually happens that when H is increased to a magnitude which already suffices to decouple I and J pretty thoroughly, it is not yet great enought to do much to the coupling between L and S. I have spoken of this range of magnitudes as "customary," on the ground that it is usual in experiments on the Zeeman effect; but there is no good single word for qualifying it, inasmuch as it is simultaneously weak with respect to the (L, S) coupling and strong with respect to the (I, J) coupling.

We have seen that in the absence of magnetic field, the normal state is subdivided into two by the influence of the nuclear angular momentum I and its coupling with J. In one of these—call it N'—the vectors I and J are nearly parallel, and their resultant, the angular momentum F of the atom-as-a-whole, has the quantum number $(I + J)$ which is $(I + \frac{1}{2})$; in the weak field where the coupling is not broken, and the state N' maintains its identity, the number of permitted orientations is $(2F + 1)$ which is $(2I + 2)$. In the other—call it N''—the vectors I and J are nearly anti-parallel, and F has the quantum number $(I - \frac{1}{2})$; in the weak field the number of permitted orientations is $(2F + 1)$ which is $2I$. Adding, we get for the total number of magnetic levels the value $(4I + 2)$, which is equal to $(2I + 1)\ (2J + 1)$ as I stated. In the weak field, the different levels are distinguished by their values of the magnetic quantum-number m, which is defined by saying that the projection F_z of the angular momentum (of the atom-as-a-whole) on the field-direction is equal to $m(h/2\pi)$. The permitted values of m are $(I + \frac{1}{2})$, $(I - \frac{1}{2})$, $(I - \frac{3}{2}) \cdots$, $-(I - \frac{1}{2})$, $-(I + \frac{1}{2})$. The first and the last of these values are attached each to a single level, belonging (in the weak field) to the state N'; each of the others is attached to a pair of levels, one belonging to the state N' and the other to the state N''.

We know that each of these levels maintains its identity as the field-strength is increased, even when the coupling of I and J into F is broken down and the separate states N' and N'' lose their identities. We wish to know how the value of M_z for each level is varying as the field increases. Let a stand for $2m/(2I + 1)$; let b stand for the energy-difference between N' and N''; let g stand for the g-factor associated with the extra-nuclear electron-system and with the angular momentum J; let x stand for $(g/b)(eH/2mc)(h/2\pi)$. The formulae of Breit and Rabi are as follows[8]:

$$M_z = \pm \frac{a + x}{2(1 + 2ax + x^2)^{1/2}}\, g(e/2mc)(h/2\pi) \qquad (16)$$

For the levels characterized by the extreme values of m(*viz.*, $\pm (I + \frac{1}{2})$) and initially belonging exclusively to N', the first factor is equal to one half and the two levels are distinguished by the two choices of sign, and M_z is independent of field-strength. With respect to the other values of m, the situation is more complex and curious. A single value of m, say $(I - \frac{1}{2})$, corresponds to two different values of M_z which are equal in magnitude and opposite in sign; the opposite value of m,

[8] Perhaps it is not superfluous to remark that in the factor $(e/2mc)$, the symbol m always stands for electron-mass, never for magnetic quantum-number.

in this case $-(I - \frac{1}{2})$, also corresponds to two different values of M_z which are equal and opposite, and each of which at zero field-strength (but not for $H = 0$) coincides with one of the previous two. Now consider any two of these values of M_z which correspond to equal and opposite values of m and coincide with one another at $H = 0$. As H is increased from zero, these two are altered in opposite senses, and one of them actually passes through zero and then reverses at its sign at a certain value of field-strength (m and hence a are negative, and M_z vanishes when $x = a$) while the other is shifted in the opposite sense and never vanishes (m and a are positive).[9] We shall presently see (page 310) that this behavior is the basis of one of the methods of evaluating I. One more peculiarity of these equations must be stressed: the magnetic moment of the nucleus nowhere appears in them! This becomes evident when the field-strength is put equal to zero and x vanishes, for then the several values of the right-hand member of (16) become simply the projections, upon the field-direction, of the magnetic moment of the extra-nuclear electron-system. Thus we have the paradox that in these experiments the magnetic field gives us information about the nucleus by virtue of the force which it exerts upon the atom, and yet this force is exerted practically upon the electrons alone, and not to any perceptible extent upon the nucleus.

The laws expressed in equation (16) have thus far assisted in three ways in the study of the nucleus:

First, in respect to the experiment which I was describing (page 306) when I began on this detour: as H is decreased from what I called the "customary" magnitude, the $(2I + 1)$ levels constituting each of the there-mentioned groups draw gradually apart—i.e. they differ more and more in respect of the value of the component of the magnetic moment along the field-direction, which is what controls the deflection. The experiment must therefore be performed with field-strengths H which are sufficiently low, much lower than those customarily employed in the Gerlach-Stern experiment or in spectroscopy; and this is one of the distinctive features of the technique of Rabi and his school. Narrowness of the beam is all the more required, since dH/dz must be large enough to produce considerable deflection, and if both its value and the breadth of the beam in the z-direction were large, H could not be small in every part of the beam. The beam must also be made nearly homogeneous in speed, and this is done by a

[9] Exception must be made for pairs of values of M_z, both members of which correspond to $m = 0$ and vanish at $H = 0$; each member of such a pair departs farther and farther from zero, to equal extents in opposite senses, as H is increased from zero.

clever utilization of the law embodied in (16), for which I must refer to the original papers.—When the experiment was performed on sodium, it was found that the beam is split into eight components instead of merely two, proving the value 4 for the factor $(2I + 1)$ and the value 3/2 for the quantum-number of the nuclear angular momentum.

Second, suppose an experiment done by measuring the number of atoms which go through the deflecting field entirely undeflected. These are the atoms for which $M_z = 0$, and in ideal conditions there would never be any such atoms, excepting at one of the particular values of field-strength for which the M_z of one (or rather, two at a time) of the levels mentioned on page 309 is passing through zero; in actual conditions one would expect the curve of the number-of-undeflected-atoms *versus* the field-strength to exhibit peaks. On examining equation (16) one may see that there would be one peak for $I = 1$ or 3/2, two for $I = 2$ or 5/2, three for $I = 3$ or 7/2, and so on.[10] The number of peaks by itself thus gives a partly ambiguous indication, but the ambiguity can be resolved by another theorem deducible from (16): if we denote by H_1, H_2, $\cdots$ the abscissæ of the consecutive peaks, then all the intervals $(H_i - H_{i-1})$ are equal in any case, but the value of H_1 is equal to the half or to the whole of their common value, according as I is a full integer or a half-integer.—The curve for caesium was found to display three peaks, and the second criterion showed that the value of I is a half-integer, therefore 7/2.

Third, when M_z is found by measuring the deflected beams in an apparatus where field and field-gradient are accurately known (not a stringent requirement in either of the two previous cases), one may use equation (16) to compute b: thus determining the "hyperfine-structure" separation between two states without an optical measurement! This has been done with both varieties of hydrogen, the heavy isotope and the light, because for these very important atoms the separation in question is far too small to be detected by any optical device: the method of magnetic deflection has proved itself the superior of the long-established arts of spectroscopy, hitherto regarded as the *ne plus ultra* of subtlety and refinement.[11] The results of these experiments are mostly quoted for their bearing on the magnetic moment

[10] Not counting the peak at $H = 0$ which (it is obvious) must always appear but has no bearing on the value of I.

[11] I should, however, perhaps make exception for the most delicate of these, which is the derivation of hyperfine structure from observations on the resonance radiation produced by polarized light acting on atoms of gases in magnetic fields, and is practiced by Ellett and his school at the University of Iowa. The complexity of the theory forbids a description of the method in this place, but several values of I have been obtained by it.

of the nucleus, that vector quality of which till now next to nothing has been said. Little indeed can be said about it with assurance, but we must consider at least that little.

I recall that the magnetic moment of the extra-nuclear electron-system is very simply ascertained by measuring the magnetic splitting-up of the stationary states, i.e. the energy-differences between the different orientations of an atom in an applied magnetic field, because it enters directly into the formula for those energy-differences; but although the nucleus itself produces a further splitting-up which in its turn is measured, the situation is so much more complicated that these measurements have no interpretable bearing on the value of the nuclear magnetic moment. For protium and for deuterium, the two isotopes of hydrogen, the magnetic moment of the nucleus has been directly measured by a magnetic-deflection method. For all the other kinds of atoms we are obliged to infer it by theory from the measured values of the energy-differences between the states of what I called a cluster, which are alike in respect of I and J and differ in respect of the mutual inclination of these vectors.

The theory can at least be illustrated by a quasi-classical derivation, though the differences between this and the quantum-mechanical method are not slight. One first visualizes the valence-electron as a charged particle running around and around its orbit, equivalent therefore to a steady current running around the orbit and producing a magnetic field at all points within the orbit and in particular at the point occupied by the nucleus; the nuclear magnetic moment is subjected to this field, and when it is shifted from one to another of its permitted orientations a certain amount of work must be done (or received) and constitutes the energy-difference in question. Supposing a circular orbit with radius r and angular momentum p, the argument commences like that of page 290; we have $pe/2\pi mr^2c$ for the strength of the equivalent current, pe/mr^3c for the field-strength which it produces at the centre of the circle where the nucleus is; we conceive the nucleus as having a magnetic moment M parallel to its angular momentum; we assign the quantum-number I to this angular momentum and the quantum-number l to that of the orbital motion of the electron, thus conceiving these as vectors having the magnitudes $\sqrt{I(I+1)}$ $(h/2\pi)$ and $\sqrt{l(l+1)}$ $(h/2\pi)$, which last is what I have been calling p. If we could ignore the spin of the electron, l could be replaced by J, and the torque exerted by the field upon the nucleus would be $M(pe/mr^3c)\sin\theta_{I,J}$. There would be two or more permitted values of $\theta_{I,J}$ corresponding to the various states of the cluster, and we should get the corresponding energy-values U by writing:

$$U = M(pe/mr^3c)\cos\theta_{I,J} = M(e/mc)(r^{-3})\sqrt{J(J+1)}h/2\pi\cos\theta_{I,J}, \quad (17)$$

and using expressions based on equation (13) for the cosine. Finally we should form the differences between the right-hand members of these equations, and equate them to the observed energy-differences between the states of the cluster, and solve for M.

Even this formula, when applied to the data, gives values of M of the same order of magnitude as do the more elaborate ones; otherwise there would be no point in quoting it. There is, however, very much to be done to improve it. There is the magnetic field produced at the nucleus by the spin of the electron. There is the alteration required by relativity. There is the task of applying quantum-mechanical rather than quasi-classical reasoning to the postulates. The procedure is strongly supported by the fact that it is copied from the argument which, in the theory of the interaction between the spin and the orbital angular momentum of the valence-electron, leads to a wonderful explanation of the fine-structure of the hydrogen spectrum. It is, however, certainly not perfect, since when applied to different states of a particular kind of atom it is likely to lead to different values of the nuclear magnetic moment, a result which either shows some of the mathematical methods of approximation to be faulty or else is a *reductio ad absurdum* of one or more of the postulates. The problem is in fact one of the great unmastered problems of atomic physics, and some believe that it is wrong to postulate that the nucleus can be regarded, in its interactions with the extra-nuclear electrons, as nothing but a simple magnet attached to a body having mass and charge. I shall therefore say nothing further about it, except for quoting the formula oftenest used in cases such as those of sodium and hydrogen, where the energy-difference b in question (to follow the notation of page 308) is that between the two members of a pair of states for both of which $L = 0$ and $J = S = 1/2$, while for one of them $F = I - J$ and for the other $F = I + J$:

$$b = (8\pi/3)\left(\frac{2I+1}{I}\right)M(eh/4\pi mc)\Psi^2(0), \quad (18)$$

the last symbol standing for the square of the value which the Schroedinger wave-function has at the nucleus, which is known exactly for hydrogen and approximately for other one-valence-electron atoms; the formula is due to Fermi. Applying this formula to the values of b for light and heavy hydrogen which they had ascertained by the magnetic-deflection method (page 310), Rabi Kellogg and Zacharias

got values for the magnetic moments of proton and deuteron [12] which are of the order of one one-thousandth of those associated with electrons. These values are stated (for a reason which may be obvious but will be set forth presently) as 3.25 and as 0.75 times the quantity $(eh/4\pi m_p c)$, where m_p stands for the mass of the proton; the uncertainty is given as 10 percent for the former, about 25 percent for the latter.

There remain the experiments whereby M was directly determined, for protons and for deuterons, from the force exerted by an inhomogeneous magnetic field on the nuclear moment itself. One cannot use the bare proton for such an experiment, since owing to its charge it would suffer, as it flies through the magnetic field, an electrodynamic force and a deflection by comparison with which the others would be trivial. One cannot use the isolated hydrogen atom, since just as in the case of sodium which we considered at such length, the force exerted by the field upon the magnetic moment of its electron would far outweigh that exerted on the nucleus. There remains the hydrogen molecule, which in its normal state has the convenient feature that the spins of its two electrons are oriented anti-parallel (in the loose sense of the term) and cancel one another out, while the angular momenta and the magnetic moments of their orbital motions likewise vanish. This seems to remove all the possible competitors to the nuclear moments, but there arises another which does not occur in individual atoms: the rotation of the molecule-as-a-whole, which has an angular momentum and a magnetic moment. This magnetic moment, however, is of the same order as those of the nuclei, and its contribution to the net magnetic moment of the nuclei can be estimated and subtracted from theirs. As for the nuclei, they may set themselves with their spins either parallel or antiparallel;[13] in the latter case their magnetic moments cancel one another, and observations on such molecules teach us only about the rotation, knowledge which is useful; in the other case their magnetic moments add, and the data of the experiment yield a value which is double the moment of the individual proton—or of the individual deuteron, according as the molecule is formed of two light atoms or two heavy atoms of hydrogen.

[12] This substitute for the names *deuton* and *diplon*, by which the nucleus of the H^2 atom (deuterium or "heavy hydrogen") has usually been known in America and England respectively, was recommended at a recent meeting of the American Physical Society by Dr. Urey, the discoverer of deuterium.

[13] This implies that nuclei conform to rules of quantization in direction relative to one another similar to those for electrons. This is true, and is superbly demonstrated by observations on band-spectra and by the chemical separation of the two kinds of hydrogen molecules ("ortho-hydrogen" and "para-hydrogen") here mentioned; but the story is much too long for this place.

The experiments were performed by Stern and his school. They are extremely delicate, owing to the excessively small value of the magnetic moment and hence of the deflection. Indeed, they might have been beyond instead of within the limits of the possible, were it not for two distinctive features of hydrogen: the just-mentioned doubling of the moment due to the parallelism of the spins of two nuclei, and the fact that a beam of slow-moving molecules can be utilized, because hydrogen does not have to be heated in a furnace in order to be vaporized, but will emerge as a molecular beam through a hole in the wall of a chamber no warmer, indeed very much colder, than the temperature of the room itself! For the magnetic moment of the proton, Frisch and Stern give 2.5 $(eh/4\pi m_p c)$ with an uncertainty of "at most 10 per cent"; for that of the deuteron, Estermann and Stern say that the value is between 0.5 and 1.0 times $(eh/4\pi m_p c)$.

We now turn to the ensemble of the estimates of nuclear angular momenta and of magnetic moments, and the laws and rules which they seem to obey.

Estimates of I have by now been made for some fifty-five elements; but this is not an adequate statement to make, for the nuclear angular momentum is one of those qualities—as I shall presently stress—which may vary from one isotope to another of a single element, and there are already several cases in which values of I have been reliably assigned to two or more different isotopes. The great majority are derived from optical analysis of hyperfine-structure; four or five from magnetic-deflection experiments; about ten from alternating intensities in band-spectra, most of these last being checked from the hyperfine structure of line-spectra.

The values are of very unequal merit, some being derived independently and concordantly from several different properties of hyperfine structure (pages 301–302), some being further sustained by deductions from band-spectra, while others are guesses based on a few rough observations of intensities or intervals. The unreliability of these last is of a type not to be described by giving a most-probable-value coupled with a probable error. A critical and analytical review of the lot by a neutral expert is badly needed.

The first thing which strikes the eye on viewing a tabulation is the immense preponderance of half-integer values of spin; but this is only one sign of the most important of the rules of nuclear momenta, which is one of the most important of the rules of nature. I recall that the "mass-number" A of an atom is the integer nearest to the value of its mass expressed in terms of one sixteenth the mass of the commonest oxygen atom as unit. The rule, then, is:

Atoms of odd mass-number have half-integer or "odd" spins; while among the atoms of even mass-number, two of the lightest have the full-integer or "even" spin of unity, while all the rest display no hyperfine-structure at all and thus probably have the "even" spin-value zero (though we must never forget the possibility that the lack of observable hyperfine-structure means that the magnetic moment is very small, the clusters of states therefore so compact that the spectroscope cannot resolve the hyperfine pattern, and the value of I therefore unascertainable).

Although not every kind of atom has yet been studied, the number of cases on which this rule is based is already so considerable that the discovery of an exception would be a sensation of the first order. Among the most striking exemplifications of the rule are those afforded by elements of many isotopes. Mercury, the outstanding example, displays a hyperfine-structure of wondrous complexity, which has been very successfully interpreted by assigning the value 3/2 of I to its odd isotope of mass-number 201 (Hg^{201}), the value 1/2 to its odd isotope Hg^{199}, and the value zero to its four principal even isotopes Hg^{198}, Hg^{200}, Hg^{202} and Hg^{204}; the lines attributed to these isotopes stand to one another in the intensity-ratios deduced from the relative abundances of the isotopes as measured by Aston. Cadmium has two odd isotopes for both of which $I = 1/2$, and several even ones exhibiting no hyperfine structure at all. Several elements have two isotopes, both of odd mass-number; in some of these cases both have the same value of I(e.g. gallium 3/2, rhenium 5/2), in other cases they differ (e.g. rubidium 3/2 and 5/2). The outstanding and very-certainly-known case of hydrogen is distinguished by the values 1/2 for the light and 1 for the heavy isotope.

(It may occur to the reader that if, say, four isotopes of mercury display no hyperfine-structure, then everything that I have heretofore said implies that their spectrum-lines and levels ought to coincide absolutely and be indistinguishable. These are, however, slightly separated, owing, it is presumed, to very slight differences in the fields surrounding nuclei of different isotopes even when they all belong to the same element and have the same nuclear charge and indistinguishable moments. The phenomenon, which is known as "isotope shift," is likely to be much studied in the theoretical physics of the near future).

As we go along the list of the atoms of odd mass-number from the lightest to the heaviest, the value 1/2 for I appears at the start; the value 3/2 at mass-number 6; 5/2 at $A = 35$; 7/2 at $A = 83$; 9/2 (the highest yet inferred) at $A = 93$. Further observations may change

these numbers somewhat, but it seems fairly assured that the higher values first appear later than the lower. Yet the lower and even the lowest appear repeatedly all through the list; for the five most massive atoms for which I is known, its values are 1/2 for Tl^{203}, Tl^{205}, and Pb^{207}, 9/2 for Bi^{209} and 3/2 for Pa^{231}. There is no sign of a periodic variation; the random fashion in which the display or the lack of detectable hyperfine structure are sprinkled among the elements would be sufficient proof that it is of nuclear origin, even were there no theory.

We recognize thus that I is a quality of nuclei which depends upon nuclear *mass* (not charge!) Moreover, everything connected with the concept of nuclear angular momentum—the behavior of the extranuclear momenta L and S and J of which it is an analogue, the nature of the phenomena which it is contrived to explain, the values of I deduced from these phenomena—all these things imply that I is a quantum-vector compounded out of the individual angular momenta of individual particles composing the nucleus, the quantum-numbers of these individual momenta being, like that of their resultant, integer multiples of 1/2. The very simplest model would consist, of course, of particles all having identical angular momenta of quantum-number 1/2.

Now there are two leading schemes for imagining a nucleus—let us say, of atomic number Z and mass-number A—as systems of minuter particles. According to the one, it consists of A protons and $(A - Z)$ negative electrons; according to the other, of Z protons and $(A - Z)$ neutrons.[11] It would be correct to say that the former was the leading scheme until about two years ago, the latter now; but as the facts about nuclear momenta and nuclear magnetic moments have had much to do with bringing on this change of favor, I will not take it for granted in advance.

If we take the nucleus to be a congeries of protons and negative electrons, then we *are* postulating a system of which all the particles are known to have spins of quantum-number 1/2—the simplest conceivable system, as I said; and we should still have in reserve the possibility of assigning orbital motions with orbital angular momenta to some or all of the particles, should it ever seem desirable. But if we take the nucleus to be a congeries of protons and neutrons, we are introducing particles of a kind for which the spin is unknown, and must be fixed by assumption. Since we can put 1/2 for the spin of the neutron, this affords little basis for choice.

[11] The fact that alpha-particles are often mentioned as constituents of nuclei is not in contradiction with this statement, since an alpha-particle is interpreted as 4 protons and 2 electrons by the one scheme, 2 protons and 2 neutrons by the other.

It is, however, an important point—a very important point—that the different schemes imply different numbers of particles for a given nucleus: $(2A - Z)$ by the former, and simply A by the latter. I have spoken of a notable rule contrasting the values of I for even mass-number with those for odd mass-number. Notice now that if we adopt the latter or proton-neutron scheme, the rule becomes: *The quantum-number of the nuclear angular momentum is a half-integer if the number of particles in the nucleus is odd, and a full-integer (if ascertainable at all) if that number is even.* With the proton-electron scheme, this could not be said.

This difference would give an advantage to the latter scheme, even if the model were no more specific. However, the rules for quantizing orientations of vectors of quantum-number 1/2 require that these shall set themselves either parallel or anti-parallel (in the loose sense) to one another. If I be built up out of such vectors, then necessarily an even number thereof implies a full-integer value and an odd number a half-integer value, and *vice versa*.

This argument for the proton-neutron scheme is therefore strong, though perhaps not so strong as it would be, were not the basis for the test narrowed down by one of the curious empirical rules of the world of atoms: it is found that mostly $(2A - Z)$ is even when A is even, and odd when A is odd. Fortunately there are some exceptions; the famous one is N^{14}, the chief isotope of nitrogen, for which it is certain (from alternating intensities in the band-spectrum) that I is full-integer (unity), whereas $(2A - Z)$ is odd but A is even. This is the only case of its kind, but there are something like ten in which $(2A - Z)$ is even but A is odd, and there is a hyperfine-structure believed to correspond to a half-integer value of I; this seems especially well established for two isotopes of tin and two of mercury. Another and very powerful argument from alternating intensities in band-spectra, unfortunately too long to be expounded here, supports the belief that the nucleus of N^{14} has an even and not an odd number of constituent particles. On the whole it is pretty likely that any nucleus-model providing an even number of particles for an even mass-number and an odd for an odd will always be preferred to any model not having this feature.

The field is now open for interpreting the observed values of I by compounding proton-spins and neutron-spins (or proton-spins and electron-spins), and trying to find reasons for the resultants which are observed. It seems natural enough to have unity for the deuteron (proton and neutron spins parallel), zero for He^4 and C^{12} and O^{16} (spins cancelling each other two by two); but farther along the list

of atoms there are curiosities enough to keep theorists busy, one guesses, for years. I leave this for the future, and close by speaking briefly of the magnetic moments, which by themselves suffice to show that the task will not be easy.

For the magnetic moments attributable to extra-nuclear electrons (or rather to their projections upon the direction of the angular momentum corresponding), we found by theory and by experiment values which we wrote as $g(e/2mc)\ \sqrt{n(n+1)}\ h/2\pi$; here m stands for the mass of the electron; g is 1 for the orbital motion of a single electron, is 2 for the spin of a single electron, and has calculable values not greater than a very few units for the extra-nuclear electron-system as a whole; n is the quantum-number of the associated angular momentum, and is 1/2 for the spin of a single electron and has various values for the other cases.

If now a proton be a simple particle with a spin 1/2, one would expect for its magnetic moment a value differing from that of the electron-spin only by the substitution of m_p the proton-mass for m, therefore about 1840 times smaller. The actual value (page 313) is about three times as great as the expected one, so that the analogy with the electron is good enough to predict the order of magnitude, but is not perfect. This is as surprising a discovery as any that has cropped up in the last several years in the field of atomic physics, and shows that even the supposedly elementary particles still have mysteries for us. The value for the deuteron is a good deal smaller, which with the proton-neutron scheme implies that the neutron has a magnetic moment pointing in the opposite direction to that of the proton. The spins of proton and neutron must, however, be parallel, else their resultant could not be unity, which is the value of I for the deuteron. This brings up a new point: I have not yet said whether magnetic moment and angular momentum should be visualized as parallel or anti-parallel, or better, whether the component of the former along the direction of the latter should be taken as positive or negative. The qualities of the deuteron indicate that whichever is the case with the proton, the opposite is the case with the neutron. Actually it is often possible to tell, from features of hyperfine-structure of which I have said little (cf. footnote 5, p. 302), which of the cases is realized for the nucleus as a whole. For the proton and the deuteron these features are unhappily either absent or inaccessible, and the question remains open. For most of the others which have been analyzed, the magnetic moment is said to be positive, meaning that it is of the sign which one expects, considering the nucleus as a whirling positive charge. For a few nuclei, however, it appears to be negative

Table of Nuclear Spins

This tabulation is presented *sous toutes réserves.* The values are of very different degrees of reliability, but it would be foolhardy for a non-specialist to grade them. They are collected from many sources, including tables by H. E. White (*Introduction to Atomic Spectra*, McGraw-Hill, 1934), by H. Schüler (*ZS. f. Physik*, 88, 323–335, 1934), by G. Beck (unpublished) and by R. F. Bacher (unpublished), and a number of papers which have appeared in the last fourteen months in *Die Naturwissenschaften*, the *Physical Review*, and other journals.

Values obtained by magnetic-deflection methods are marked R; those deduced from alternating intensities in band-spectra are labelled B; those marked H or not marked at all are inferred from optical observations on hyperfine-structure. Two values for a single mass-number signify either data compatible with both, or discordance between two authorities. X signifies that hyperfine-structure has been sought but not found, which may mean that the spin is zero, but on the other hand may mean that the magnetic moment is too small to produce an observable spread of the hyperfine pattern. When a value of zero for spin is deduced from band-spectra, it is not subject to that doubt.

A	Z	Element	I	A	Z	Element	I
1	1	H	1/2 B	119	50	Sn	1/2
2	1	H	1 B	121	51	Sb	5/2
4	2	He	0 B	123	51	Sb	5/2, 7/2
6	3	Li	X	127	53	I	5/2, 9/2
7	3	Li	3/2 B, H, R	129	54	Xe	1/2
9	4	Be	X	131	54	Xe	3/2
12	6	C	0 B	133	55	Cs	7/2 R, H
14	7	N	1 B	136	56	Ba	X
16	8	O	0 B	137	56	Ba	5/2
19	9	F	1/2	138	56	Ba	X
20	10	Ne	X	139	57	La	5/2, 7/2
22	10	Ne	X	141	59	Pr	5/2
23	11	Na	3/2 R, H	144	62	Sa	X
27	13	Al	1/2	147	62	Sa	$\neq 0$
31	15	P	1/2 B	148	62	Sa	X
32	16	S	X	149	62	Sa	$\neq 0$
35	17	Cl	5/2 B	150	62	Sa	X
39	19	K	3/2 B, R	151	63	Eu	$\neq 0$
41	19	K	$> 1/2$ R	152	62	Sa	$\neq 0$
40	20	Ca	X	153	63	Eu	$\neq 0$
45	21	Sc	7/2	159	65	Tb	$3/2, \geq 5/2$
51	23	V	$\geq 5/2$	165	67	Ho	$7/2, \geq 5/2$
55	25	Mn	5/2	169	69	Tu	1/2
59	27	Co	7/2	175	71	Lu	$5/2, \geq 3/2$
63	29	Cu	3/2	177	72	Hf	1/2, 3/2
65	29	Cu	3/2	178	72	Hf	X
67	30	Zn	3/2	179	72	Hf	1/2, 3/2
69	31	Ga	3/2	180	72	Hf	X
71	31	Ga	3/2	181	73	Ta	7/2
75	33	As	3/2	185	75	Re	5/2
79	35	Br	3/2, 5/2 B	187	75	Re	5/2
81	35	Br	3/2, 5/2 B	197	79	Au	3/2
83	36	Kr	7/2, 9/2	198	80	Hg	X
85	37	Rb	5/2	199	80	Hg	1/2
87	37	Rb	3/2	200	80	Hg	X
87	38	Sr	3/2?	201	80	Hg	3/2
89	39	Y	1/2	202	80	Hg	X
93	41	Cb	9/2	203	81	Tl	1/2
103	45	Rh	X	204	80	Hg	X
110	47	Cd	X	204	82	Pb	X
111	47	Cd	1/2	205	81	Tl	1/2
112	47	Cd	X	206	82	Pb	X
113	47	Cd	1/2	207	82	Pb	1/2
114	47	Cd	X	208	82	Pb	X
116	47	Cd	X	209	83	Bi	9/2
115	49	In	9/2	231	91	Pa	3/2
117	50	Sn	1/2				

Table of Nuclear Magnetic Moments

The values are expressed in terms of the unit $(eh/4\pi m_p c)$ customary in such tables (though for no very good reason) and sometimes called the "nuclear magneton." They are collected from the same sources as those in the Table of Nuclear Spins, and reduced to at most two significant figures; in cases of discrepancy, Schüler's value is the one usually taken. Those marked S are the ones obtained by Stern and his school; those marked R are obtained by Rabi and his school; those marked H or not at all are deduced from optical observations on hyperfine structure. I have heard the uncertainty of these last put at twenty per cent by one expert in the field, but some authorities would set it even higher. For the isotopes of hydrogen, the sign is unknown.

A	Z	Element		A	Z	Element	
1	1	H	2.5 S	93	41	Cb	3.5
			3.2 R	111	48	Cd	− 0.5
2	1	H	0.5–1.0 S	113	48	Cd	− 0.5
			0.77 R	115	49	In	5.3
7	3	Li	3.3 H, R	117	50	Sn	− 0.9
19	9	F	2.4	119	50	Sn	− 0.9
23	11	Na	2.1	121	51	Sb	2.7
27	13	Al	1.9	123	51	Sb	2.1
39	19	K	0.4 H, R	129	54	Xe	− 1
45	21	Sc	3.5	133	55	Cs	2.5
59	27	Co	2.8	137	56	Ba	1.0
63	29	Cu	2.7	197	79	Au	0.2
65	29	Cu	2.7	199	80	Hg	0.5
69	31	Ga	2.1	201	80	Hg	− 0.6
71	31	Ga	2.7	203	81	Tl	1.5
75	33	As	0.9	205	81	Tl	1.5
83	36	Kr	− 1.0	207	82	Pb	0.5
85	37	Rb	1.5	209	83	Bi	3.6
87	37	Rb	3.1				

The ratio of the magnetic moments of two isotopes of a single element may sometimes be calculated from hyperfine-structure with less uncertainty of theory than the value of either moment separately. For such ratios the following values are available:

$Cu^{63}/Cu^{65} = 1.00$; $Ga^{69}/Ga^{71} = 1.27$; $Rb^{85}/Rb^{87} = 2.04$; $Cd^{111}/Cd^{113} = 1$;
$Sn^{117}/Sn^{119} = 1$; $Sb^{121}/Sb^{123} = 1.36$; $Hg^{199}/Hg^{201} = -1.11$; $Tl^{203}/Tl^{205} = 1.02$.

The experiments of Rabi's school and those of Stern's indicate a value of about *four* for the ratio of the moments of proton and deuteron; this is substantiated by L. and A. Farkas (*Nature*, **135**, 372; 9 March, 1935) who measure the rates of the reactions whereby ortho-hydrogen and ortho-deuterium transform themselves into the para-forms (and reversely) in the presence of oxygen; from measurements made at three different temperatures they get for the ratio the three values 3.85, 4.03, 4.07, which agree (they say) "within the limits of the experimental error, which is less than five per cent."

(it is perhaps significant that all of these have even atomic numbers and odd mass-numbers).[15] As for its numerical magnitude, I recall that the insufficiencies of the theory render the recorded values—of which there are some twenty-five—subject to much doubt. Such as they are, it is a striking fact that none of them is more than a couple of times as great as that of the proton, and some are a good deal smaller. This injures still more the scheme of constructing nucleus-models out of electrons and protons, for the magnetic moment of a single electron not compensated by some other should of itself make the nuclear moment enormously larger. Yet if we adopt the proton and the neutron as the sole constituents of nuclei for the sake of banishing this difficulty, it will return to haunt us so soon as we attempt to reduce the number of elementary particles by conceiving either the neutron as a proton united with a negative electron, or the proton as a neutron with which a positive electron is combined.

Acknowledgments

I am much indebted to Professor I. I. Rabi and Dr. R. F. Bacher for having read over the manuscript of this article and having made many helpful comments, and to Professors G. Beck, S. Goudsmit and A. E. Ruark for correspondence.

[15] Two isotopes of Cd (A = 111 and 113), two of Sn (117 and 119), one of Hg (201), one of Kr (83), one of Xe (129).

Ferromagnetic Distortion of a Two-Frequency Wave

By ROBERT M. KALB and WILLIAM R. BENNETT

Frequency components are found for the ferromagnetic induction produced by a small magnetizing force of two incommensurable frequencies. Because of hysteresis the results depend intimately upon the ratios of these frequencies and of their amplitudes. With these ratios as criteria, two solutions are provided, adequate for most modulation problems of this character occurring in the field of communications.

The development is based on Madelung's empirical propositions. From these are deduced the forms of complex hysteresis loops occasioned by two-frequency magnetomotive forces, and from the loops sinusoidal components of the flux wave are derived by means of Fourier's series. The various voltages generated in a coil by such a flux are then calculated and next correlated with analyses for a single applied frequency. The resulting changes in the impedances to the two fundamental frequencies are also evaluated. The most important results are given in graphs and tables.

Experimental data on a number of specimens show close agreement with curves computed by the theory.

The analysis discloses several interesting features. It is shown that Madelung's conclusions imply Rayleigh's law of loop similarity; as a consequence the parameters of a Rayleigh loop suffice to describe a complex loop to the extent that it conforms to Madelung's results. Hysteresis suppression is found not to occur at low fields, although harmonic suppression may. The generated side frequencies of the flux appear in unequal pairs, the lower one being the stronger in each instance. Such inequality is a general property ascribable to the multivaluedness of the loop.

FOR precisely evaluating the performance of communication circuits containing ferromagnetic materials, methods for taking into account the non-linear effects of these materials are needed. To this end there have been devoted certain investigations of the behavior of such materials at the low flux densities usual in communication. These early disclosed that hysteresis is a governing factor for weak fields and led to attempts to solve the problem of its bearing on speech.

The complexity of speech and of hysteretic phenomena has made desirable the use of simple testing methods, which in turn require for their interpretation a quantitative theory. Since tests of this sort are usually made with one or more sinusoidal test waves, a theory of single-frequency magnetic performance has already been evolved as a first step toward fulfilling this need. For many purposes single-frequency tests are inadequate, and two-frequency waves are often used to obtain better information bearing on the design or performance of communication systems. It is the purpose of the present paper to take a further step by furnishing the theory of magnetic behavior under a two-frequency force.

The very simplicity which sometimes makes sinusoidal waves valuable for analyzing or testing the non-linear properties of channels of communication makes such waves worthless when applied in other instances, and more complex test waves must then be employed. The harmonics produced by a sine wave furnish an index of the distorting properties of a system, but the side frequencies produced by two such waves are needed to indicate its modulating properties or to give a measure of the interference between carrier channels. When two waves are used, one may be thought of as the carrier, modulated by the other, whose amplitude is chosen proportionate to the square root of the energy in the more complex modulating wave it represents, or both may be thought of as component waves in the same channel, or as carriers in different channels, producing interference in certain other channels. The amplitudes of the several product frequencies then give a measure of the energy falling in their respective regions of the spectrum under actual operation. The effect of the presence of one fundamental upon the transmission of the other can also be ascertained. Increasing the complexity of the test wave by the superposition of additional frequencies can be seen to afford little added advantage at the cost of much complication, unless the character of the waves actually transmitted is simulated, in which case statistical methods of study can perhaps be applied.

From the foregoing circumstances the utility of information pertaining to the application of two-frequency inputs as well as single-frequency inputs to non-linear circuit elements is apparent; many investigations have been conducted in this field to provide such information. When the current-voltage relation is not single-valued a more intricate treatment is necessary in carrying out the analysis. A general method of attack for double-valued characteristics has been provided and applied to hysteresis loops by E. Peterson [1] to determine the flux in ferromagnetic materials under single-frequency magnetizing forces. The fundamental dependence of loop form upon wave shape precludes immediate extension of Peterson's results to the case of a multi-frequency force except for certain harmonic combinations, one of which he considers.[2] A study of flutter effect has been published by Walter Deutschmann,[3] who analyzed a complex loop made up of straight lines. Both instances serve to emphasize the desirability of a broader investigation of the theoretical aspects of two-frequency magnetization including hysteresis. While no general and rigorous

[1] *B. S. T. J.*, Vol. 7, pp. 762–796, Oct. 1928.
[2] *Ibid.*, p. 773.
[3] *Wiss. Ver.* a.d. Siemens-Konzern, Vol. 8, No. 2, pp. 22–44, 1929; *E. N. T.*, Vol. 6, pp. 80–86, Feb. 1929.

method has yet been developed for handling this problem in a manner analogous to that applicable for a single frequency, practically important cases are solved here by means extensible to more complicated ones.

Considerations of Wave Shape and Loop Forms

Scope of the Two-Frequency Analysis

The envelope of a wave affords a means for classification. Waves whose envelopes change gradually during any oscillation form a class apart from those waves which have envelopes subject to abrupt changes. Members of each class can be segregated into those with envelopes nearly symmetrical with respect to the average magnetizing force, those with envelopes of almost uniform width, and so on. One of the two last-mentioned properties in a wave with a gradually changing envelope is essential to successful analysis by the methods about to be detailed.

The cases of magnetization analyzed, which include all the two-frequency wave shapes that can qualify under the foregoing criterion of tractability to analysis, are the following:

Case 1. The ratio between the geometric and arithmetic means of the amplitudes much smaller than the ratio between the sum and difference of the frequencies.

Case 2. One fundamental frequency high relative to the other; the product of the higher frequency with its amplitude large relative to the product of the lower frequency with its amplitude.

Between the two cases there exist intermediate ratios of frequencies and amplitudes over which the theory does not extend; however, the most frequent problems are usually entirely within the domain of a single case. The inequalities involved in the foregoing case definitions are not susceptible to simple explicit statement as limiting numerical ratios, in advance of the development of the theory. Much depends upon the accuracy required in predicting performance. From data supplied in the paper, following the theory, it is possible to determine the practical limitations of the mathematical treatment.

Formation of Complex Loops

Considering the complex hysteresis loops arising from multi-frequency magnetizing forces to be many-valued characteristics for determining the flux density, it is pertinent to study their formation and to correlate their parameters in so far as may be possible with those of single-frequency loops. The latter at low fields are known

to consist approximately of two parabolic branches, the exact shape of each dependent upon its point of origin. For fields confined somewhat below maximum permeability, the situation customarily obtaining in communication circuits, this representation has proved to be sufficiently exact to warrant the neglect of higher order terms in analyses. In conformity, terms of higher than second order will not be retained in equations used here; the development so presented can be extended to include them without other change in procedure. The induction at any point on a simple loop centered on the origin is expressed as a function of the instantaneous magnetizing force h by means of a formula developed by Peterson,

$$B = (a_{10} + a_{11}H)h \pm a_{02}(H^2 - h^2). \tag{1}$$

The upper sign of the double sign is used for the descending (upper) branch of the loop, and the lower sign for the ascending (lower) branch. H is the maximum magnetizing force and the coefficients are constants of the ferromagnetic material, determinable by single-frequency measurements. They have the following significance: a_{10} is the initial permeability, a_{11} the rate of change of permeability with magnetizing force, and a_{02} a factor of proportionality between the hysteresis loss and the cube of the maximum magnetizing force. The concepts in terms of which these parameters are defined acquire extended meanings for complex loops.

In the absence of an adequate theory of ferromagnetism the question of whether branches of complex loops and of simple loops have similar forms must be answered by experiment. The steady state of retracing alternately the two branches of a simple loop may eventuate in a different relation of B versus h than results from the first cycle; such a condition would mean that transient branches compose the complex loop, inasmuch as it is not retraced. It is also possible that the biasing effect of one sinusoidal component of the magnetizing force upon the other might cause the branches of the two types of loops to be dissimilar. The coefficients which specify the branches of the simple loop are evaluated with it centered at the origin of the B–h plane, using single-frequency methods, and cannot be assumed *a priori* to apply to to other situations, or to an unrepeated branch.

According to experiments by R. Goldschmidt [4] the superposed field necessary to cause much change of either the shapes or axial slopes of loops exceeds the weak fields to which this development is limited. Likewise, Lord Rayleigh [5] in his original investigation found small super-

[4] *Zeits. f. Techn. Physik*, Vol. 11, pp. 8–12, 1930.
[5] *Phil. Mag.*, Vol. 23, 1887.

posed fields to have no measurable effect upon the form of the loops for the specimens he investigated. Based upon these results, the branches of simple loops can be taken to be independent of their location in the B–h plane, and in so far as the complex magnetizing force gives rise to branches of simple loops, they have the same shape they would have in such a loop centered at the origin.

The formation of complex loops has been determined experimentally by E. Madelung.[6] He found that after reversing the magnetizing force at any point along the branch of a hysteresis loop a new curve is traced which, if continued, passes through the tip of the loop. A second reversal before the tip is reached causes another new curve to be traced back to the point of reversal on the original branch, which is followed thenceforth as if the two reversals have not occurred. The return to a reversal point makes all subsequent traces of the loop the same as if no changes of magnetizing force intervened between the two transits through that point. Any branch of the complex loop is then, in accordance with Madelung's determinations, completely specified by two points of reversal—the one from which it starts and the one through which it must pass if continued far enough; after passing the latter point it becomes the continuation of another branch similarly specified by different points of reversal.

The foregoing principles furnish sufficient information for deducing the form of the branches of complex loops. If such a branch be extended to one of the reversal points defining it and a trace then be carried back to the other, the loop so formed will be retraced by repeating the cycle. As these repetitions can be carried on indefinitely, the path must comprise a simple loop, the branch of the complex loop forming a portion of it. Every complex loop can therefore be considered as composed of adjoined sections of simple loops. Each branch of the complex loop is representable on suitably transformed axes by formula (1) with H taken as half the change in magnetizing force between the reversal points which specify the branch. In general, a different pair of axes will be required for each branch; they can subsequently be referred to a common origin.

The application of this analysis to the complex loop discloses the requirement that the relation $a_{11} = 2a_{02}$ must be true if Madelung's propositions are to hold, because the values of H differ for the two branches of a subsidiary loop. If this equality is not satisfied, the return to the original branch does not take place at the point of departure. Madelung's observations do not include this possibility,

[6] *Annalen der Physik*, Vol. 17, pp. 861–890, 1905. See also *Handbuch der Physik*, Vol. 15, pp. 106–107, Berlin, 1927.

and sufficient evidence of behavior for extending them in such circumstances is not available at present. Some experiments by Lehde [7] and others indicate that subsidiary loops do not quite close at their junctions with the major loop and often show departures both ways on different parts of the same complex loop. This sort of behavior is not explainable by inequality between a_{11} and $2a_{02}$ for the subsidiary loops, as it would cause the departure of one specimen to be always the same way depending upon which quantity was the larger. In all cases, even near saturation, Lehde's results show this departure to be small and the connecting branches between successive subsidiary loops to form approximately a simple loop. On this basis Madelung's results can be considered confirmed to a sufficient degree of approximation.

The ratio a_{11}/a_{02}, which has been taken as a measure of the validity of Rayleigh's relation in single-frequency theory, becomes a criterion of the usefulness of Madelung's propositions concerning loop form in multi-frequency theory. Those substances which most closely accord with Rayleigh's analysis can also be expected to be in best agreement with Madelung's results. The relation between coefficients required on the basis of Madelung's and Rayleigh's experiments will be used hereafter to simplify the analysis. The simplification will be evidenced by the customary nomenclature, in which

$$\mu_0 \equiv a_{10}, \quad \nu \equiv a_{02} = \tfrac{1}{2}a_{11}.$$

Any attempt to distinguish here between the two latter constants would be meaningless because beyond the scope of Madelung's empirical rules. Fortunately the constants of most commercial materials conform closely to the above equality.

Types of Two-Frequency Loops

The aspect of a hysteresis loop formed by a two-frequency wave changes greatly with the frequencies and their amplitudes. Different pairs of frequencies having equal ratios give rise to families of loops which are identical except as affected by eddy currents. These, for the purpose of this study, are supposed to be so small that the flux is substantially uniform over a cross-section of the magnetic circuit. If the two frequencies have a common source or are synchronized, the hysteretic phenomena are singly periodic and subject to simpler treatment than developed here for independent sources.

For detailed analysis of the effects of hysteresis with two applied frequencies from independent sources, the phase angles of both may

[7] *Rev. of Sci. Instr.*, Vol. 2, pp. 16–43, Jan. 1931.

be taken as zero, equivalent to measuring the time from a point where the two components of the magnetizing force become maximum simultaneously. Then

$$h = P \cos pt + Q \cos qt \tag{2}$$

is the instantaneous magnetizing force. Phase angles can be made arbitrary by replacing pt and qt by $pt + \theta_p$ and $qt + \theta_q$, respectively, in this equation and the subsequent results.

The configurations of complex loops can be followed by altering single-frequency loops to accord with the results of investigations of loop formation. If to a single-frequency magnetizing force a relatively small one of slightly lower frequency is added, the resultant is an oscillatory force whose peaks undulate around the value reached by those of the original wave. Their maximum will be the sum of the amplitudes of the two components and their minimum the difference. On a hysteresis loop (Fig. 1*a*) this means that portions between successive reversal points will differ slightly from one another, being formed approximately as if belonging to successively smaller loops until a minimum peak is passed, thenceforth as if belonging to successively larger loops, and so on cyclically. Such behavior is sketched in the figure.

As the amplitude of the lower frequency component of the magnetizing force is increased the undulations become more pronounced and the preceding picture more inexact. When both amplitudes are equal the envelope of the resultant magnetizing force vanishes periodically and the portion of the hysteresis loop formed while this envelope goes from its maximum to zero is in the nature of a spiral (Fig. 1*b*), a similar curve being developed outwardly as the envelope increases again to its maximum. Provided only that successive peaks of the magnetizing force do not differ greatly in magnitude, each portion of such a loop between adjacent reversal points may be assumed to have the form of a branch of a single-frequency hysteresis loop having its point of origin coincident with that of the portion of the complex loop. Then the induction may be derived from the magnetizing force by the use of single-frequency data in accordance with the known manner of formation of complex loops.

If now the amplitude of the higher frequency component be decreased to a relatively small value, the undulations in the envelope subside, and a condition similar to the original one is seen to obtain. This time, however, the amplitude of the lower frequency will be found to be the one about which these undulations occur, and the characteristic will again look like that in Fig. 1*a*.

As long as one frequency is less than twice the other the undulations in the peaks of the magnetizing force will be regular and gradual. If the higher frequency be raised to more than twice the lower, the undulations become more abrupt and complex, and increasingly so

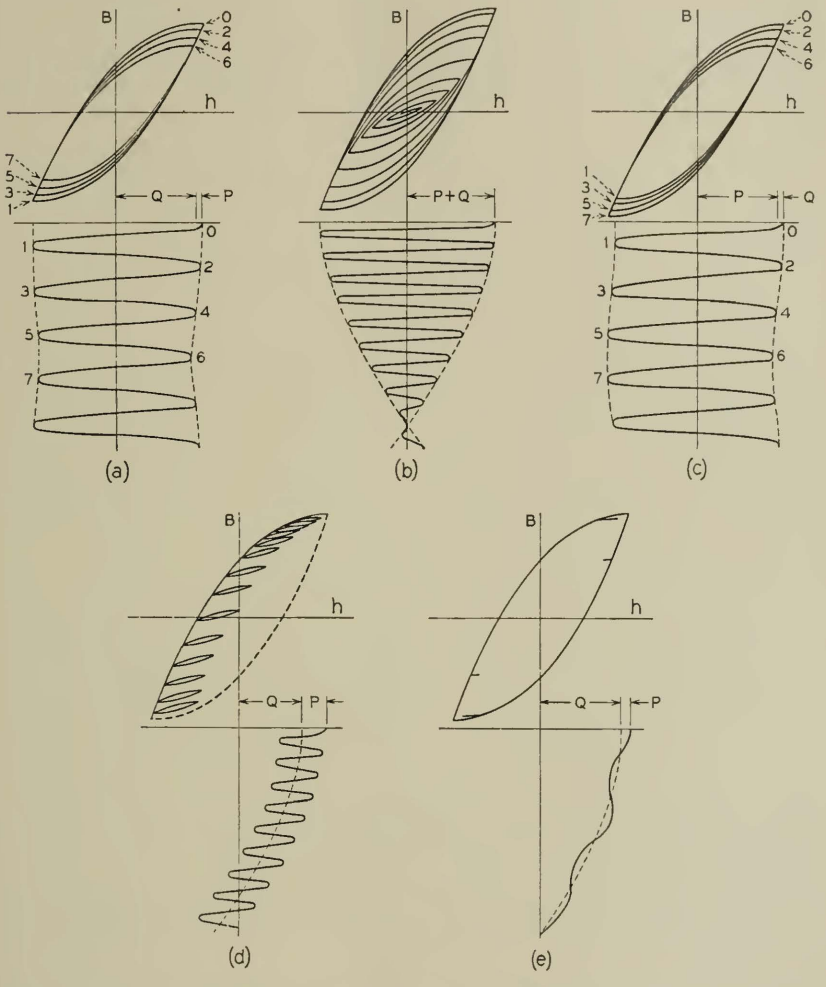


Fig. 1—Types of hysteresis loops characteristic of a two-frequency magnetizing force.

as it is raised still more. The complexity attendant upon the formation of the hysteresis loops becomes greater and the simplifying artifices heretofore suggested no longer apply.

When one frequency has become several times the other, the higher

frequency component must have one of its maxima very near each maximum of the lower one so that a maximum value practically equal to the sum of the amplitudes of the two components is reached every cycle of the lower one; likewise for the minima. Between these two extremes other reversals in the magnetizing force would be expected to cause the hysteresis loop to comprise additional small loops not necessarily closed. Experiments confirm this conjecture and indicate that not only for weak fields but even for fields near saturation the small loops nearly close and the paths between them are traced nearly the same as portions of a large loop.

If the amplitude of the higher frequency component is considerably the larger, all the loops composing the characteristic are virtually the same size and shifted slightly with respect to each other on account of the lower frequency component. The characteristic will be that depicted in Fig. 1*c*.

When the amplitudes of the two components are not grossly unequal, the hysteresis loop is of the type represented by Fig. 1*d*. Small loops formed when the magnetizing force is near an extreme value are longer than those formed when it is near zero, aside from any effects of superposition, because of its different rates of change in the two positions. In general these loops will not occur in the same place for different cycles and the distances between them will be lessened by increasing the higher frequency. By considering the complex loop to consist of a major loop, such as a single frequency would generate, encompassing a number of minor loops, the induction may be derived from the magnetizing force since the form of the minor loops is known. When these are not too widely spaced, each may be assigned a mean position in a loop fixed for all time and the induction calculated therefrom. It is evident that such an undertaking is vastly more complicated than the ones suggested heretofore and that unlike them it requires information in addition to that obtained from single-frequency measurements performed with no superposed field.

By decreasing the amplitude of the higher frequency component of the magnetizing force, the amplitudes of the minor loops may be reduced until those in the neighborhood of zero magnetizing force vanish entirely. Further decrease of the same component causes more and more of the minor loops to disappear, so that finally only a few small ones remain at each end of the major loop. This condition is shown in Fig. 1*e*. As the wave form of the induction is only slightly affected by the presence of these loops, they may safely be omitted and the characteristic simply taken as the major loop, determined from single-frequency results alone.

The point where the minor loops first vanish and the requirement that they vanish at all may be readily determined. Two adjacent extremes of the magnetizing force approach a common value as the amplitude of the higher frequency component is decreased. When they attain this value, the slope is no longer reversed between them and consequently no minor loop is formed. The instantaneous magnetizing force is expressed by equation (2), and in this instance $p \gg q$. The minor loops first appear where $dh/dt = 0$ when $pt \doteq -\frac{\pi}{2}$. Solved simultaneously, these equations yield

$$qt = \sin^{-1} \frac{Pp}{Qq}, \tag{3}$$

where the minor loops vanish. They reappear at an equal negative value of $\sin qt$, and other intervals during which they vanish are apparent from symmetry. If $(Pp/Qq) > 1$, no real solution of equation (3) exists, so the minor loops do not vanish anywhere. The appearance and non-appearance of minor loops is seen to be governed by the ratio of amplitudes in the same way as by the ratio of frequencies as long as the restriction on the latter is observed, and the product of these ratios

$$\kappa = \frac{Qq}{Pp}$$

determines the type of hysteresis loop (1*d* or 1*e*, or an intermediate form) obtained.

Induction with a Two-Frequency Magnetizing Force

General Expression for the Induction

As a function of time, the induction for any type of loop described will consist of intermodulation products of the two fundamental frequencies. Because of the kind of symmetry the characteristic has these products will include only the odd orders, but because it is a loop, quadrature terms must be expected. The induction then will ultimately be in the form

$$B = \sum_{m=0}^{\infty} \sum_{n-\infty}^{\infty} [c_{mn} \sin (mp + nq)t + d_{mn} \cos (mp + nq)t] \tag{4}$$

with all even order coefficients zero. The odd order coefficients remain to be determined from the hysteresis loop. In doing this only third order products in addition to the fundamentals will be evaluated explicitly, as they are stronger than the higher orders and therefore

of most importance in measurements of distortion. Any higher order products would be less precisely evaluated, more numerous, and probably of less interest; their computation is clearly evident as an extension of the processes later carried out.

The Multi-Branched Hysteresis Loop

Under certain circumstances mentioned earlier, the portions of a complex loop between adjacent reversal points are representable by

$$B = (\mu_0 + 2\nu H)h \pm \nu(H^2 - h^2). \tag{5}$$

Referred to the origin of the B–h plane, the induction on such a portion originating at the jth reversal point from $t = 0$ is

$$B_j = G_j + \mu_0(h - H_j) - (-1)^j\nu(h - H_j)^2. \tag{6}$$

Here H_j is the magnetizing force and G_j the induction at the jth reversal point. The latter quantity satisfies the difference equation

$$G_j = G_{j-1} + \mu_0(H_j - H_{j-1}) + (-1)^j\nu(H_j - H_{j-1})^2,$$

arrived at by evaluating the induction on the $(j - 1)$st branch at the jth reversal point. Subject to the initial condition

$$G_0 = (\mu_0 + 2\nu H_0)H_0,$$

this can be solved by the method of successive substitutions; the solution is

$$G_j = \mu_0 H_0 + 2\nu H_0^2 + \mu_0 \sum_{i=1}^{j} (H_i - H_{i-1}) + \nu \sum_{i=1}^{j} (-1)^i(H_i - H_{i-1})^2. \tag{7}$$

The foregoing expressions define the induction everywhere on the complex loop. Equations (7) and (6) combined to eliminate G_j give

$$B_j = \mu_0 h + (-1)^j 2\nu H_j h + (-1)^j\nu(H_j^2 - h^2). \tag{8}$$

The problem remaining is to develop this equation into the equivalent of equation (4).

The instantaneous magnetizing force is

$$h = P \cos pt + Q \cos qt. \tag{2}$$

By a trigonometric transformation this may be put in the form

$$h = \sqrt{(P+Q)^2 - 4PQ \sin^2 \frac{p-q}{2} t} \times \cos\left[\frac{p+q}{2} t + \tan^{-1}\left(\frac{P-Q}{P+Q} \tan \frac{p-q}{2} t\right)\right], \quad (9)$$

which is sometimes more convenient. The envelope of the wave is represented by the two branches of the radical. If its magnitude does not change much between adjacent maxima and minima of the wave, these extremes lie close to the points of tangency between the wave and its envelope; the latter condition is the one necessary in order that the envelope may be used to evaluate the extreme magnetizing force H_j.

This force acquires its values at the reversal points, which are situated at the zeros of dh/dt. Put into a form like equation (9) by the same transformations as were used above, this derivative is

$$\frac{dh}{dt} = -\sqrt{(Pp+Qq)^2 - 4PpQq \sin^2 st} \times \sin\left[rt + \tan^{-1}\left(\frac{Pp-Qq}{Pp+Qq} \tan st\right)\right], \quad (10)$$

where $2r = p + q$, $2s = p - q$. Except when $\kappa = 1$, this vanishes only at

$$rt + \tan^{-1}\left[\frac{1-\kappa}{1+\kappa} \tan st\right] = j\pi, \quad (11)$$

j integral or null. Substituting equation (11) into equation (9) yields the magnetizing force at the jth reversal from $t = 0$:

$$H_j = \sqrt{(P+Q)^2 - 4PQ \sin^2 st} \times \cos\left[j\pi + \tan^{-1}\left(\frac{1-k}{1+k} \tan st\right) - \tan^{-1}\left(\frac{1-\kappa}{1+\kappa} \tan st\right)\right],$$

letting $k = Q/P$. Upon combining the arc tangents this becomes

$$H_j = P\sqrt{(1+k)^2 - 4k \sin^2 st} \times \left[\cos j\pi - \tan^{-1} \frac{2(k-\kappa) \tan st}{(1+k)(1+\kappa) + (1-k)(1-\kappa) \tan^2 st}\right]. \quad (12)$$

According to equation (12) the magnetizing force and its envelope are tangent at reversal points provided the arc tangent is always zero. This it is if $k = \kappa$, a trivial solution. By certain choices of these two parameters, however, it is possible to keep the angle between any reversal point and the nearest extreme of the magnetizing force from exceeding any prescribed limit. The maximum value of the angle in

equation (12) is

$$2 \tan^{-1} \sqrt{\frac{(1-k)(1+\kappa)}{(1+k)(1-\kappa)}} - \frac{\pi}{2},$$

which can be made small by making k and κ each small compared to unity, or each large compared to unity, or both approximately equal. For the previously excluded instance $\kappa = 1$, this angle can be limited and the last condition fulfilled by keeping k nearly equal to one, as is obvious from the equations. So for each of these three conditions on the parameters, to a definite degree of approximation the envelope at each point of tangency becomes the magnetizing force for the nearest reversal point, a feature useful for the transformation of equation (8) into a function of time.

Calculation of the Induction—Case 1

The three conditions are confluent and will be seen to set the limiting bounds for case 1. When the fundamental frequencies lie close together the ratio of their amplitudes is practically unrestricted; as more widely spaced frequencies are chosen it becomes necessary to require an increased (or diminished) amplitude ratio in order that the phase angle in equation (12) does not exceed the chosen limit. This limit must be such that the cosine of that phase angle is substantially unity.

The maximum magnetizing force is thereupon

$$H_i = (-1)^i P \sqrt{(1+k)^2 - 4k \sin^2 st}\,. \tag{13}$$

As a periodic even function of st, H_i may be expanded in a Fourier's series

$$H_i = (-1)^i (P+Q) \left[\frac{A_0}{2} + \sum_{\eta=1}^{\infty} A_\eta \cos \eta st \right], \tag{14}$$

where

$$A_\eta = \frac{4}{\pi} \int_0^{\pi/2} \sqrt{1 - k_1^2 \sin^2 \lambda} \cos \eta\lambda d\lambda \tag{15}$$

in terms of the parameter

$$k_1 = \frac{2\sqrt{PQ}}{(P+Q)} = \frac{2\sqrt{k}}{(1+k)} = \frac{2\sqrt{1/k}}{(1+1/k)},$$

which never exceeds unity and which diminishes as k is either increased or decreased from unity. The integral (15) reduces immediately to elliptic form by the substitution $z = \sin \lambda$. All odd order coefficients are zero. For the first three significant values of η:

$$\left.\begin{aligned} A_0 &= \frac{4}{\pi} E_1 \\ A_2 &= \frac{4}{3\pi k_1^2}[(2 - k_1^2)E_1 - 2(1 - k_1^2)K_1] \\ A_4 &= \frac{4}{15\pi k_1^4}[8(2 - k_1^2)(1 - k_1^2)K_1 - (16 - 16k_1^2 + k_1^4)E_1] \end{aligned}\right\} \quad (16)$$

Here K_1 and E_1 are complete elliptic integrals of the first and second kind, respectively, with modulus k_1.

The series (14) may be used to evaluate the variable permeability anywhere on the loop, for upon substitution in equation (8) reference to a particular branch is eliminated by the disappearance of the double sign on the second term.

The square of the maximum magnetizing force needed in the final term of equation (8) comes directly from equation (13); to determine the sign of this term at any instant remains the only problem. Interpretation of $(-1)^j$ seems simple when it is remembered to be positive for decreasing h and negative for increasing h, and therefore an odd function of time. The rate of change of the magnetizing force is

$$\frac{dh}{dt} = -Pp \sin pt - Qq \sin qt,$$

so it follows that

$$\left.\begin{aligned} (-1)^j &= +1, \quad \sin pt + \kappa \sin qt > 0 \\ &= -1, \quad \sin pt + \kappa \sin qt < 0 \end{aligned}\right\}. \quad (17)$$

The solution may be completed by expanding this quantity in a Fourier's series:[8]

$$(-1)^j = \sum_{m=0}^{\infty} \sum_{n=-\infty}^{\infty} A_{mn} \sin (mp + nq)t. \quad (18)$$

When $m = 0$ the summation is to be extended over only positive values of n. With this convention the coefficients are

$$A_{mn} = \frac{1}{2\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} (-1)^j \sin (mx + ny) dx\, dy \quad (19)$$

or, with the use of equation (17),

$$A_{mn} = (-1)^{\frac{m+n-1}{2}} \frac{4}{\pi^2} \int_0^{\pi} dy \int_0^{\cos^{-1}(-\kappa \cos y)} \cos mx \cos ny\, dx, \quad (20)$$

[8] W. R. Bennett, "New Results in the Calculation of Modulation Products," *B. S. T. J.*, Vol. 12, pp. 228–243, April, 1933.

where p and q are to be so assigned that $\kappa \leq 1$. Then the coefficients are all expressible in terms of complete elliptic integrals with modulus κ; these will be designated as K_2 and E_2. Coefficients of even order vanish, while those of the first three odd orders are found to be

$$
\begin{aligned}
A_{10} &= \frac{8}{\pi^2} E_2 \\
A_{01} &= \frac{8}{\pi^2 \kappa} [E_2 - (1 - \kappa^2) K_2] \\
A_{2\underline{1}} = -A_{21} &= \frac{8}{3\pi^2 \kappa} [(1 - 2\kappa^2) E_2 - (1 - \kappa^2) K_2] \\
A_{1\underline{2}} = A_{12} &= \frac{8}{3\pi^2 \kappa^2} [(2 - \kappa^2) E_2 - 2(1 - \kappa^2) K_2] \\
A_{30} &= \frac{8}{9\pi^2} [(7 - 8\kappa^2) E_2 - 4(1 - \kappa^2) K_2] \\
A_{03} &= \frac{8}{9\pi^2 \kappa^3} [(8 - 3\kappa^2)(1 - \kappa^2) K_2 - (8 - 7\kappa^2) E_2] \\
A_{4\underline{1}} = -A_{41} &= \frac{8}{15\pi^2 \kappa} [(1 - 16\kappa^2 + 16\kappa^4) E_2 \\
&\qquad - (1 - 8\kappa^2)(1 - \kappa^2) K_2] \\
A_{1\underline{4}} = A_{14} &= \frac{8}{15\pi^2 \kappa^4} [8(2 - \kappa^2)(1 - \kappa^2) K_2 \\
&\qquad - (16 - 16\kappa^2 + \kappa^4) E_2] \\
A_{3\underline{2}} = A_{32} &= \frac{8}{15\pi^2 \kappa^2} [2(1 + 2\kappa^2)(1 - \kappa^2) K_2 \\
&\qquad - (2 + 3\kappa^2 - 8\kappa^4) E_2] \\
A_{2\underline{3}} = -A_{23} &= \frac{8}{15\pi^2 \kappa^3} [(8 + \kappa^2)(1 - \kappa^2) K_2 - (8 - 3\kappa^2 - 2\kappa^4) E_2] \\
A_{50} &= \frac{8}{75\pi^2} [(43 - 168\kappa^2 + 128\kappa^4) E_2 \\
&\qquad - 4(7 - 16\kappa^2)(1 - \kappa^2) K_2] \\
A_{05} &= \frac{8}{75\pi^2 \kappa^5} [(128 - 168\kappa^2 + 43\kappa^4) E_2 \\
&\qquad - (128 - 104\kappa^2 + 15\kappa^4)(1 - \kappa^2) K_2]
\end{aligned}
\tag{21}
$$

Negative digits of the subscripts are underscored in the coefficients for lower side frequencies.

Upon putting the various quantities into equation (8) from equations (2), (13), (14), and (18) it thus becomes

$$
\begin{aligned}
B = {} & \mu_0 P \cos pt + \mu_0 Q \cos qt \\
& + \nu P(P+Q) \sum_{m=0}^{\infty} (A_{2m} + kA_{2m+2}) \cos [(m+1)p - mq]t \\
& + \nu P(P+Q) \sum_{m=0}^{\infty} (kA_{2m} + A_{2m+2}) \cos [mp - (m+1)q]t \\
& - \tfrac{1}{4}\nu P^2 \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} (A_{m+2,\,n} - 2A_{mn} + A_{m-2,\,n}) \\
& \qquad \times \sin [mp + nq]t \\
& - \tfrac{1}{4}\nu Q^2 \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} (A_{m,\,n-2} - 2A_{mn} + A_{m,\,n+2}) \\
& \qquad \times \sin [mp + nq]t \\
& + \tfrac{1}{2}\nu PQ \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} (A_{m+1,\,n-1} - A_{m+1,\,n+1} \\
& \qquad + A_{m-1,\,n+1} - A_{m-1,\,n-1}) \sin [mp + nq]t
\end{aligned}
\tag{22}
$$

with the understanding that $A_{rs} \equiv 0$ for $r < 0$ and for $r = 0$, $s < 0$. The first line is the linear portion of the induction, given by a permeability constant at its initial value; the first two summations arise from the variation of the permeability with the maximum magnetizing force; the remaining terms comprise the results of distortion attributable to hysteresis *per se*. The coefficients of the induction, c_{mn} and d_{mn} of equation (4), may now be evaluated by selecting the necessary quantities from equation (22).

General expressions for the coefficients can be evolved in series known as hypergeometric functions. These are all of the type

$$F(\alpha, \beta; \gamma; z) = 1 + \frac{\alpha\beta}{\gamma}\frac{z}{1!} + \frac{\alpha(\alpha+1)\beta(\beta+1)}{\gamma(\gamma+1)}\frac{z^2}{2!} + \cdots; \tag{23}$$

a particular one is chosen by specifying the parameters. The coefficients needed here are

$$A_{mn} = \frac{2}{\pi} \frac{\Gamma\left(\frac{m+n}{2}\right) \kappa^n}{\Gamma(n+1)\,\Gamma\left(\frac{m-n}{2}+1\right)} F\left(\frac{m+n}{2}, \frac{n-m}{2}; n+1; \kappa^2\right) \tag{24}$$

for m and n both positive and $m + n$ odd, and

$$A_\eta = \frac{\Gamma\left(\frac{\eta+1}{2}\right) k_1{}^\eta}{\Gamma(\eta+1)\,\Gamma\left(\frac{1-\eta}{2}+1\right)} F\left(\frac{\eta+1}{2}, \frac{\eta-1}{2}; \eta+1; k_1{}^2\right) \tag{25}$$

for η even. When n is negative, the coefficient may be found by using the relation

$$A_{m\underline{n}} = (-1)^n A_{mn}; \tag{26}$$

for all other values of the subscripts excluded the coefficients are zero. A recurrence formula for computing products of higher order is

$$A_{mn} = \frac{1}{m+n}\left[2\left(\frac{n-1}{\kappa} + (m-1)\kappa\right)A_{m-1,\ n-1} - (m+n-4)A_{m-2,\ n-2}\right] \tag{27}$$

with $m - 2$ positive. Comparison of equation (24) with equation (25) reveals that if in the former κ is replaced by k_1,

$$A_\eta = \frac{\pi}{2}A_{1,\eta}, \tag{28}$$

so the equations (21) and the recurrence formula (27) suffice for computing all the coefficients in the series for the induction.

Calculation of the Induction—Case 2

For case 2 the two branches of a minor loop and an adjoined portion of the major loop are combined into a Fourier's series whose coefficients are functions of position in the major loop. By developing these coefficients into Fourier's series, a double series in time is obtained. For this case, as for the other just considered, the induction thus is developed in the form of equation (4) and the coefficients are determined through the third order.

When minor loops are formed throughout the lower frequency cycle, an expression for each minor loop and the portion of the major loop joining it to the next one is found relative to an origin at the junction of the major and minor loops. A succession of such loops is then referred to the origin of the major loop by transforming the coordinates of a typical minor loop, the transformation being a function of the position of the minor loop.

Attention first will be devoted to a single high-frequency cycle occurring while the lower-frequency component of the magnetizing force is decreasing. Let the time of occurrence of the maximum in the higher frequency component of the magnetizing force during this cycle be designated by τ. By restricting application to characteristics with sizeable minor loops, i.e. $\kappa \ll 1$ when $p > q$, consistent with the stipulation that the minor loops do not vanish anywhere, this maximum

can be made practically coincident with the corresponding maximum in the magnetizing force.

Writing

$$t = \tau + \lambda, \tag{29}$$

the lower-frequency component in the vicinity is expressible by the Taylor's series

$$Q \cos qt = Q\left[1 - \frac{q^2\lambda^2}{2!} + \frac{q^4\lambda^4}{4!} - \cdots\right]\cos q\tau - Q\left[q\lambda - \frac{q^3\lambda^3}{3!} + \frac{q^5\lambda^5}{5!} - \cdots\right]\sin q\tau. \tag{30}$$

Over one cycle of the higher frequency this component is very nearly linear, so its variation in this range is

$$-4\Delta = -2\pi\kappa P \sin q\tau. \tag{31}$$

Since τ has been so chosen as to be an integral multiple of $2\pi/p$, when equations (29) and (30) are substituted in the magnetizing force given by equation (2) it becomes

$$h = P \cos p\lambda + Q \cos q\tau - \lambda Qq \sin q\tau. \tag{32}$$

Its value referred to a new set of coordinates, B', h', with their origin at the junction of the major and minor loops is

$$h' = P(1 + \cos p\lambda) - 2\Delta\left(1 + \frac{p\lambda}{\pi}\right). \tag{33}$$

According to Madelung's findings previous minor loops will not influence the one under consideration, so its lower branch will proceed toward the upper tip of the major loop as indicated in Fig. 2. A transformation of equation (1), simplified by the use of Rayleigh's relation, then gives for this branch

$$B_2' = \mu_0 h' + \nu h'^2. \tag{34}$$

The upper branch is

$$B_1' = [\mu_0 + 4\nu(P - \Delta)]h' - \nu h'^2. \tag{35}$$

The small portion of the major loop traversed during the last of the cycle may be expanded in a Taylor's series

$$B' = B'(0) + h'\left.\frac{\partial B_1}{\partial h}\right|_{h=h_r} + \frac{h'^2}{2!}\left.\frac{\partial^2 B_1}{\partial h^2}\right|_{h=h_r} + \cdots, \tag{36}$$

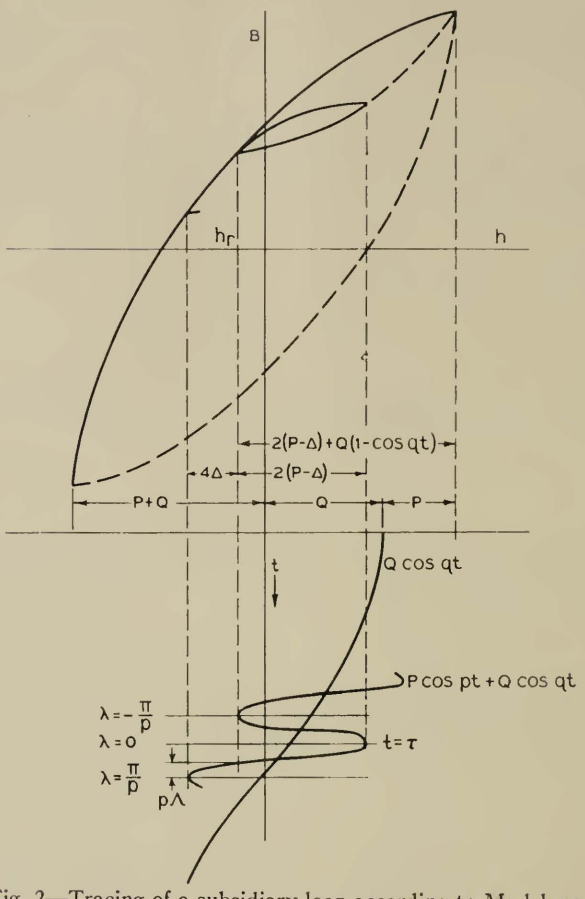


Fig. 2—Tracing of a subsidiary loop according to Madelung.

denoting by h_r the value of h at $h' = 0$, which is

$$h_r = 2\Delta - P + Q \cos q\tau. \tag{37}$$

The upper branch of the major loop as a function of h (not h') is

$$B_1 = (\mu_0 + 2\nu H)h + \nu(H^2 - h^2) \tag{38}$$

with

$$H = P + Q.$$

The expansion (36) is now found to be

$$B' = \{\mu_0 + 2\nu[2(P - \Delta) + Q(1 - \cos q\tau)]\}h' - \nu h'^2. \tag{39}$$

Equations (34), (35), and (39) define the induction in terms of h' over an entire high-frequency cycle; the first is valid for $-\pi < p\lambda < 0$, the second for $0 < p\lambda < \pi - p\Lambda$, and the last for $\pi - p\Lambda < p\lambda < \pi$.

For combining these three expressions a Fourier's series may be developed applicable over the entire interval $-\pi < p\lambda < \pi$, and the induction expressed by this series can be referred to the central axes, B, h, of the major loop by including in its constant term the value of the induction at the junction of the major and minor loops. The series is

$$B_1 = \tfrac{1}{2}b_0' + b_1' \cos p\lambda + b_2' \cos 2p\lambda + b_3' \cos 3p\lambda + a_1' \sin p\lambda + a_2' \sin 2p\lambda + a_3' \sin 3p\lambda. \quad (40)$$

The coefficients are determined by the integrals

$$a_n' = \frac{1}{\pi}\int_{-\pi}^{\pi} B \sin np\lambda \, d(p\lambda), \qquad b_n' = \frac{1}{\pi}\int_{-\pi}^{\pi} B \cos np\lambda \, d(p\lambda), \quad (41)$$

where

$$B = B' + B_r.$$

B_r is the induction at the junction of the major and minor loops, found by inserting equation (37) in equation (38). The resulting quantity together with expressions previously found for B' over various parts of the cycle are substituted in the integrands of equation (41), and the integrations are performed, using h' given by equation (33). All terms of order higher than the third and those containing the square of the frequency ratio as a factor are rejected as they occur. The resulting coefficients are functions of $q\tau$ both explicitly and also through Δ and $p\Lambda$.

To determine $p\Lambda$ as a function of $q\tau$, the vanishing of h' at the tip of the minor loop on the major loop gives an equation for use along the descending branch of the major loop. By equation (33)

$$P(1 - \cos p\Lambda) = 4\Delta\left(1 - \frac{p\Lambda}{2\pi}\right),$$

$$(1 - \cos p\Lambda) = (2\pi - p\Lambda)\kappa \sin q\tau. \quad (42)$$

An approximation to the general solution can be got by transforming equation (42) into a quadratic algebraic equation and solving. This is done by means of the first two terms of the power series expansion for $\cos p\Lambda$. The approximation will be best for small values of the angle, but very good over all its admissible range. This reduction gives

$$(p\Lambda)^2 + 2\kappa \sin q\tau \, (p\Lambda) - 4\pi\kappa \sin q\tau = 0, \quad (43)$$

the roots of which are

$$p\Lambda = -\kappa \sin q\tau \pm \sqrt{\kappa^2 \sin^2 q\tau + 4\pi\kappa \sin q\tau},$$

the positive one being that sought. By expanding the quantity under the radical according to the binomial theorem this root reduces to

$$p\Lambda = 2\sqrt{\pi}\ \sqrt{\kappa \sin q\tau} - \kappa \sin q\tau \tag{44}$$

when higher powers of κ are dropped.

The coefficients of the series then become, using the value of Δ from equation (31) and the value of $p\Lambda$ from equation (44),

$$\begin{aligned}
\frac{1}{2} b_0' &= \mu_0 Q \cos q\tau + 2\nu(2P + Q)Q \cos q\tau + \left(\pi - \frac{4}{\pi}\right) \nu PQ \frac{q}{p} \sin q\tau \\
&\quad + 2\pi\nu Q^2 \frac{q}{p} \sin q\tau\,(1 - \cos q\tau) + \nu Q^2 \sin^2 q\tau, \\
a_1' &= -2\mu_0 Q \frac{q}{p} \sin q\tau - 8\nu PQ \frac{q}{p} \sin q\tau + \frac{8}{3\pi} \nu P^2, \\
b_1' &= \mu_0 P + 2\nu P^2, \\
a_2' &= \mu_0 Q \frac{q}{p} \sin q\tau + \frac{14}{3} \nu PQ \frac{q}{p} \sin q\tau, \\
b_2' &= -\frac{40}{9\pi} \nu PQ \frac{q}{p} \sin q\tau, \\
a_3' &= -\frac{2}{3} \mu_0 Q \frac{q}{p} \sin q\tau - \frac{8}{15\pi} \nu P, \\
b_3' &= 0.
\end{aligned} \tag{45}$$

The relation (29) can be used to return equations (45) to the general time coordinate. Replacing τ by $t - \lambda$ gives

$$\begin{aligned}
\cos q\tau &= \cos q\lambda \cos qt + \sin q\lambda \sin qt, \\
\sin q\tau &= \cos q\lambda \sin qt - \sin q\lambda \cos qt.
\end{aligned}$$

As $|\lambda|$ never exceeds π/p these equations can be simplified to

$$\begin{aligned}
\cos q\tau &= \cos qt + \frac{q}{p}\,(2 \sin p\lambda - \sin 2p\lambda) \sin qt, \\
\sin q\tau &= \sin qt - \frac{q}{p}\,(2 \sin p\lambda - \sin 2p\lambda) \cos qt,
\end{aligned} \tag{46}$$

using the first terms of the Fourier's series for $\sin q\lambda$ in multiples of $p\lambda$. Upon substitution of equation (29) trigonometric functions of $p\lambda$ become the same functions of $(pt - 2j\pi)$, j an integer, since τ is defined as an even integral multiple of $2\pi/p$. The phase angle is

therefore immaterial, so $p\lambda$ can be replaced by pt in equations (40) and (46). Combination of equations (45) and (46) with equation (40) results in an expression of the induction on the upper half of the complex loop in terms of time.

For the lower half of the loop a mean position must be found. Having started with incommensurable frequencies at zero phase angles, the reversals at the lower tip of the major loop will occur at

$$pt = 2m\pi + R,$$

where $0 < R < 2\pi$ and all values of R within these limits are equally probable. The lower frequency component at the instant of reversal will have a time angle

$$qt = (2n + 1)\pi + S,$$

where $-(q/p)\pi < S < (q/p)\pi$ and all values of S within the limits are equally probable. The expected medians of the time angles are therefore $(2m + 1)\pi$ and $(2n + 1)\pi$ for the higher and lower frequency components respectively. These values and the point symmetry of the characteristic specify that the induction during the ascendency of the lower-frequency component of the magnetizing force will be equal in magnitude and opposite in sign to the induction for the descent with the phase of each component increased by its expected median. The induction for an increasing lower-frequency component is therefore given by

$$B_2[pt, qt] = -B_1[pt + \pi, qt + \pi], \tag{47}$$

where the right-hand member is evaluated for a decreasing lower-frequency component.

The coefficients of equation (40) may be altered accordingly to furnish a set for use when the lower-frequency component is increasing by replacing qt by $(qt + \pi)$ and pt by $(pt + \pi)$. In series form, then, the induction on the lower half of the loop is

$$B_2 = \tfrac{1}{2}b_0'' + b_1'' \cos pt + b_2'' \cos 2pt + b_3'' \cos 3pt + a_1'' \sin pt + a_2'' \sin 2pt + a_3'' \sin 3pt. \tag{48}$$

One pair of coefficients is necessary to specify completely the amplitude of each component of the induction when it is split into in-phase and quadrature terms harmonic in pt. Coefficients of corresponding terms in equations (40) and (48) are all functions of qt, each series applying over one half of a lower-frequency cycle. Each pair of coefficients can therefore be developed into a Fourier's series in qt. so that the single expression

$$B = \tfrac{1}{2}b_0 + b_1 \cos pt + b_2 \cos 2pt + b_3 \cos 3pt + a_1 \sin pt + a_2 \sin 2pt + a_3 \sin 3pt \quad (49)$$

defines the induction everywhere.

The coefficients of equation (49) are given by the expressions

$$\begin{aligned} b_n &= \frac{1}{2}[b_n' + b_n''] + \frac{2}{\pi}[b_n' - b_n''] \sum_{m=0}^{\infty} \frac{\sin (2m+1)q\tau}{(2m+1)}, \\ a_n &= \frac{1}{2}[a_n' + a_n''] + \frac{2}{\pi}[a_n' - a_n''] \sum_{m=0}^{\infty} \frac{\sin (2m+1)q\tau}{(2m+1)}. \end{aligned} \quad (50)$$

After putting the values of the primed coefficients into the respective terms, changing the arguments of functions of $q\tau$ to qt by using equations (46), and expanding the result into multiple angles, there remains when terms beyond the third order are dropped

$$\begin{aligned} B = {} & [\mu_0 + 2\nu P]P \cos pt + \frac{8}{3\pi}\nu[P + 3\kappa Q]P \sin pt \\ & + \left[\mu_0 + 2\nu\left(2 + k - \frac{4}{3}\kappa\right)P\right] Q \cos qt \\ & + \frac{8}{3\pi}\nu\left[k + \frac{3\pi^2}{4}\kappa + \frac{3}{2}\left(\frac{\pi^2}{4} - 1\right)\frac{\kappa}{k}\right] PQ \sin qt \\ & - \frac{1}{3}\nu[1 - 6k]\kappa P^2 \cos (2p + q)t \\ & - \frac{40}{9\pi}\nu\left[1 + \frac{3}{10}k\right]\kappa P^2 \sin (2p + q)t \\ & + \frac{1}{3}\nu[1 - 6k]\kappa P^2 \cos (2p - q)t \\ & + \frac{40}{9\pi}\nu\left[1 - \frac{3}{10}k\right]\kappa P^2 \sin (2p - q)t \\ & + \frac{32}{5\pi^2}\nu\kappa PQ \cos (p + 2q)t \\ & - \frac{2}{3\pi}[\mu_0 + 2\nu(2 + 3k)P]\kappa P \sin (p + 2q)t \\ & - \frac{32}{5\pi^2}\nu\kappa PQ \cos (p - 2q)t \\ & - \frac{2}{3\pi}[\mu_0 + 2\nu(2 + 3k)P]\kappa P \sin (p - 2q)t \\ & - \frac{4}{3\pi}\left[\mu_0\kappa - \frac{2}{5}\nu P\right] P \sin 3pt \\ & + \frac{8}{5}\nu\kappa PQ \cos 3qt - \frac{8}{15\pi}\nu Q^2 \sin 3qt. \end{aligned} \quad (51)$$

Recapitulation of Principal Results

General formulæ have now been made available for calculating the flux density over a wide range of conditions of two-frequency magnetization. For many ordinary purposes a table or graph of some of the results is convenient; useful ones are therefore included.

The hypergeometric expansions in the coefficients of case 1 can be put to further use to examine the behavior of the induction for special ratios of fundamental amplitudes and frequencies. When $k \ll 1$ and $\kappa \ll 1$ the coefficients of the several frequencies in the induction reduce to simple, rational, algebraic expressions of the amplitudes. These coefficients likewise reduce when $k \doteqdot 1$ and $\kappa \doteqdot 1$, since

$$F(\alpha, \beta; \gamma; 1) = \frac{\Gamma(\gamma)\Gamma(\gamma - \beta - \alpha)}{\Gamma(\gamma - \alpha)\Gamma(\gamma - \beta)}.$$

Third order coefficients for case 1 with restricted parameters, and also those for case 2 are tabulated in the accompanying table through terms in the lowest power of the smaller amplitude. Underlined subscripts distinguish lower side frequencies; a bar under a digit indicates it is to be taken with a negative sign.

When the ratio of amplitudes is unrestricted, graphs of the coefficients which specify the induction enable their magnitudes to be determined most readily. The strongest products are found to be the third order lower side frequencies; Fig. 3, calculated by A. G. Tynan, may be used to get both components of either of these. The corresponding upper side frequencies are almost as strong; their amplitudes can be found from the figure by virtue of the relations $c_{12} = c_{\underline{12}}$ and $c_{21} = c_{\underline{21}}$, since their other components are zero.

By interchanging P and Q and likewise p and q in either the table or the graphs, the subscripts are also reversed. In the table the inequalities restricting the columns are reversed thereby, since the interchanged quantities are arbitrarily assignable. This procedure does not extend the applicability of either case, but does permit the application of case 1 directly to both extremes of the amplitude ratio and the use of the curves to evaluate the amplitudes of both lower side frequencies. The scope of the table and curves given has been extended in this manner. A field of usefulness sufficiently extensive for most purposes of present-day communication is thereby achieved.

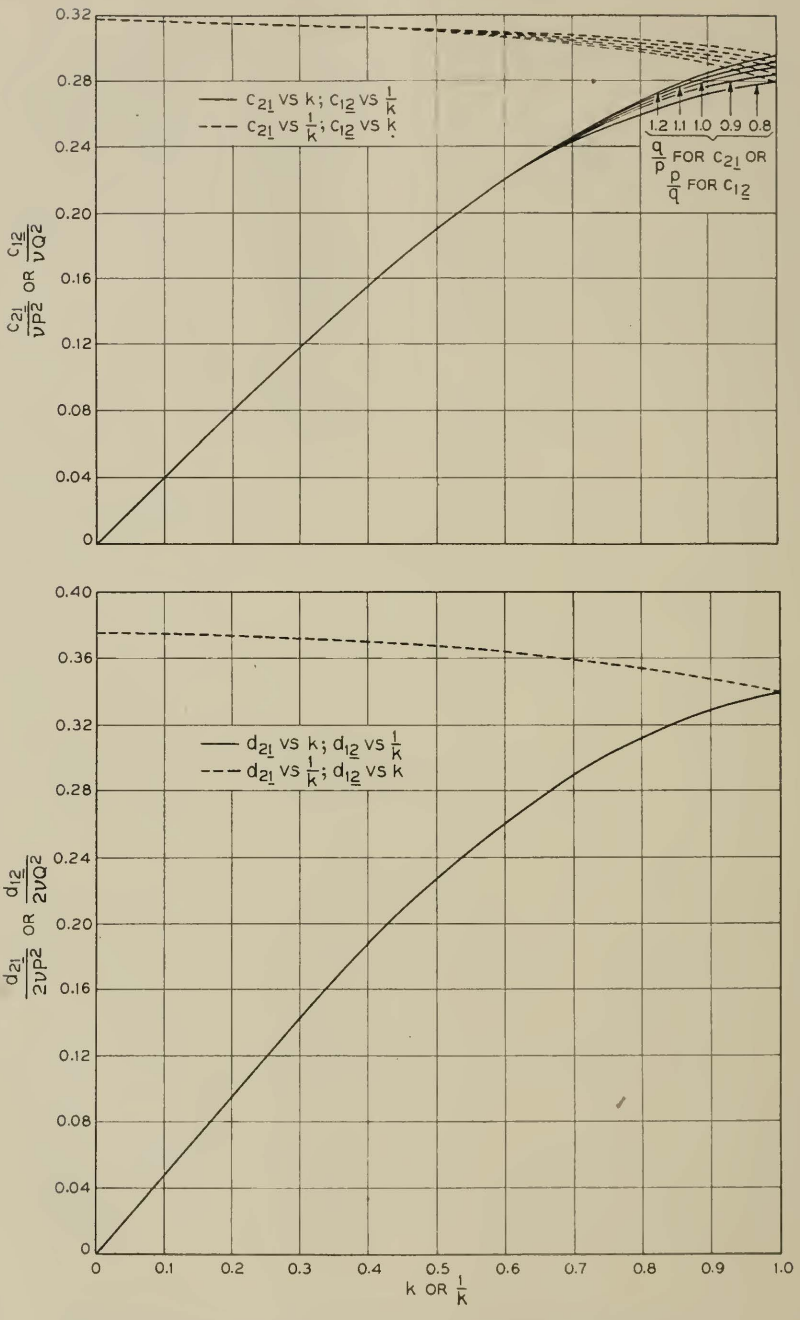


Fig. 3—Components of third order lower side frequencies—case 1.

TABLE I

	Case 1a $\kappa \ll 1$, $k \ll 1$	Case 1b $\kappa \doteq 1$, $k \doteq 1$	Case 1c $\kappa \gg 1$, $k \gg 1$	Case 2 $\kappa \ll 1$, $k \gg \kappa$
c_{10}	$\frac{8}{3\pi}\nu P^2$	$\frac{128}{9\pi^2}\nu PQ$	$\frac{4}{\pi}\nu PQ$	$\frac{8}{3\pi}\nu P^2$
d_{10}	$[\mu_0 + 2\nu P]P$	$\left[\mu_0 + \frac{32}{3\pi}\nu P\right]P$	$[\mu_0 + 3\nu Q]P$	$[\mu_0 + 2\nu P]P$
c_{01}	$\frac{4}{\pi}\nu PQ$	$\frac{128}{9\pi^2}\nu PQ$	$\frac{8}{3\pi}\nu Q^2$	$\frac{8}{3\pi}\nu Q^2$
d_{01}	$[\mu_0 + 3\nu P]Q$	$\left[\mu_0 + \frac{32}{3\pi}\nu Q\right]Q$	$[\mu_0 + 2\nu Q]Q$	$[\mu_0 + 2\nu(2P + Q)]Q$
c_{12}	$-\frac{1}{\pi}\nu Q^2$	$-\frac{128}{45\pi^2}\nu Q^2$	$-\frac{4}{3\pi}\nu PQ$	$\frac{2}{3\pi}[\mu_0 + 2\nu(2P + 3Q)]\kappa P$
d_{12}	$\frac{3}{4}\nu(1 + k)Q^2$	$\frac{32}{15\pi}\nu Q^2$	νPQ	$-\frac{32}{5\pi^2}\nu\kappa PQ$
c_{12}	$-\frac{1}{\pi}\nu Q^2$	$-\frac{128}{45\pi^2}\nu Q^2$	$-\frac{4}{3\pi}\nu PQ$	$\frac{2}{3\pi}[\mu_0 + 2\nu(2P + 3Q)]\kappa P$
d_{12}	0	0	0	$\frac{32}{5\pi^2}\nu\kappa PQ$
c_{21}	$\frac{4}{3\pi}\nu PQ$	$\frac{128}{45\pi^2}\nu P^2$	$\frac{1}{\pi}\nu P^2$	$-\frac{4}{9\pi}\nu[5P - 6Q]\kappa P$
d_{21}	νPQ	$\frac{32}{15\pi}\nu P^2$	$\frac{3}{4}\nu(1 + 1/k)P^2$	$\frac{1}{3}\nu[P - 6Q]\kappa P$
c_{21}	$-\frac{4}{3\pi}\nu PQ$	$-\frac{128}{45\pi^2}\nu P^2$	$-\frac{1}{\pi}\nu P^2$	$-\frac{4}{9\pi}\nu[5P + 6Q]\kappa P$
d_{21}	0	0	0	$-\frac{1}{3}\nu[P - 6Q]\kappa P$
c_{30}	$-\frac{8}{15\pi}\nu P^2$	$-\frac{128}{225\pi^2}\nu P^2$	0	$-\frac{4}{3\pi}\left[\mu_0\kappa - \frac{2}{5}\nu P\right]P$
d_{30}	0	0	0	0
c_{03}	0	$-\frac{128}{225\pi^2}\nu Q^2$	$-\frac{8}{15\pi}\nu Q^2$	$-\frac{8}{15\pi}\nu Q^2$
d_{03}	0	0	0	$\frac{8}{5}\nu\kappa PQ$

Intermodulation Products

Generated Modulation Voltages

From the foregoing results the voltage generated in a coil of N turns on a closed ferromagnetic core of cross-sectional area A can be found by the use of

$$e(t) = NA\ 10^{-8}\frac{dB}{dt}. \tag{52}$$

Components of this voltage segregated according to frequency are of the type

$$e_{mn}(t) = (mp + nq)NA\ 10^{-8}\left[c_{mn}\cos\ (mp + nq)t - d_{mn}\sin\ (mp + nq)t\right], \tag{53}$$

each proportional to its frequency and in general having two components in quadrature. The amplitudes of these will be designated by

$$E_{mn}' = (mp + nq)NA\ 10^{-8}\ c_{mn}, \tag{54}$$

$$E_{mn}'' = (mp + nq)NA\ 10^{-8}d_{mn}. \tag{55}$$

The amplitude of their resultant is

$$E_{mn} = (mp + nq)NA\ 10^{-8}\sqrt{c_{mn}^2 + d_{mn}^2}. \tag{56}$$

One component, if it greatly exceeds the other, may be taken as the generated distortion voltage. The various coefficients to which the components of the voltage are proportional have already been calculated, and also given in tabular or graphical form for specific instances. The relations

$$P = \frac{0.4\pi NI}{l}, \quad Q = \frac{0.4\pi NJ}{l},$$

where l is the mean length of magnetic path, may be used to convert these results into terms of the current amplitudes I and J. Where r-m-s quantities are used, they will be distinguished by bars over them.

Several features of the distortion are particularly outstanding. Perhaps chief among these is the dissimilarity of corresponding upper and lower side-frequency voltages. Inasmuch as these are products of a reactance modulator, they might be expected to be in the ratio of their frequencies, as they are found to be for one component. Often, however, a predominating component appears at each lower side frequency with no counterpart at the upper side frequency. This

component can be traced to the different axial slopes of the several branches of the loop, caused by their different points of origination. The slopes are fixed by the envelope of the magnetizing force; since this envelope is periodic in the difference of the fundamental frequencies, difference products will appear without corresponding sum products in the induction. Such a phenomenon is a fundamental property of the multivalued characteristic, and will occur wherever the envelope of a complex wave is instrumental in selecting the branch to be traversed.

No simple yet general rule seems to embrace the behavior of the various products in an iron core coil as governed by the amplitudes of the fundamental currents. Each voltage component is proportional to its frequency and to the product of two amplitudes, but these often enter in a complicated way. For fundamental frequencies close together all the higher order voltages vary directly with the hysteretic coefficient ν; for widely separated frequencies the distortion may depend also on the permeability through its effect on the axial slope of the minor loops. At the extremes of the amplitude ratio certain products or their components are found to be independent of one of the fundamental currents, the stronger one in some instances. When case 1 is applicable the third harmonic of the weaker fundamental current is suppressed below the value it would have without the stronger current superposed, while the third harmonic of the stronger fundamental is affected only slightly by the presence of a second frequency. Perusal of the table will disclose more detailed relations.

Distortion Measured in Coils

Voltages calculated by the theory have been compared with measured values for several coils using two common core materials. The agreement found provides a check of the theoretical predictions.

The two third order lower side frequencies of fundamental frequencies $p/2\pi = 760$ cycles per second and $q/2\pi = 600$ cycles per second are plotted in Fig. 4 for a higher frequency current of ten milliamperes in an iron dust coil of special design. The frequencies of the products are 920 and 440 cycles per second. These data were taken by I. E. Wood and the calculations were made by A. G. Tynan. These curves show the product as a function of the amplitude ratio more directly than the curves of Fig. 3.

Both upper and lower third order side frequencies have been measured by A. G. Landeen. The results are given in Figs. 5 and 6 for an annular core of iron dust. It is so wound that the magnetizing force is 0.04 times the current in milliamperes. The figures show two third

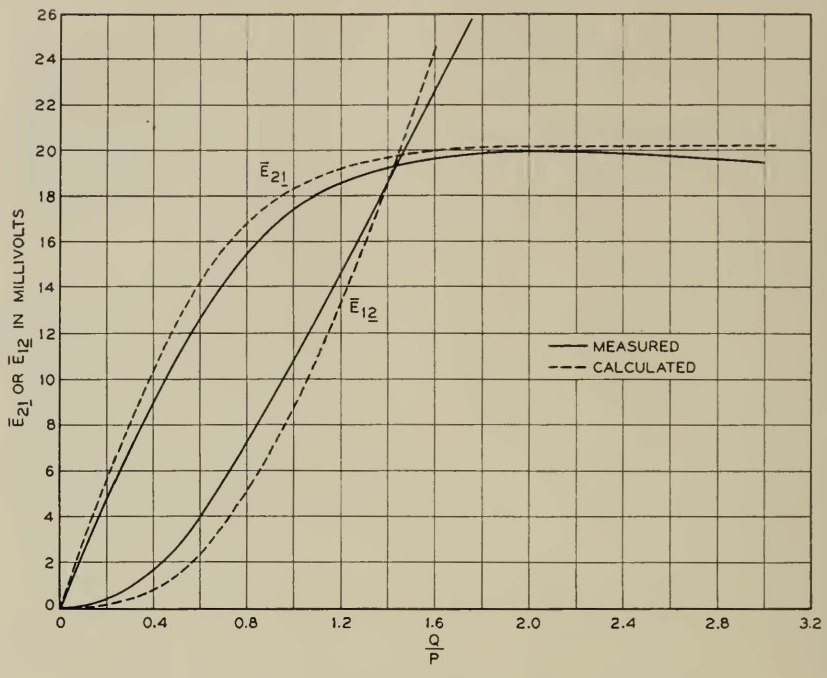


Fig. 4—Lower side-frequency voltages in an iron dust coil.

order products plotted against the fundamental current of higher frequency in each instance. For these measurements the current of one frequency was maintained at some fixed value and the amplitude of the other one varied. Both sum and difference products, but with different fundamental frequencies, are exhibited. The upper side

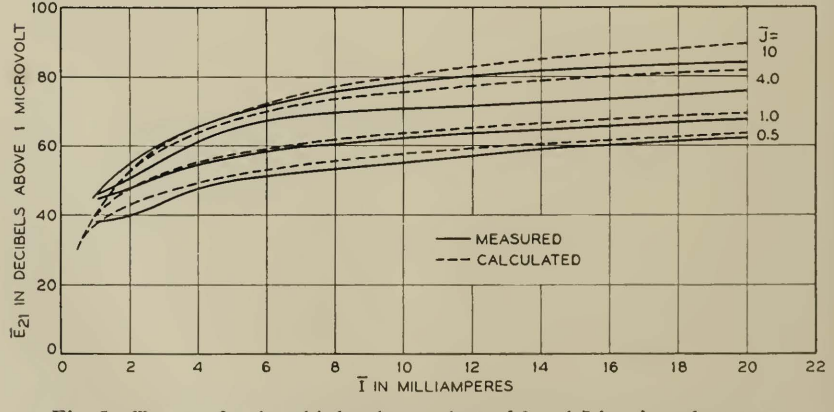


Fig. 5—Twenty-five kc. third order product of 9 and 7 kc., iron dust core. $\mu_0 = 24.5$, $\nu = 0.18$, $L_0 = 4.54$ mh., $II = 38.9\ I$.

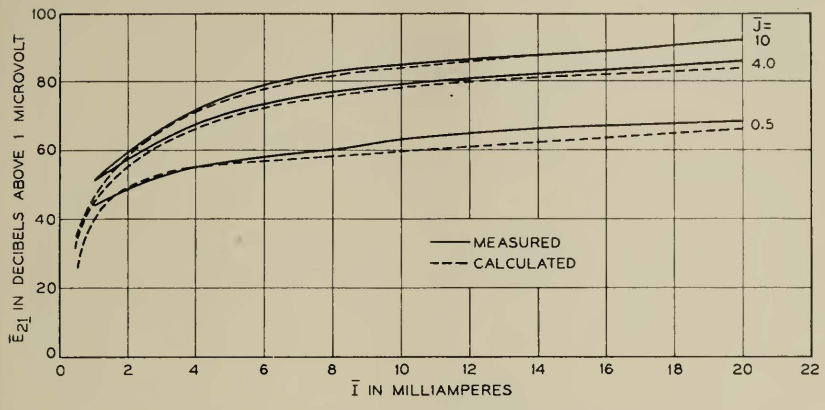


Fig. 6—Seventeen-kc. third order product of 13 and 9 kc., iron dust core. $\mu_0 = 24.5$, $\nu = 0.18$, $L_0 = 4.54$ mh., $H = 38.9\ I$.

frequency is the 25-kilocycle product of 9 and 7 kilocycles; the lower side frequency is the 17-kilocycle product of 13 and 9 kilocycles.

Some measurements to which subcase 1*b* is applicable are given in Figs. 7 and 8. Each curve gives a third order product for a permalloy dust core. The upper side frequency is 25 kilocycles generated by 9 and 7 kilocycle fundamental frequencies. The lower side frequency

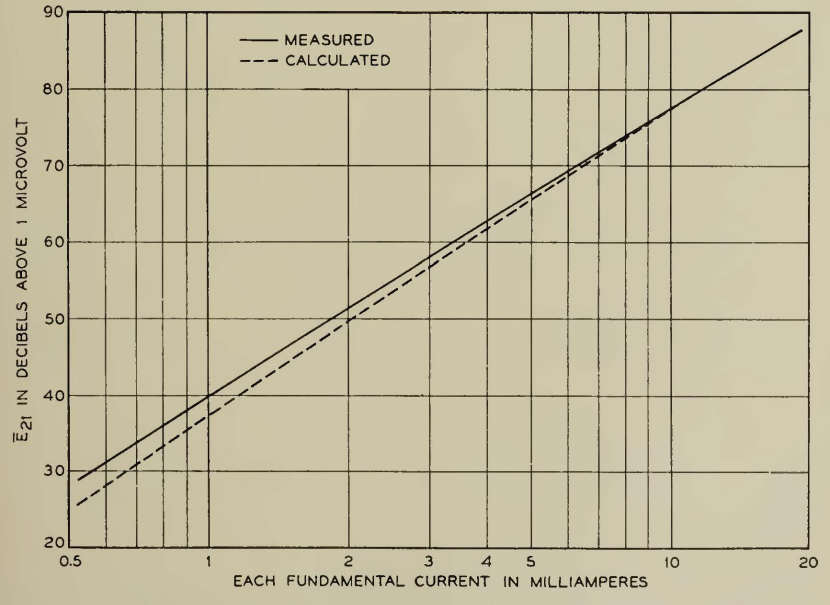


Fig. 7—Twenty-five kc. third order product of equal currents at 9 and 7 kc., permalloy dust core. $\mu_0 = 75$, $\nu = 0.41$, $L_0 = 5.45$ mh., $H = 36.0\ I$.

is 13 kilocycles from fundamental frequencies at 21 and 17 kilocycles. In these measurements both fundamental currents were changed simultaneously so as to be kept equal throughout. The approach to saturation at currents above ten milliamperes is apparent on these curves; below, the distortion voltage is proportional to the square of the current.

The measured curves are seen to agree well with the calculated ones in every instance, confirming theoretical values of the products within close limits. Eddy currents were negligible in all these coils because of the dust cores. The use of the formulae to determine important

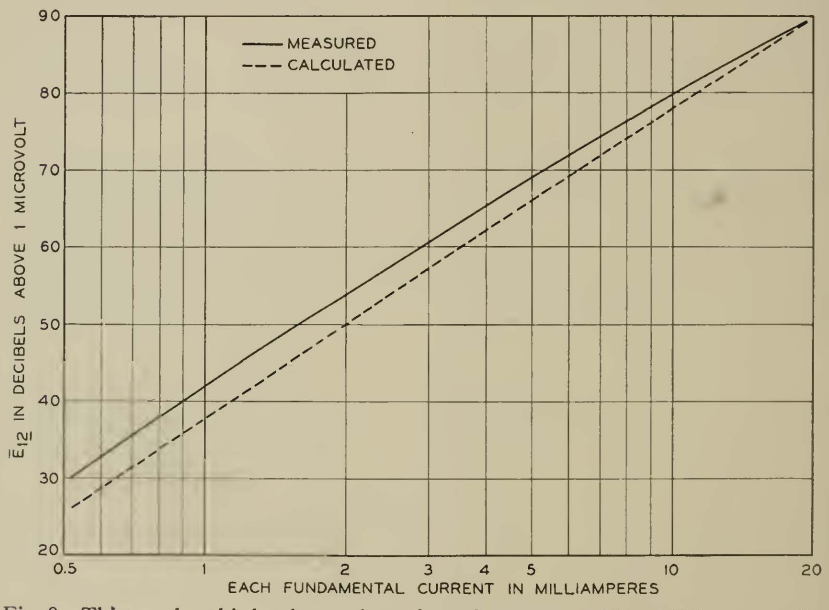


Fig. 8—Thirteen-kc. third order product of equal currents at 21 and 17 kc., permalloy dust core. $\mu_0 = 75$, $\nu = 0.41$, $L_0 = 5.45$ mh., $H = 36.0\ I$.

intermodulation products from the constants of the coils therefore seems to be justified.

Correlation with Single-Frequency Results

In the single-frequency case Peterson found the resistance and reactance at a fundamental frequency $\omega/2\pi$ to be increased by

$$\Delta R(\omega, II) = \frac{8}{3\pi}\omega L_0 \frac{\nu}{\mu_0} II,$$

$$\Delta X(\omega, II) = 2\omega L_0 \frac{\nu}{\mu_0} II,$$

respectively, on account of hysteresis and variable permeability, II being the maximum of the applied magnetizing force. The total non-linear reactance is then

$$X(\omega, H) = X_0 + \Delta X(\omega, II)$$

with

$$X_0 = \omega L_0,$$

the constant part, representing the reactance the coil would have if the permeability remained constant at its initial value.

The distortion voltages for a two-frequency input may be written in terms of these non-linear impedances. With some simplication this is done in Table II for special cases, using equation (56) and the relations

$$P = \frac{0.4\pi NI}{l}, \quad Q = \frac{0.4\pi NJ}{l}.$$

TABLE II

	Case 1a $\kappa \ll 1$ $k \ll 1$	Case 1b $\kappa \doteqdot 1$ $k \doteqdot 1$
E_{12}	$0.960\ \Delta R(p - 2q, Q)J$	$0.870\ \Delta R(p - 2q, Q)J$
E_{12}	$0.375\ \Delta R(p + 2q, Q)J$	$0.340\ \Delta R(p + 2q, Q)J$
E_{21}	$1.28\ \ \Delta R(2p - q, Q)I$	$0.870\ \Delta R(2p - q, P)I$
E_{21}	$0.500\ \Delta R(2p + q, Q)I$	$0.340\ \Delta R(2p + q, P)I$
E_{30}	$0.200\ \Delta R(3p, P)I$	$0.068\ \Delta R(3p, P)I$
E_{03}	—	$0.068\ \Delta R(3q, Q)J$
	Case 1c $k \gg 1$ $\kappa \gg 1$	Case 2 $\kappa \ll 1$ $k \gg \kappa$
E_{12}	$1.28\ \ \Delta R(p - 2q, P)J$	$0.212\ \kappa X(p - 2q, 2P + 3Q)I$
E_{12}	$0.500\ \Delta R(p + 2q, P)J$	$0.212\ \kappa X(p + 2q, 2P + 3Q)I$
E_{21}	$0.960\ \Delta R(2p - q, P)I$	$1.06\ \ \kappa\Delta R(2p - q, P - 0.944Q)I$
E_{21}	$0.375\ \Delta R(2p + q, P)I$	$1.23\ \ \kappa\Delta R(2p + q, P + 2.73Q)I$
E_{30}	—	$[0.425\ \kappa X_0(3p) - 0.200\ \Delta R(3p, P)]I$
E_{03}	$0.200\ \Delta R(3p, Q)J$	$0.200\ \Delta R(3p, Q)J$

The general formulæ of case 1 can be similarly represented, but not as concisely. Besides exhibiting the connection between intermodulation products and impedance changes, this table provides a convenient means for computing voltage components directly from data obtainable by single-frequency bridge measurements.

Hysteretic Impedances

The fundamental voltages are not included in the table. For them each component is separately significant. The singly primed E's are in phase with the corresponding magnetizing current and the doubly primed E's in quadrature. Hence, the former are resistance drops and the latter reactance drops, defining incremental components of impedance analogous to those mentioned for a single-frequency input.

For each of the fundamental frequencies the results developed previously may be used to determine these components. They represent the hysteretic resistance and the hysteretic reactance to the fundamental at hand, specified by a subscript p or q. They are tabulated in Table III for the special cases considered. The total resistance of the coil

TABLE III

	Case 1a $\kappa \ll 1$ $k \ll 1$	Case 1b $\kappa \doteq 1$ $k \doteq 1$	Case 1c $k \gg 1$ $\kappa \gg 1$	Case 2 $\kappa \ll 1$ $k \gg \kappa$
ΔR_p	$\frac{16}{15}\frac{N}{l}\frac{\nu}{\mu_0} pL_0I$	$\frac{256}{45\pi}\frac{N}{l}\frac{\nu}{\mu_0} pL_0J$	$\frac{8}{5}\frac{N}{l}\frac{\nu}{\mu_0} pL_0J$	$\frac{8}{5}\frac{N}{l}\frac{\nu}{\mu_0} pL_0I$
ΔX_p	$\frac{4\pi}{5}\frac{N}{l}\frac{\nu}{\mu_0} pL_0I$	$\frac{64}{15}\frac{N}{l}\frac{\nu}{\mu_0} pL_0I$	$\frac{6\pi}{5}\frac{N}{l}\frac{\nu}{\mu_0} pL_0J$	$\frac{4\pi}{5}\frac{N}{l}\frac{\nu}{\mu_0} pL_0I$
ΔR_q	$\frac{8}{5}\frac{N}{l}\frac{\nu}{\mu_0} qL_0I$	$\frac{256}{45\pi}\frac{N}{l}\frac{\nu}{\mu_0} qL_0I$	$\frac{16}{15}\frac{N}{l}\frac{\nu}{\mu_0} qL_0J$	$\frac{16}{15}\frac{N}{l}\frac{\nu}{\mu_0} qL_0J$
ΔX_q	$\frac{6\pi}{5}\frac{N}{l}\frac{\nu}{\mu_0} qL_0I$	$\frac{64}{15}\frac{N}{l}\frac{\nu}{\mu_0} qL_0J$	$\frac{4\pi}{5}\frac{N}{l}\frac{\nu}{\mu_0} qL_0J$	$\frac{4\pi}{5}\frac{N}{l}\frac{\nu}{\mu_0} qL_0[2I + J]$

to either fundamental current can be calculated by adding to the value of ΔR from the table the resistance of the windings, the eddy current resistance, and the initial (viscosity) resistance, all evaluated for the frequency of the fundamental. The eddy currents must be so small that the flux density is substantially uniform across the cross section of the core. The reactance X_0 of the coil can be diminished by the eddy current reduction factor for the fundamental frequency and added to the hysteretic reactance to give the net reactance of the coil under these conditions.

The table is helpful in evaluating the effect of one fundamental current upon the other. Within the limits of the analysis, which in substance limits the permeability to linear variation with the field intensity, the hysteresis loss at any frequency is either increased or unchanged by the superposition of a second frequency. Nearly equal input currents whose frequencies do not differ greatly share equally the hysteresis loss. This amounts to about twice what it would if either fundamental flowed alone. If the frequencies differ greatly, the

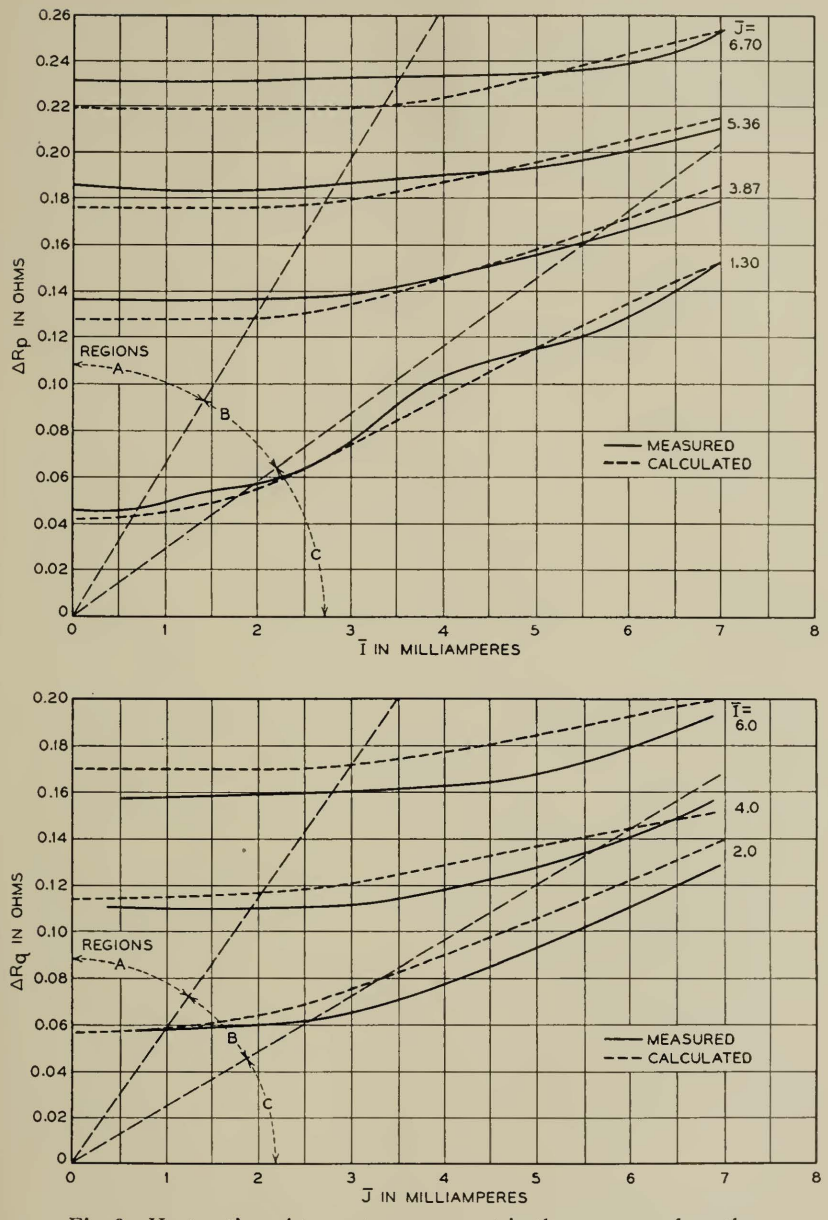


Fig. 9—Hysteretic resistance to one current in the presence of another.

loss to either is not affected by the other; if the amplitudes differ greatly, the loss to either is governed by the stronger current. In no case, at these weak fields, does the increased loss at one frequency

reduce the loss at another frequency, contrary to well established experiments at considerably higher fields, for which the hysteresis loss at one frequency may be reduced by superposing a magnetizing force at a different frequency. The inductance is the same to both funda-

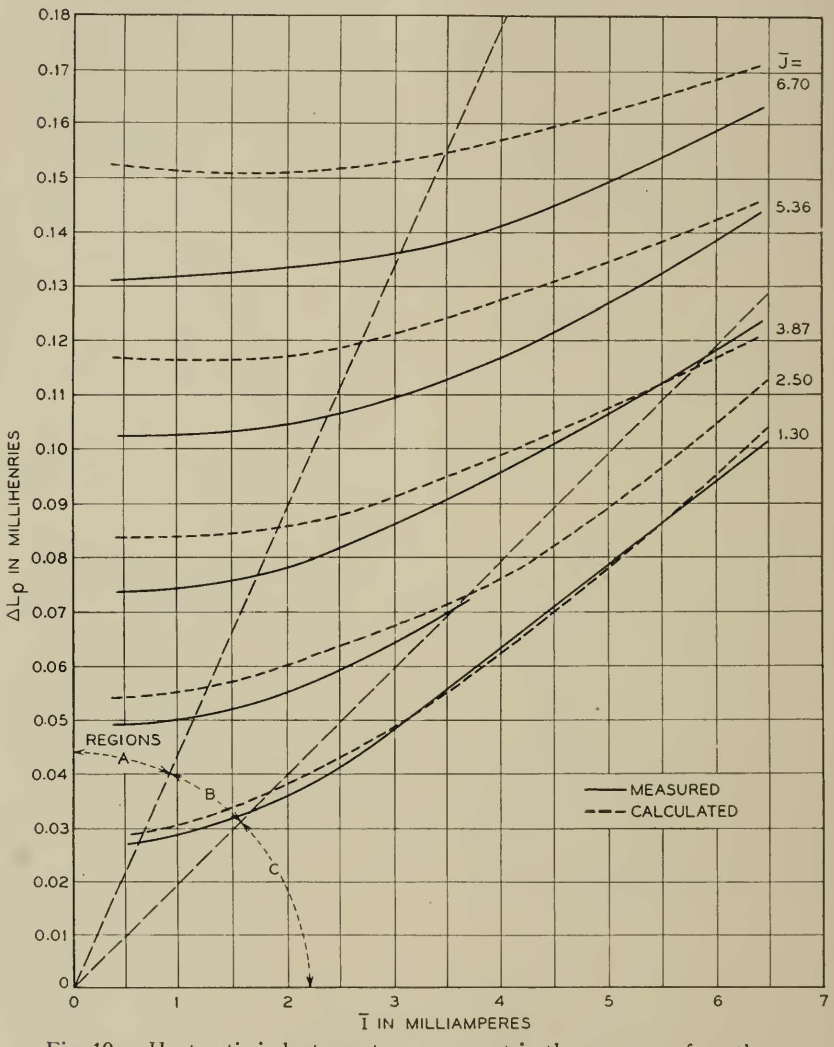


Fig. 10a—Hysteretic inductance to one current in the presence of another.

mentals in all cases except the second, for which a small difference exists. The effect of a superposed alternating current is always apparent through an increased inductance, although sometimes the increase may be slight; it is determined by the larger current.

The influence of one fundamental current upon another has been termed mutual crowding. Because of the increased attenuation, and at times because of resulting unbalance or phase shift, crowding becomes important when different frequencies or bands of frequencies

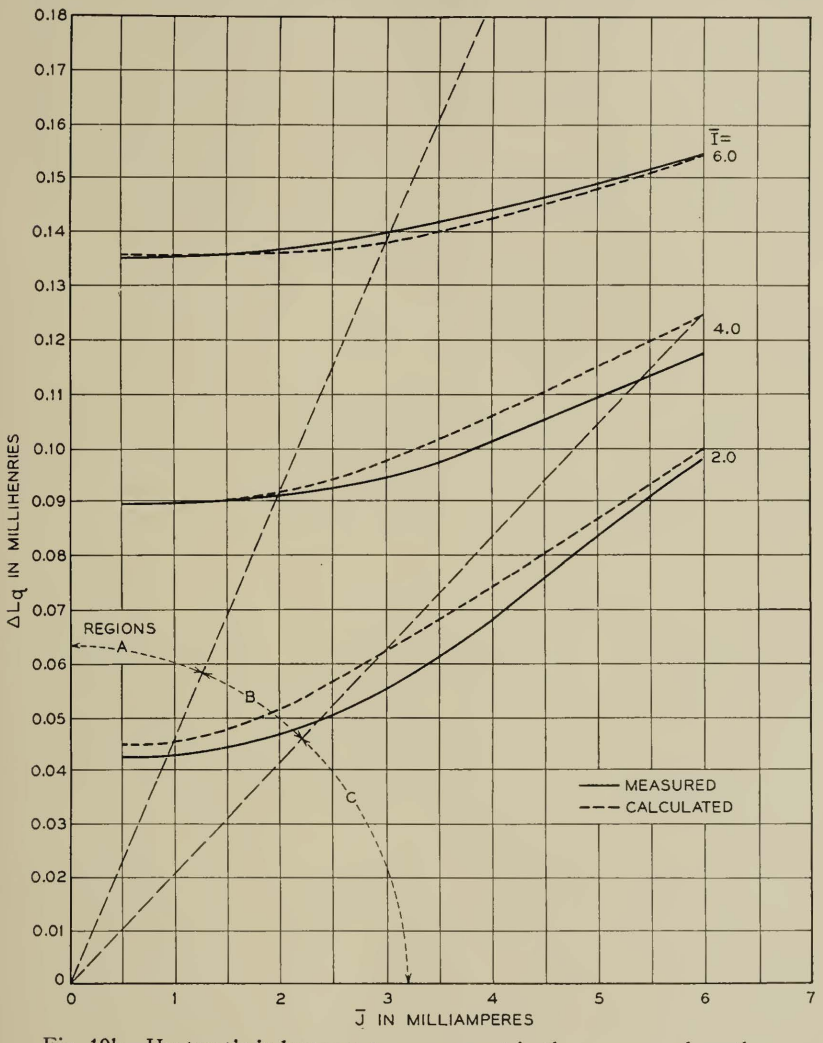


Fig. 10b—Hysteretic inductance to one current in the presence of another.

are transmitted simultaneously through a circuit including ferromagnetic material.

Incremental impedances for a twenty-two millihenry permalloy dust loading coil are given in Figs. 9 and 10. The current *I* had a frequency

of 550 cycles per second, and J 475 cycles per second. Measured and calculated values are plotted for comparison, with regions of applicability of the special subcases indicated. Portions of calculated curves falling in regions A or C were computed by the two sets of formulæ for subcases 1a and 1c, portions in the middle of region B by the formulæ for subcase 1b. The trend of the quantities measured is accurately portrayed by the calculations and agreement of the values is good.

All the curves commence at those values of resistance or inductance which would obtain in a single-frequency case with a current having the magnitude of the one here superposed in fixed amount. Upon increasing the variable current the measured quantities show an increase as it begins to preponderate, and eventually they approach asymptotically the values they would have if it flowed alone.

The measurements were made by L. R. Wrathall using a Maxwell inductance bridge with two imputs and a tuned detector. Eddy currents being of no consequence at the low frequencies employed, the chief sources of possible error are calibrations of the standards used and variation in the temperature of the coil during taking of the data. Changes in winding resistance caused by the latter are of the same order of magnitude as the changes in hysteretic resistance being observed. Precautions against both possibilities were taken.

Conclusion

The multiplicity of forms of complex hysteresis loops makes their analysis in general a complicated and difficult matter if indeed possible at all. Extensive experiments with two frequencies must be completed and the results classified according to the types of loops before an acceptable method of taking their form into account can be formulated. The parameters k and κ seem to be effective quantities for denoting concisely a particular form of loop in many instances.

A way of representing the behavior of complex loops more exactly than do Madelung's propositions is needed, and might be the fruit of precise experiments designed to clear up also the early closure and lack of closure which Lehde apparently found in minor loops. The tracing of complex loops is not simply cyclic, and only when a nearly complete magnetic cycle is executed between successive maxima in the magnetizing force can conditions approaching a cyclic state be expected to exist. Some experimental evidence of performance in other conditions is a present need which can perhaps be met by a thorough investigation of spiral characteristics. These seem to have been ignored entirely in the past, the literature dealing with sub-

sidiary loops only along a magnetization curve or a branch of a major loop.

Correlation of certain magnetic phenomena to a degree not heretofore attainable is made possible by the preceding development. Among at least some of these a qualitative connection has been well recognized. Flutter and allied effects are known to be aspects of modulation, and crowding is observed to accompany non-linear distortion quite generally. These features are related quantitatively by means of the theory, and are linked with their single-frequency counterparts. It thus becomes feasible to evaluate some of the more abstruse occurrences in terms of readily understandable effects simpler in nature; ultimate in this direction is the use of steady state results to forecast the behavior of transmitted speech or music.

Abstracts of Technical Articles from Bell System Sources

Reverberation Time and Absorption Measurements with the High Speed Level Recorder.[1] E. H. Bedell and K. D. Swartzel, Jr. It has been common in reverberation theories to neglect the effect of the stationary wave pattern in a room and to assume a logarithmic decay of the sound energy. In many cases this assumption of a constant decay rate is not fulfilled, and in particular it is well known that the decay curves as obtained with available instruments, which indicate either the pressure or the velocity at a point in the sound field, show marked fluctuations in the decay rate. The rate of decay, in general, varies during the decay period, from point to point in the room, and may depend upon the position of the sound source, and the location of absorbing materials. Very rapid fluctuations in the decay rate have commonly been averaged out by the measuring apparatus itself, either by making the indicating instrument sluggish in its action, or by measuring the average intensity over finite time intervals during the decay period. The slower, and perhaps more important, deviations from linearity in the decay curves have been either reduced, or averaged out, by a number of expedients. Among these are the use of rotating sound reflectors, or vanes, to break up the stationary wave pattern in the room; the use of frequency modulated, or warble, tones in place of a constant single frequency; moving the microphone to a number of positions in the room; and moving the sound source. Hunt has given some quantitative data on the effect of the warble tone, but similar data on other methods of smoothing out the decay curves are not available. This paper presents some data on the relative value of the above methods of improving the measured decay curves, and on the use of a motor driven rotating switch to connect in rapid sequence a number of microphones, placed in different parts of the room, into the measuring apparatus, for obtaining on a single curve a space average of the time decay pattern. Since many of the deviations from linearity in the decay curves have a "period" of 30 db or more, we should expect the accuracy of our values to be a function of the range through which the decay is measured, particularly when the range is not large compared to the period of the deviations. This effect is discussed for three values of the decay range, 30, 60 and 90 db.

[1] *Jour. Acous. Soc. Amer.*, January, 1935.

Standardization of Noise Meters.[2] R. G. McCURDY. A brief review of the present status of standardization of noise meters and measurements, and progress made to date by the technical committee on noise meters and noise levels of the American Standards Association.

A Rotating Mirror Oscilloscope.[3] R. F. MALLINA. When studying sound it is sometimes useful to project the wave-form of electrical or acoustical phenomena on a screen. A rotating mirror in combination with a vibrating mirror and a light source provide a convenient means of showing such waves. The problem of building an instrument for such a purpose is comparatively simple if a small screen is used in a dark chamber. However, when the screen is large enough to be viewed by a dozen or more persons, many difficulties arise.

The paper describes how the various parts of the apparatus may be coordinated in order to produce a comparatively bright, clearly defined wave with a small incandescent lamp in a room of average illumination. The vibrator used in the apparatus may be so constructed that its response is either inversely proportional to or independent of the frequency.

Shot Effect and Thermal Agitation in an Electron Current Limited by Space Charge.[4] G. L. PEARSON. The space current in a thermionic vacuum tube is not a steady flow of electricity but is subject to minute irregular fluctuations. The two most fundamental causes for these fluctuations are the random distribution of instants of emission of the individual electrons and the distribution of these electrons in velocity. The random emission produces shot noise which may be reduced by the space charge surrounding the cathode, while the velocity distribution produces thermal noise and is dependent upon the temperature of the cathode.

Although plausible theories of these effects have been given they have never been checked by accurate experiments because of the difficulties involved. By using two electrode tubes capable of producing a large space charge such measurements have now been made and are reported in this paper.

Simple Theory of the Three-Electrode Vacuum Tube.[5] H. A. PIDGEON. The physical principles upon which the operation of the three-element vacuum tube depends are presented in simple form and the terms

[2] *Elec. Engg.*, January, 1935; *Indus Standardization*, January, 1935.
[3] *Jour. S.M.P.E.*, December, 1934.
[4] *Physics*, January, 1935.
[5] *Jour. S.M.P.E.*, February, 1935.

usually applied to the tube, its operation as an amplifier, and a simple approximate method for computing the power output and percentage of distortion are explained.

No new material is presented in the paper although some of it is presented from a somewhat different point of view from that usually found in the literature. An effort has been made to present in reasonably compact form the essential features of the subject most useful to engineers interested in vacuum-tube applications.

The subjects discussed include: the portion of electron theory upon which the fundamental principles of vacuum-tube operation are based; space charge, the three-halves power law, temperature and voltage saturation; characteristics of the three-element tube; definition and physical significance of the terms plate resistance, transconductance, and amplification factor; dynamic characteristics, power output, and distortion; various means of coupling the vacuum tube to its associated circuits; and means for testing vacuum tubes for adequate thermionic emission.

Coaxial Communication Transmission Lines.[6] S. A. SCHELKUNOFF. A non-mathematical discussion of the mechanism whereby energy may be transmitted over long distances at high frequencies by the use of "coaxial conductors" is presented in this paper. A coaxial system consists of a cylindrical conducting tube within which a smaller conductor is coaxially placed. Such conductors, which reduce interference and crosstalk, are applicable for the transmission of telephone, telegraph, and television signals over a very wide range of frequencies.

Some Aspects of Quality Control.[7] W. A. SHEWHART. The object of this paper is to make clear what is meant by quality in a practical objective way that is subject to experimental verification and to consider some aspects of the problem of control. As a basis for judging the quality of current product it is necessary to obtain first of all adequate information, in the most efficient manner, on which to render a judgment. This can be accomplished by providing an inspection specification which is distinct from the design specification. One specifies the quantity and kind of evidence that is required as a basis for judging whether or not the quality of the product will attain its goal; the other specifies the goal. Certain elements of uncertainty must be allowed for in setting the goal. The discussion closes by pointing out the necessity of keeping a running report or record of the

[6] *Elec. Engg.*, December, 1934.
[7] *Mech. Engg.*, December, 1934.

evidence used in judging the quality of current product as a part of any scientific plan of making use of hindsight as well as foresight in controlling quality.

The Ionizing Effects of Meteors.[8] A. M. SKELLETT. It is shown that a meteor of average velocity has enough energy to cause ionization of atmospheric gases by impact. Recent experimental work by Frische and others on collisions of ions is interpreted as supporting the hypothesis that meteoric collisions do result in ionization. The afterglow of nitrogen is considered as a possible example of the process by which a meteor train remains glowing for a period of minutes and the coincidence of the region in which such trains are generally observed and of the E region of the upper atmosphere is pointed out. The spectra of bright meteors, while not showing atmospheric lines, are shown not to be inconsistent with the above hypothesis.

The behavior of the transatlantic short-wave radio telephone circuits of the American Telephone and Telegraph Company, during 1930, 1931, and 1932, is examined for possible meteoric effects. It is concluded that, in general, a rather large shower is necessary to affect them appreciably. This was to be expected since these circuits are normally under a continuous bombardment by random meteors. It seems possible that a certain degree of the variability (rapid fading, etc.) of received signals over such paths is due to this bombardment.

Results of radio pulse studies of the upper atmosphere, particularly by Schafer and Goodall, which are strongly suggestive of meteoric ionization, especially at times of special meteoric activity, are (1) sudden increases in ionization in the E region lasting for a period of minutes or less, and (2) increases of longer duration with maxima coincident in time with those of observed meteoric activity. Such tests made during the Leonid shower of November, 1932, were successful in correlating sudden increases in ionization in the E region with the visual observations of a number of bright meteors passing overhead. For the brightest meteor observed, the ionization increased to a value in excess of summer noon conditions.

It is pointed out that meteoric showers might take place in the F region which would be unobservable by ordinary visual means.

Taking into account the energy spent by the meteor in ionization, a mass for the brightest meteor, for which correlative data was obtained is roughly calculated to be 0.3 gram. Its estimated brightness was -1 magnitude.

The recombination coefficient at the height of the E region is calcu-

[8] *Proc. I.R.E.*, February, 1935.

lated from the rate of decrease of ionization after the passage of a meteor, to be less than 0.2×10^{-8} cubic centimeters per second.

Small Sapling Method of Evaluating Wood Preservatives.[9] R. E. WATERMAN and R. R. WILLIAMS. Permanence and toxicity are probably the most necessary characteristics of a wood preservative. Ease of injection, freedom from corrosive properties, cleanliness, cost, and the like are all important, but no material can be considered unless it displays a high degree of resistance to wood-destroying fungi and unless this toxic potency persists when the treated wood is exposed to the weather for long periods of time. The problem under discussion is that of appraisal of wood preservatives for these two characteristics within a reasonably short time.

In order to expedite tests of the permanency of pole preservatives, use is made of groups of small pine saplings treated with the preservative in question and set in the ground as miniature telephone poles. In these specimens weathering is relatively rapid on account of the large ratio of surface to volume, and poorly preserved material begins to decay in one or two years. Analyses and toxicity tests as well as observations of decay are made periodically. Seven years' experience indicates that the comparative preservative values of various salts, creosotes, oils, etc., may be estimated relatively cheaply, quickly, and with considerable reliability by this method.

A High Speed Level Recorder for Acoustic Measurements.[10] E. C. WENTE, E. H. BEDELL and K. D. SWARTZEL, JR. Two quite accurate means for recording rapid variations in sound intensity in a form suitable for visual inspection have been available for a number of years. One of these is the phonodeik, or one of its variants, and the other is a combination of a microphone and an oscillograph. When properly designed these devices record the actual wave form of the sound. However, for many acoustic measurements, a knowledge of the wave form is of secondary interest, whereas it is important that one should be able to record rapidly varying mean intensities over a wide range of values. From a record of the wave form it is not easy to determine the intensity with any degree of accuracy for a range greater than 20 or 30 db, but in some types of acoustic measurements it is highly desirable that the record cover a range of at least 60 db. Recently several types of instruments have been built which record, on a logarithmic scale, the mean power of the electrical input. These instruments, like that described here, may be used to plot the intensity

[9] *Indus. and Engg. Chem.* (*Analytical Edition*), November 15, 1934.
[10] *Jour. Acous. Soc. Amer.*, January, 1935.

level in db as a continuous function of either time, frequency, or any other variable. The adaptability of such level recorders to acoustic measurements depends, among other factors, upon the range and accuracy of the logarithmic scale, and upon the effective speed of the recording mechanism. This recording speed is most conveniently expressed in terms of the rate, in db per second, at which the recorder is capable of following changes in the input power.

The level recorder described here consists essentially of an amplifier and rectifier, the output current of which is held at a substantially constant value automatically by a change in the gain of the amplifier, following changes in input power. The gain is varied by means of motor driven slide wire potentiometers graduated in logarithmic steps, the gain settings of which are recorded.

Some Applications of Modern Acoustic Apparatus.[11] S. K. Wolf and W. J. Sette. Within the past two years there have been developed at the Bell Telephone Laboratories several electro-acoustic instruments designed to facilitate accurate measurement of a wide variety of acoustic phenomena. Three of these instruments are: an automatic level recorder, a crystal analyzer, and an acoustic spectrometer. Some of the types of acoustic studies for which these modern devices are well adapted may be of general interest and hence specific applications made at Electrical Research Products are described here. These include reverberation measurements, loud speaker response measurements, noise analyses, piano tone analyses, and studies on the singing voice. A brief description of the operating characteristics of the instruments is first given.

[11] *Jour. Acous. Soc. Amer.*, January, 1935.

Contributors to this Issue

ALFRED C. BECK, E.E., Rensselaer Polytechnic Institute, 1927. Test Department, New York Edison Company, summers 1926 & 27. Instructor in mathematics, Rensselaer Polytechnic Institute, 1927–28. Member, Technical Staff, Bell Telephone Laboratories, 1928–.

W. R. BENNETT, B.S., Oregon State College, 1925; A.M., Columbia University, 1928. Bell Telephone Laboratories, 1925–. Mr. Bennett has been engaged in the study of the electrical transmission problems of communication.

H. W. BODE, A.B., Ohio State University, 1924; M.A., 1926; Ph.D., Columbia University, 1935. Bell Telephone Laboratories, 1926–. Dr. Bode has been engaged in the study of transmission networks, such as wave filters, attenuation equalizers, and phase correctors.

R. P. BOOTH, S.B. in Electrical Engineering, Massachusetts Institute of Technology, 1925. American Telephone and Telegraph Company, Department of Development and Research, 1925–34; Bell Telephone Laboratories, 1934–. Mr. Booth's work has had to do mainly with studies of crosstalk in open-wire and toll cable circuits.

EDMOND BRUCE, B.S., Massachusetts Institute of Technology, 1924. Radio service, U. S. Navy, 1917–19. Western Electric Company, 1924–25; Bell Telephone Laboratories, 1925–. Mr. Bruce has been engaged in the development of short-wave radio receivers and field-strength measuring equipment, and has specialized in directive antenna systems for short-wave radio communication. He was awarded the Morris Liebmann Memorial Prize by the Institute of Radio Engineers in 1932.

CHARLES R. BURROWS, B.S. in Electrical Engineering, University of Michigan, 1924; A.M., Columbia University, 1927. Western Electric Company, Engineering Department, 1924–25; Bell Telephone Laboratories, Research Department, 1925–. Mr. Burrows has been associated continuously with radio research and is now in charge of a group investigating the propagation of ultra-short waves.

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.

Alfred Decino, B.S., University of Colorado, 1928. Bell Telephone Laboratories, 1928–. Mr. Decino has been associated continuously with radio research and chiefly in studies of the propagation of radio waves.

R. L. Dietzold, S.B., Massachusetts Institute of Technology, 1925; Ph.B., Yale University, 1927; Department of Mathematics, Polytechnic Institute of Brooklyn, 1928–30. Bell Telephone Laboratories, 1930–. Mr. Dietzold has been concerned with research and design studies on transmission networks.

Loyd E. Hunt, A.B. in physics, Reed College, 1927. Wireless operator and ship wireless inspector, 1918–26. Radio Instructor, Oregon Institute of Technology, 1926–27. Director Radio-Electric Division, United Y.M.C.A. Schools, Seattle, Washington, 1927–29. Bell Telephone Laboratories, 1929–. Since coming with the Laboratories, Mr. Hunt has been employed in short-wave antenna development and is now engaged in ultra-short-wave propagation studies.

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 toll cable circuits.

Robert M. Kalb, B.E.E., Ohio State University, 1927; Ohio State University, 1927–28. Bell Telephone Laboratories, 1928–. Mr. Kalb has been engaged in analyzing transmission problems, principally those of magnetic character.

H. P. Lawther, Jr., B.A., University of Texas, 1912; M.A., Harvard, 1913; Ph.D., Harvard, 1916. Southwestern Bell Telephone Company, 1919–. Transmission and Protection Engineer, Texas Area, 1928–32; General Transmission and Protection Engineer, 1932–.

VOLUME XIV JULY, 1935 NUMBER 3

THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

Further Results of a Study of Ultra-Short-Wave Transmission Phenomena—
C. R. Englund, A. B. Crawford and W. W. Mumford 369

Acoustical Instruments—*E. C. Wente* 388

Thermionic Electron Emission—*J. A. Becker* . . . 413

Radio Propagation Over Spherical Earth—
Chas. R. Burrows 477

A Single-Sideband Short-Wave System for Transatlantic Telephony—
F. A. Polkinghorn and N. F. Schlaack 489

Mutual Impedances of Parallel Wires—
Ray S. Hoyt and Sallie Pero Mead 509

An Unattended Ultra-Short-Wave Radio Telephone System—*N. F. Schlaack and F. A. Polkinghorn* 534

Around the World by Telephone 542

Abstracts of Technical Papers 544

Contributors to this Issue 550

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

50c per Copy *$1.50 per Year*

THE BELL SYSTEM TECHNICAL JOURNAL

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

EDITORIAL BOARD

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

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, 1935

PRINTED IN U. S. A.

The Bell System Technical Journal

Vol. XIV *July, 1935* *No. 3*

Further Results of a Study of Ultra-Short-Wave Transmission Phenomena *

By C. R. ENGLUND, A. B. CRAWFORD and W. W. MUMFORD

Earlier published work has shown that, while the chief features of ultra-short-wave transmission over "air line" ranges are calculable from optical theory, there are deviations from this theory at the greater distances where a bending around the earth occurs. In this paper the results of a further study of these phenomena are given. It is shown that transmission to regions beyond the optical range is determined by conditions which are not constant and which, in fact, can produce great signal strength changes. The variable percentage of water vapor normally present in the atmosphere is suggested as a possible cause. The explanation seems, therefore, to involve a combination of diffraction and refraction, this latter variable with time, and at times predominant.

IN a recently published paper [1] results obtained at the Holmdel Laboratory during a survey of ultra-short-wave transmission phenomena have been given. In this report it was shown that, while the chief features of ultra-short-wave transmission over "air line" ranges are calculable from optical theory, there are deviations from this theory at the greater distances where a diffraction around the earth occurs. The results of a further study of these diffraction phenomena form the data of this paper.

It is probable that a diffraction around the earth will be distorted by major topographical irregularities, at or near the area of grazing incidence for the waves, and hence that the ocean surface is preferable for a study of this kind. It hardly seems likely that the ocean contour can be rough enough to give results differing markedly from those for a smooth water surface.

An obvious experimental setup, therefore, is to locate a transmitter at or very near the ocean shore and to record the transmitter field as a mobile receiver is carried towards or away from this transmitter, on paths that go well below the horizon. The receiver can be carried

* Presented at April 1935 meeting of Union Radio Scientifique Internationale, Washington, D.C.

[1] *Proc. I. R. E.* 21, 464, 1933.

either by boat or airplane. It is probable that the wave structure of the ocean surface would produce irregularities in reception if the receiving antenna be too near this surface, as on a boat, and for this type of receiver transport the time occupied by an experiment is rather long. Naturally the slow motion makes a fine grained record possible. In an airplane the time of transit is very much reduced but the vibration and unsteadiness are not favorable for accurate recording and the electrical noise level is high. There is also an increase in range necessary to get the same angular distance below the horizon as with a boat. This range extension, however, is relatively not as great as might appear since the falling away below the horizon is proportional to the square of the distance. For example, a line tangent to the earth is 100 feet up at 14 miles and 1000 feet up at 45 miles from the tangent point.[2] If these be boat and airplane antenna altitudes respectively, the airplane must always travel 31 miles farther to get the same angle of refraction below the horizon as the boat does. Since it is necessary to travel about 92 miles to get one degree below the former horizon (angle between the two earth radii) and the transmitting antenna height will further increase the range for a given diffraction geometry, it is evident that the difference in antenna altitude as between a boat and an airplane, is not of serious effect either in space covered or in accompanying signal attenuation.

We were fortunate in being located so that a land plane could be used to give us an over-water transmission. As a glance at the map (Fig. 1) will show, it is possible so to locate a transmitter on the New Jersey shore that there is an over-water path for an airplane flying along the Long Island shore. Owing to the curvature of the Long Island beach, an over-water path for the entire distance to Montauk Point requires a location of the transmitter at or south of Long Branch, New Jersey, and such a location makes the minimum path length possible (Long Branch to Rockaway Beach) about 20 miles. This was too great a distance to be satisfactory to us and we elected to locate north of Long Branch. Although the curvature of the Long Island shore then interposed land between Montauk Point and the transmitter, this land lay well below the horizon, as viewed from the transmitter, and it was thought, therefore, that its effect might be small or negligible.

North of Long Branch the favorable shore transmitter sites are restricted to the Sandy Hook region and the stretch between Sea

[2] Air refraction is included by increasing the apparent radius of the earth to 5260 miles. See Schelleng, Burrows and Ferrell, *Proc. I. R. E.* **21**, 427, 1933; *Bell Sys. Tech. Jour.*, XII, 125, 1933.

Bright and Long Branch. Between these two regions the narrow sand strip is almost entirely occupied by breakwater, railroad, and highway with a number of pole-carried transmission lines. Sandy Hook has several splendid locations, but housing and 60-cycle power would have had to be supplied, and we elected, therefore, to transmit from the Calef estate in North Sea Bright. This house, the last one along the beach north of Long Branch, was within 50 feet of the ocean and already had electric power connections.

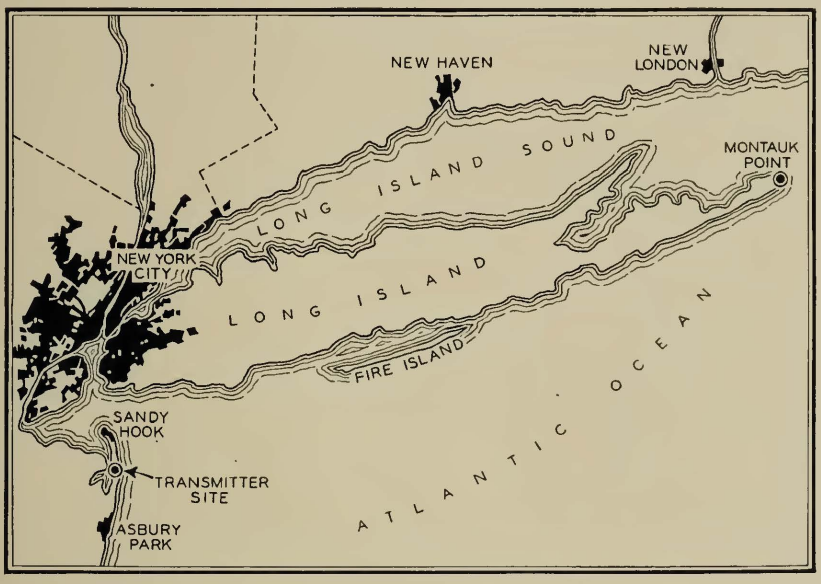


Fig. 1—Map showing transmitter location. Airplane flights were made along the south shore of Long Island.

Experimental

Our experimental setup was as follows: In the Bell Laboratories' tri-motor Ford plane, placed at our disposal by Mr. F. M. Ryan and his staff, two receivers operating at 1.58 and 4.6 meters respectively, and equipped with manual recorders as described in the earlier cited paper, were installed. The insulated mast antenna support in the tail of the plane was used as the 4.6 meter receiving antenna and connected as an approximately ½ wave, end tapped, conductor. The 1.58 meter antenna was a tubular half wave antenna, cut in the center and connected to an internal two-wire transmission line. It was unbalanced but apparently not seriously so. Its mounting socket was up forward between the wings. These receivers were double

detection sets with 100 decibels or more gain at the intermediate frequency and the spring operated recording mechanism recorded the set gain as the operator varied it manually, to hold the set output constant.

At Sea Bright two transmitters were installed in separate rooms on the top floor of the house and the high-frequency power was fed to the antennas by transmission lines. These antennas were center-driven vertical half-wave units mounted on wood beams which were erected in the gables of the house and extended above the roof. The antenna centers were about 8 feet above the roof peak and some 60 feet above mean sea level. Both sets were simple " push-pull " oscillators, the 4.6 meter one using two UX852 tubes and generating something like 80 watts, the 1.58 meter set using two 149Y tubes and generating 12 watts. Meters were arranged so as to maintain a check on the constancy of the antenna currents. Modulating equipment and a 3–5 megacycle receiver were provided so that contact with the plane

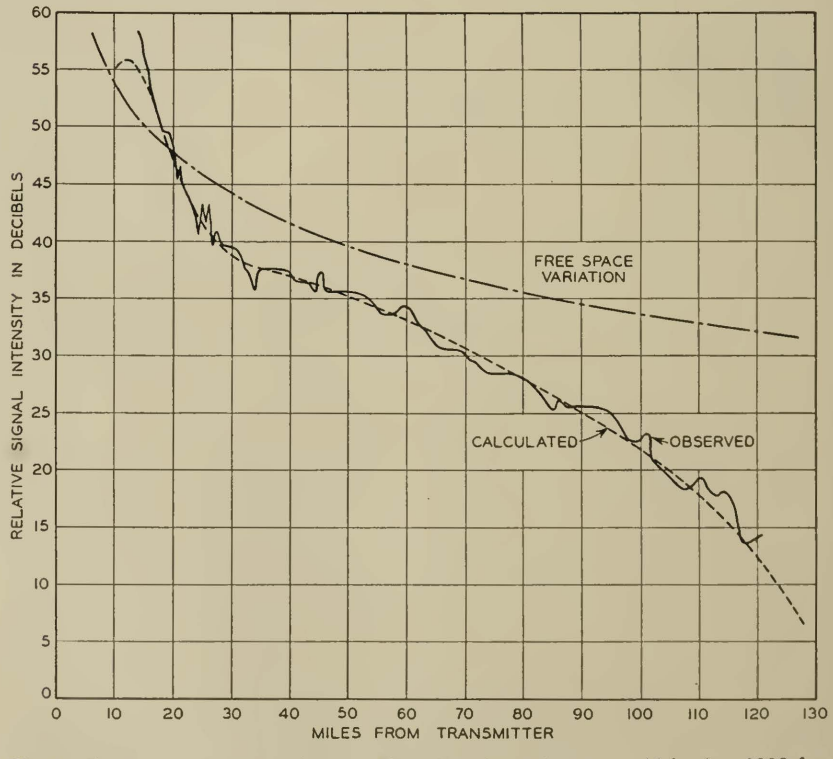


Fig. 2—Flight toward transmitter. Wave-length—4.6 meters; Altitude—8000 feet; September 27, 1933; 11:30 a.m. to 1:00 p.m.

could be maintained at all times, transmitting from Sea Bright on 4.6 or 1.58 meters and receiving on the plane's regular service wave. Contact of this kind is well nigh indispensable.

From September 25, 1933 to November 20, 1933, inclusive, fourteen airplane runs were made, ten of which were recording trips. Measurements were made both " go " and " return," and of the twenty observations resulting, three were made at 8000 feet, four at 2500 feet, two

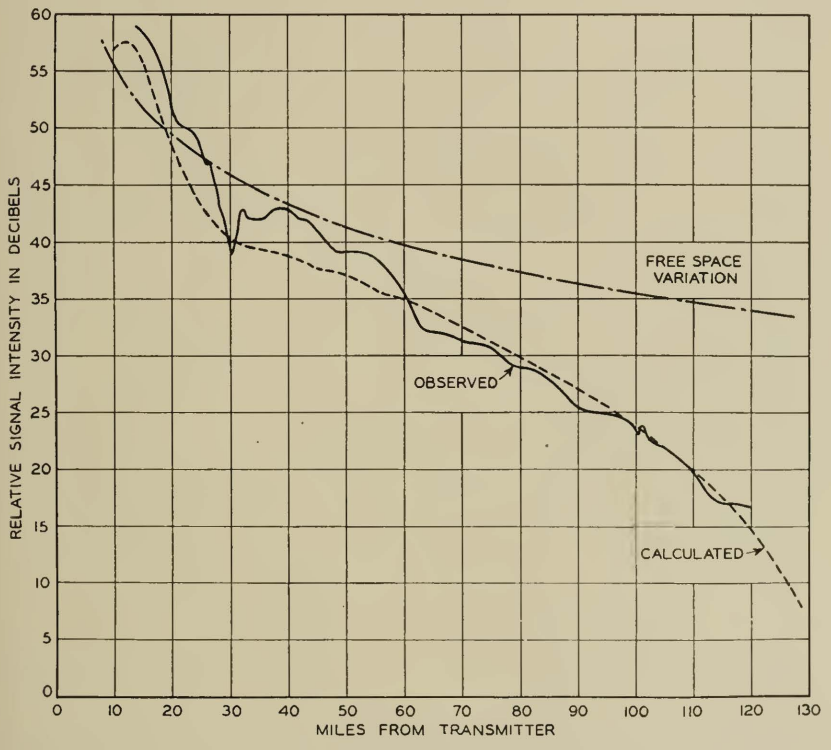


Fig. 3—Flight from transmitter. Wave-length—4.6 meters; Altitude—8000 feet; October 3, 1933; 10:20 a.m. to 11:35 a.m.

at 2000 feet, and the remainder, or eleven, at 1000 feet. No observations at 2000 feet had been scheduled but on September 28, when a 2500-foot run was begun, clouds forced a drop to 2000 feet. Each round trip lasted from two to five hours and due to the exigencies of airplane operation was completed between 9 a.m. and 5 p.m. In Figs. 2 to 12 inclusive, a set of typical observations is plotted. Superposed on the observational curves are theoretical curves cal-

culated from the height and distance data and the constants [3] of the sea water, assuming ordinary optical reflection from an earth of 5260 miles radius. These theoretical curves are adjusted best to fit with the observations, the ordinates for all the curves being the decibels left in the receiver attenuator.[4] These results can be summarized briefly as follows:

At 8000 feet the fit with theory is excellent at both wave-lengths. The grazing distance for this altitude is 137 miles and the entire

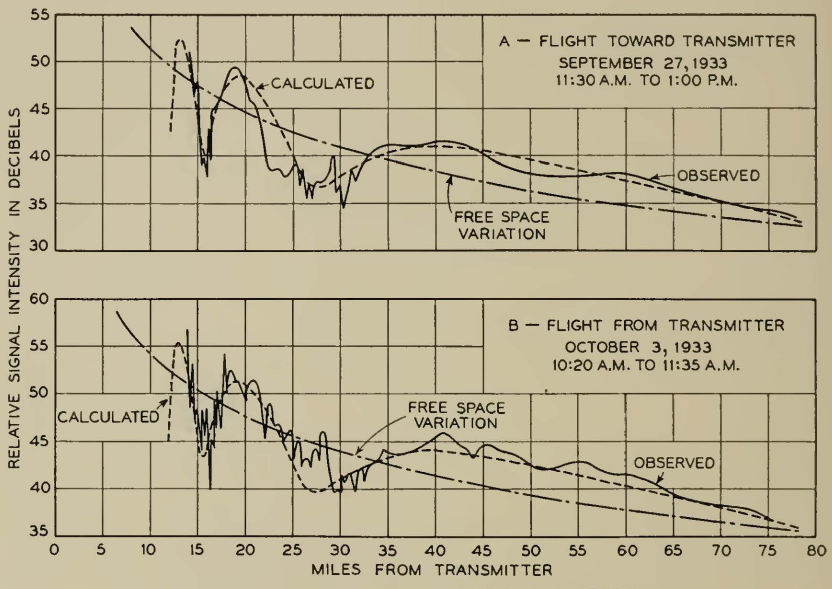


Fig. 4—Wave-length—1.58 meters; Altitude—8000 feet.

reception is, as it should be, optical. The " out " curves for the 1.58 meter reception show a middle distance roughness which characterizes all the " out " curves for this wave-length. This roughness is due to a minimum in the polar characteristic of the plane which was, unfortunately, directed at the transmitter for the first part of the outward flights.

At 2500 feet the fit with theory is good for the greater part of the optical range for the 4.6 meter wave transmission. Both curves (5 and 6) show a definite diffraction effect.

[3] Dielectric constant = 80.
Ohms per cm. cube = 20.

[4] The set gain was determined together with the average transmitter ammeter readings for each run. With these and the experimentally determined polar characteristics of the plane antennas, the curves are corrected to set gains of 100 and 110 db respectively for the 4.6 and 1.58 meter receivers and for specified transmitter currents.

The 1.58 meter observations at 2500 feet fall off with distance in fair agreement with theory (Fig. 7). The " out " curve is rough at the shorter distances, as explained above.

Observations at an altitude of 1000 feet occupied most of our flying time. This was the lowest altitude which we cared to try, since the plane had to remain within gliding distance of the shore. The

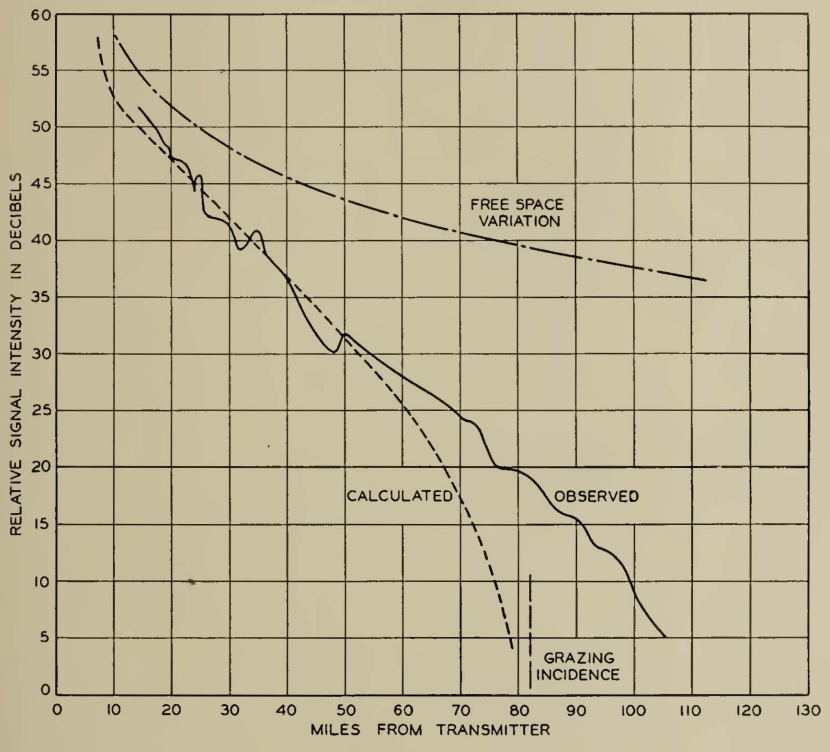


Fig. 5—Flight from transmitter. Wave-length—4.6 meters; Altitude—2500 feet; October 11, 1933; 10:35 a.m. to 11:30 a.m.

grazing distance for this altitude is 56 miles and, as the Montauk end of Long Island is a little over 120 miles out, a considerable distance where the transmission is below the earth's horizon was available. At 1000 feet, at Montauk, the plane was 5000 feet, or 0.71 degree, below the ocean grazing line from the transmitter.

At 4.6 meters the first run (the flight of September 27), Fig. 8, carried all the way out to Montauk. Subsequent runs (as Fig. 9) carried barely half way, before the plane noise drowned out the signal.

By going over the electrical system of the plane, tightening old bonds and loose metal pieces, and adding new bonds and shielding, the noise was reduced to such an extent that the results shown in Fig. 8 for the flight of November 1 were obtained, where Montauk Point was almost reached. Both of these curves fit theory well in the optical range;

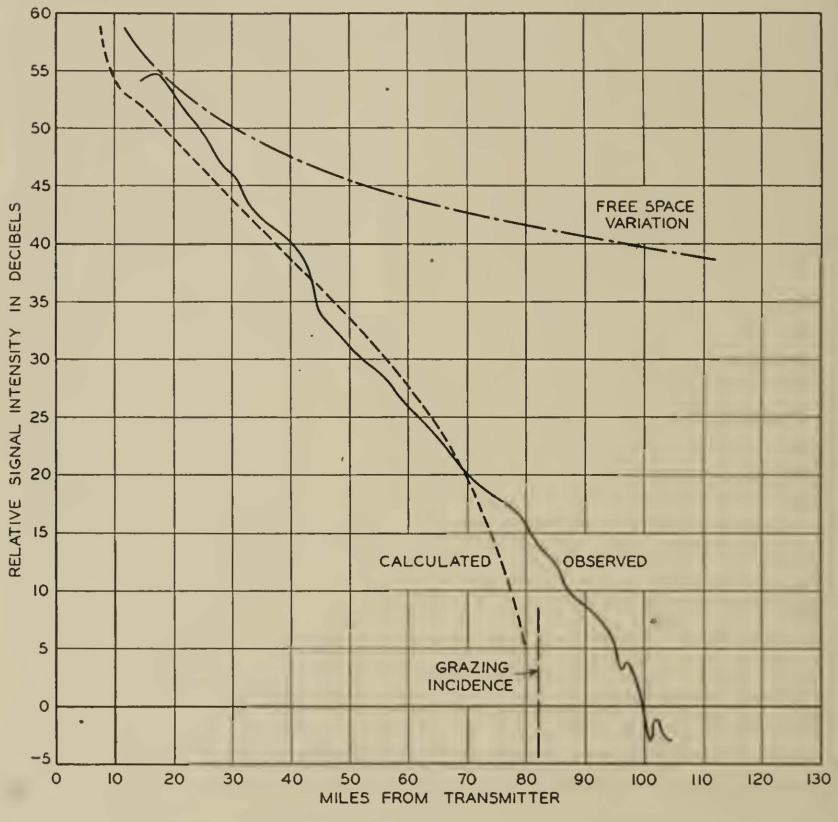


Fig. 6—Flight toward transmitter. Wave-length—4.6 meters; Altitude—2500 feet; November 16, 1933; 12:10 p.m. to 1:05 p.m.

beyond this the second curve lies 10 decibels below the first one. Evidently the plane noise troubles were due simply to the lower signal level which had to be received. This level fell lower as the cold weather came on, and additional work on the plane electrical system had to be done. The 4.6 meter receiving set was also overhauled and realigned. Finally, on November 20, we obtained the bottom curve of Fig. 8, which appeared to be the best we could hope for, and the

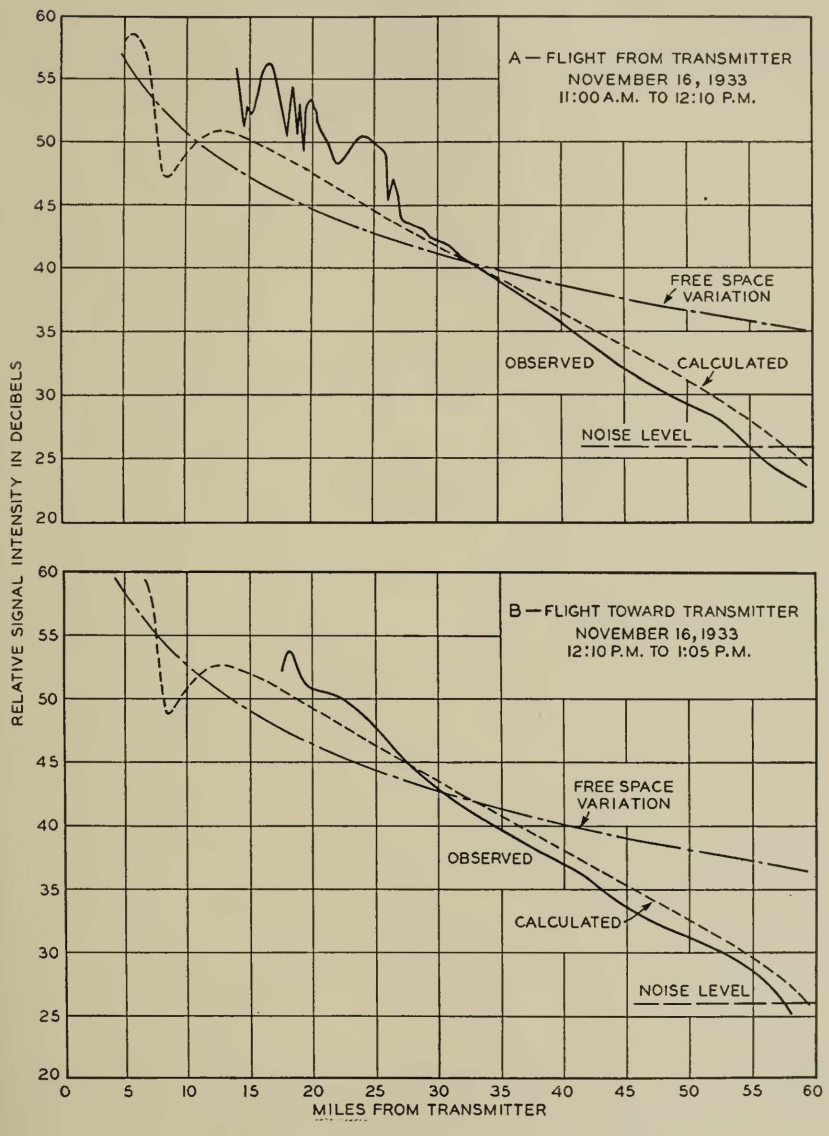


Fig. 7—Wave-length—1.58 meters; Altitude—2500 feet.

work was accordingly discontinued. For this flight, the signal strength at 80 miles was ten decibels below that observed on November 1, and twenty-one decibels below that recorded September 27. The conclusion is inescapable that transmission to regions beyond the optical range is determined by conditions which are not constant and which in fact can produce great signal strength changes.

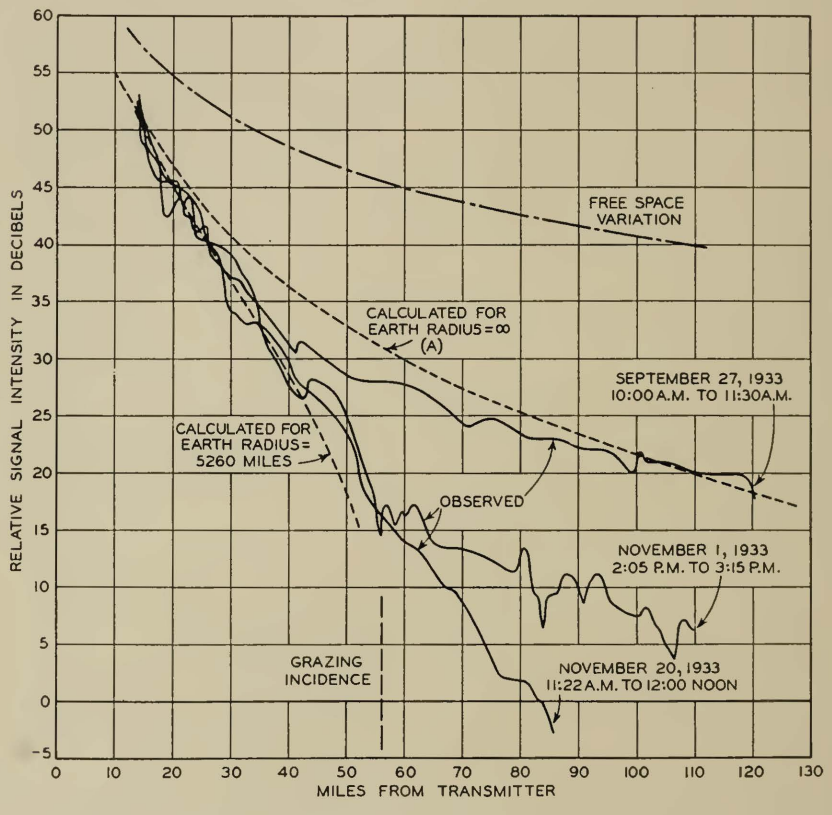


Fig. 8—Flights from transmitter. Wave-length—4.6 meters; Altitude—1000 feet.

The 1.58 meter observations at 1000 feet check the 4.6 meter results fairly well. They follow theory within the optical range. Figure 10, corresponding to Fig. 8 for the 4.6 meter observations, shows some indication of diffraction, and Fig. 11 checks the rapidly falling signal intensity of the bottom curve of Fig. 8. At no time was the half way distance to Montauk reached. There are several reasons for this. The power level available at 1.58 meters was nearly 10 db below that

for 4.6 meters, and the plane noise level was higher. It appears, from some rough tests made, that a metal plane is likely to have a peak noise range determined by the natural period of the smaller metal parts, which can vibrate and make variable contact during operation. The " out " curves show the same roughness that was found at the other altitudes. If this effect had been discovered in time, it would

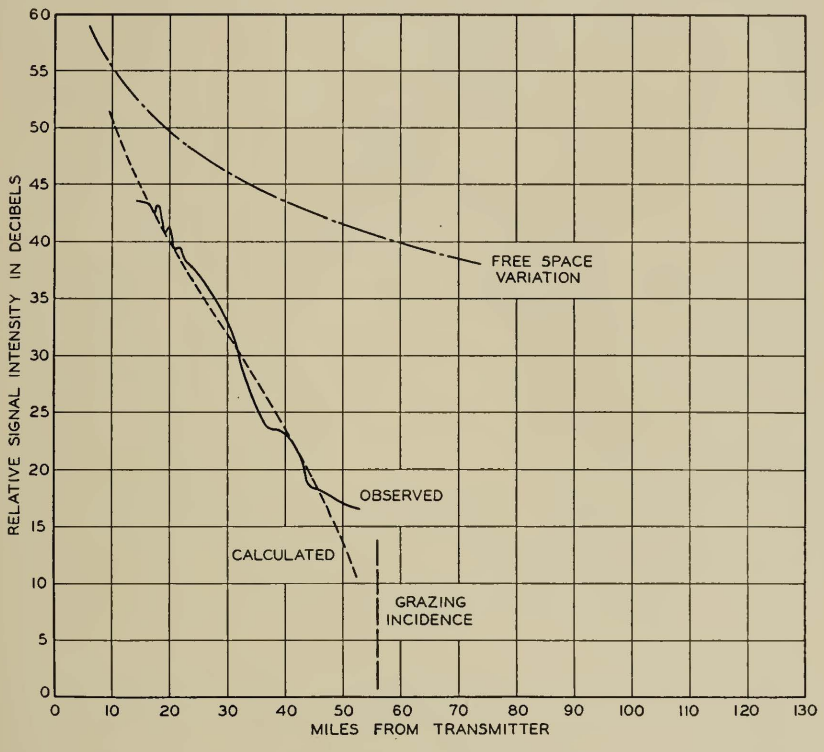


Fig. 9—Flight toward transmitter. Wave-length—4.6 meters; Altitude—1000 feet; October 3, 1933; 11:30 a.m. to 12:35 p.m.

certainly have been advisable to move the receiving antenna, so as to shift the polar characteristic minimum to some other angle.

The curves of Fig. 12 were taken at 4.6 meters on passing from the 1000- to the 8000-foot level and vice versa. The first one, taken at Montauk Point on September 27, shows very little variation in signal strength in spiraling up from 1000 to 8000 feet. This was the day when our maximum atmospheric refraction was encountered. If we assume a refraction sufficient to bend the radiation into a circle around

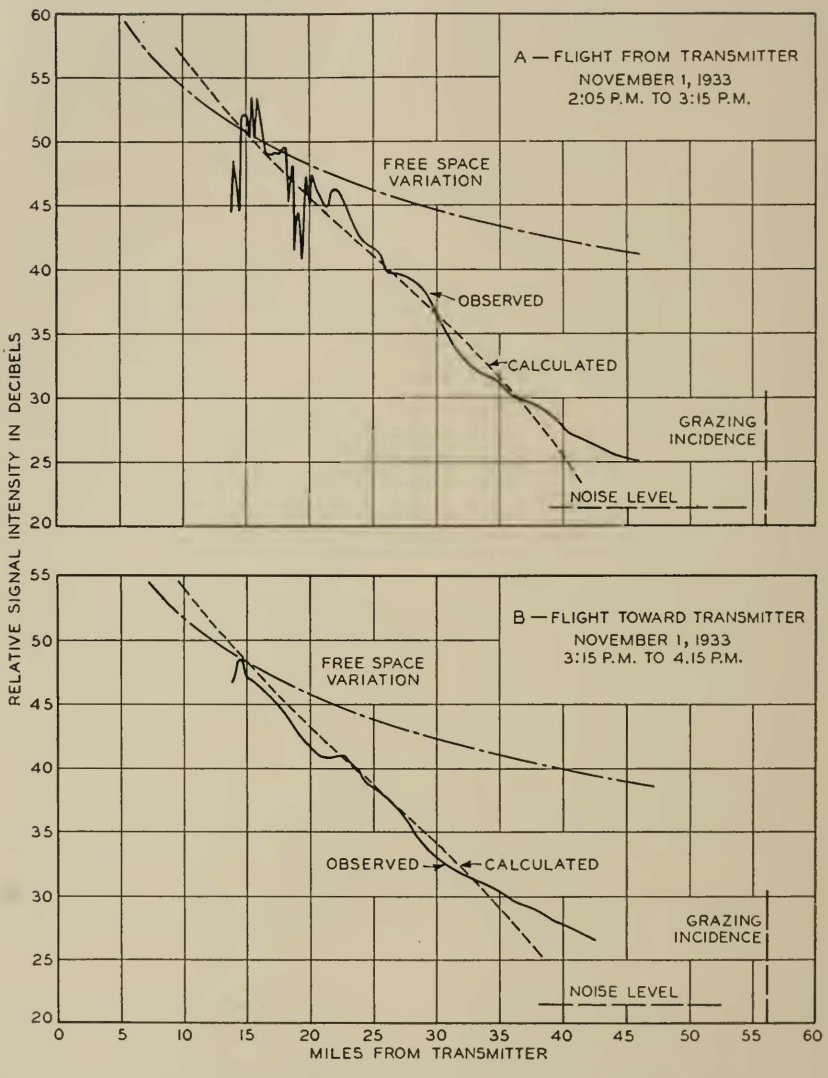


Fig. 10—Wave-length—1.58 meters; Altitude—1000 feet.

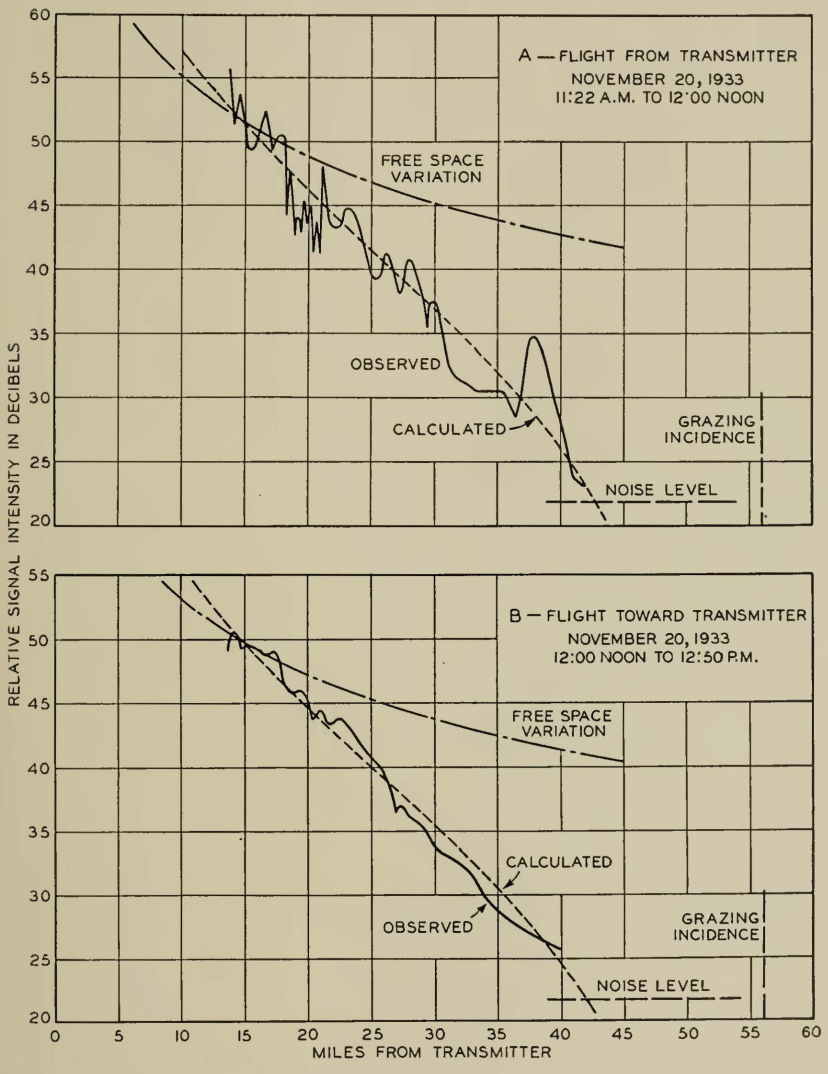


Fig. 11—Wave-length—1.58 meters; Altitude—1000 feet.

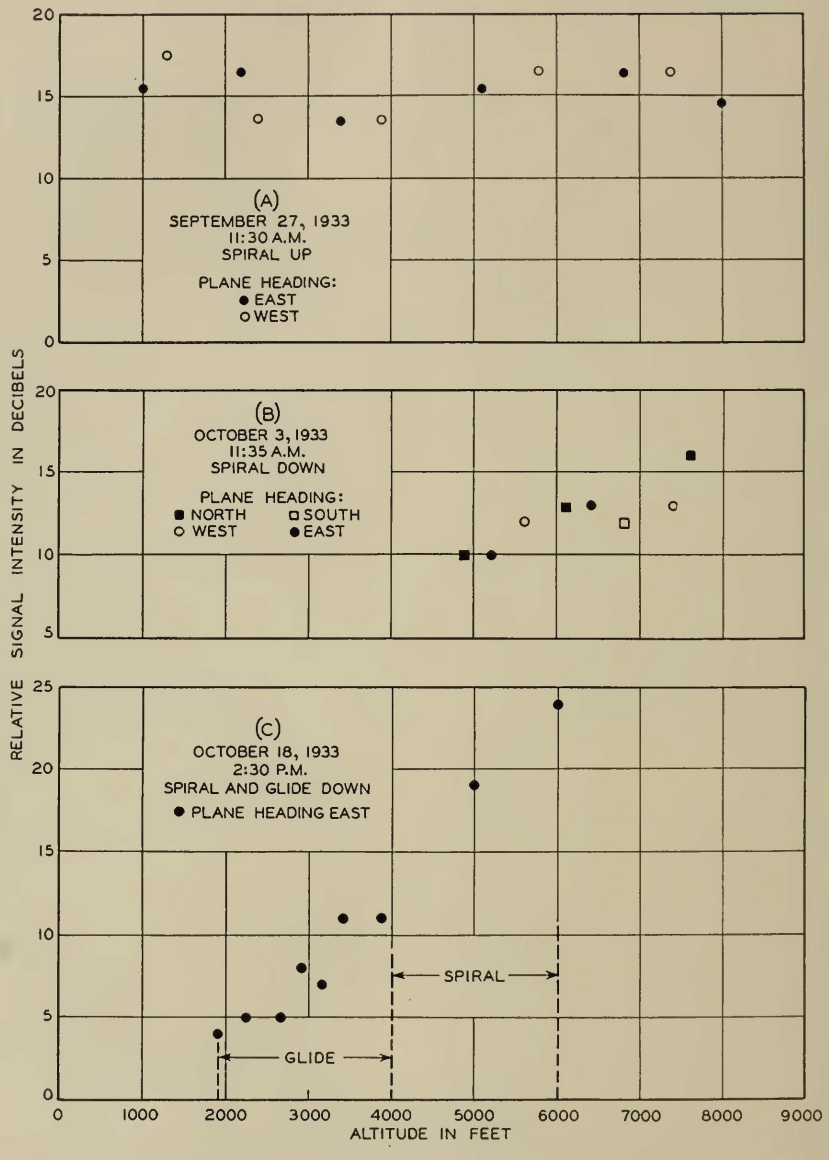


Fig. 12—Signal strength as a function of altitude. Wave-length—4.6 meters. Data for Figs. *A* and *B* obtained at Montauk Point. Data for Fig. *C* obtained 90 miles from transmitter.

the earth, we shall have the theoretical curve " A " of Fig. 8. Evidently the actual refraction was something of this order.

The second curve of Fig. 12, was taken spiraling down at Montauk Point on October 3. The third curve was taken at a distance of 90 miles from the transmitter while spiraling and gliding down, headed " out," on October 18. Unfortunately this was a day of low humidity and the signals were lost before reaching the 1000-foot level and were not picked up again on the return trip. The reduction in plane noise while idling the engines, after the descent had begun, was pronounced, and permitted following the signal almost to 1000 feet.

Discussion

The check with theory for the optical range, though gratifying, was expected from our earlier work. At these low angles of incidence the reflectivity does not vary rapidly with dielectric constant and conductivity changes so that the values assumed by us are adequate. The results for the range beyond grazing incidence are rather unexpected. Diffraction was looked for, and anticipated, but the results themselves seem most readily explicable by a combination of diffraction and refraction, the latter variable with time, and at times predominant. Apparatus variations are ruled out; no effect of ocean roughness has been discernible, and calculations show that the height of the tide is not the explanation. There remain changes in the constants of the air, or in other words, changes in air refraction to consider. Not enough data are available to predict correctly an air refraction effect, but that this is the most plausible explanation is shown by the following.

Because of the change in air density with height, the effect of air refraction is to bend the radio ray into a curved path. This bending is proportional to the gradient of the dielectric constant of the atmosphere, which in turn is proportional to the sum of the gradients of the dry air and water vapor constituting the atmosphere. The dielectric constant of a gas can be written as,

$$\epsilon - 1 = K\frac{p}{T}.$$

In the table below some calculated values of " K " are given. They indicate that water vapor is some 18 times as effective as air, as a refractive medium. They are for wave-lengths greater than 100 meters; no measurements at about 5 meters wave-length have been found, so far, in the scientific literature.

760 Mm. Barometric Pressure

	K	
Temp.	Air	Water Vapor
45° F.	0.000211	0.00381
63° F.	0.000211	0.00366
83° F.	0.000211	0.00356

As is shown in the appendix, the small percentage of water vapor, present normally in the atmosphere, has a very marked effect on the radius of curvature of the radio ray. While the bending, there calculated, does not quantitatively explain our September 27, 1933 results at 1000-foot altitude, it is qualitatively in the right direction. The weather bureau data given us were,

Date	Temp.	Bar.	Per Cent Water Vapor by Vol.
Sept. 27, 1933	83° F.	760	2.46
Nov. 1, 1933	63° F.	763	0.935
Nov. 20, 1933	45° F.	757	0.617

and were taken on top of a New York City building. The humidity and its gradient, at the ocean surface may well have been greater. It is not impossible, either, that water vapor absorption bands occur in the ultra-short-wave region. High and irregular refraction effects would then occur.

It is evident that a slight change in ray curvature under grazing incidence transmission conditions will give a marked increase in range. The fading of weak signals under these conditions, which has been observed in this country by Bell Laboratories engineers at Deal Beach, New Jersey, and by Radio Corporation of America observers, and in Europe by Senatore Marconi and International Telephone and Telegraph Company engineers, may possibly be explained in this manner.

This work is being continued.

APPENDIX

1. From the accepted theory [5] the dielectric constant of a gas is given by

$$\frac{\epsilon - 1}{\epsilon + 2} = \frac{4\pi N}{3}\left(\lambda e^2 + \frac{A\mu}{3RT}\right).$$

[5] See Debye, "Polar Molecules," Chemical Catalogue Company.

For a perfect gas $pv = \frac{R}{M} T$ or $\rho = \frac{1}{v} = \frac{Mp}{RT}$ and since $N = \frac{A\rho}{M}$

$$\frac{\epsilon - 1}{\epsilon + 2} = \frac{4\pi}{3} \cdot \frac{A}{R} \cdot \frac{p}{T} \left(\lambda e^2 + \frac{A\mu}{3RT} \right).$$

The symbols are:

ϵ = dielectric constant
N = no. of molecules per cm.^{3}
λ = elastic binding constant of optical electrons
e = electron charge ($= 4.774 \times 10^{-10}$ E.S. Units)
μ = electric moment of molecule
R = gas constant ($= 8.314 \times 10^7$)
A = Avogadro's const. ($= 6.064 \times 10^{23}$ molecules per mole)
T = absolute temperature
p = pressure in dynes per cm.^{2}
v = specific volume (cc. per gm.)
M = mole (molecular wt. in gms.)
ρ = density (gms. per cm.^{3})

For practical purposes this equation can be simplified. For most gases $\epsilon + 2 = 3$, to a high degree of accuracy and, as the material constants "λ" and "μ" are not readily separately measurable, it is convenient to lump them in a single constant. It is also convenient to change to another unit of pressure, the millimeter of mercury. In this unit and with the above simplifications

$$\epsilon - 1 = K \cdot \frac{p}{T} \text{ and the gas equation becomes } \frac{p}{T} = \frac{62370}{M} \rho.$$

2. From the Smithsonian tables we have the value of "K" for air practically constant and equal to

$$K_{\text{air}} = 211 \times 10^{-6}.$$

From the results of Jona, Zahn, Stuart, Sänger and Stranathan,[6] the value of "K" for water vapor is

$$K_{H_2O} = 182 \times 10^{-6} \left(1 + \frac{5582}{T} \right).$$

From these values the table below is calculated, assuming 760 mm. of mercury pressure. The temperatures chosen are those encountered in our airplane work, on the dates given.

[6] Jona, *Phys. Zeit.* **20**, 14, 1919. Zahn, *Phys. Rev.* **27**, 329, 1926. Stuart, *Zeit. f. Phys.* **51**, 490, 1928. Sänger, *Phys. Zeit.* **31**, 306, 1930. Stranathan, *Amer. Phys. Soc. Bull.*, Vol. 9, No. 2, abstract No. 7.

Temp.	ϵ_{air}	$\epsilon_{water\ vapor}$	Date
45° F.....	1.000574	1.01033	11–20–33
63° F.....	1.000554	1.00965	11– 1–33
83° F.....	1.000534	1.00898	9–27–33

3. The velocity of propagation, of electromagnetic waves in a gas, is given by $v = \frac{3 \times 10^{10}}{\sqrt{\epsilon}}$, and the radius of curvature of the ray, in the plane of " h," is,

$$R = \frac{v}{\frac{\partial v}{\partial h}} = -\frac{2\epsilon}{\frac{\partial \epsilon}{\partial h}} = -\frac{2}{\frac{\partial\left(K\frac{p}{T}\right)}{\partial h}} = -\frac{2M}{62370\frac{\partial(K\rho)}{\partial h}}.$$

From the linear addition theorem, the " K " of a composite gas like moist air will be

$$100K = \left[(100 - \alpha)211 + \alpha\left(182\left(1 + \frac{5582}{T}\right)\right)\right] \times 10^{-6}$$

or

$$K = \left[211 + \alpha\left(\frac{10159}{T} - 0.293\right)\right] \times 10^{-6},$$

where " α " is the percentage of water vapor, in the air, by volume. Hence

$$R = -\frac{M \times 10^{6}}{31185} \cdot \frac{1}{\frac{\partial}{\partial h}\rho\left[211 + \alpha\left(\frac{10159}{T} - 0.293\right)\right]},$$

where " M " is the molecular weight of the gas involved, and " ρ," " α," and " T " are to be determined as functions of " h," the altitude above the earth.

4. From Humphrey's " Physics of the Air " we obtain the following data:

On page 38 average summer and winter temperature versus height curves are given. For the first two kilometers a good fit to the summer curve is given by the equation,

$$T = -6.19h + 288 \quad \text{where " } h \text{ " is in miles.}$$

On page 72, average summer and winter air and water vapor pressure, and total density tables, as a function of the height, are given. For the first two kilometers the density is given by the equation,

$$\rho = -0.000185h + 0.001224$$

and the percentage of water vapor by

$$\alpha = -0.405h + 1.372.$$

The water vapor percentage gradient increases (in absolute value) up to 1.25 kilometers, after which it decreases; the other two curves ("T" and "ρ") do not have a point of inflection. The curve for "ρ" has a continuously falling slope above 2 kilometers, that for "T" has a rising slope (both in absolute value). Either rising slope curve should show, by itself, a certain converging lens effect.

5. Carrying out the differentiations indicated in paragraph 3 gives

$$\frac{\partial}{\partial h}\rho\left[211 + \alpha\left(\frac{10159}{T} - 0.293\right)\right]$$
$$= 211\frac{\partial\rho}{\partial h} + \left(\frac{10159}{T} - 0.293\right)\left(\alpha\frac{\partial\rho}{\partial h} + \rho\frac{\partial\alpha}{\partial h}\right) - \frac{10159\alpha\rho}{T^2}\frac{\partial T}{\partial h}.$$

Three terms result, distinguishable, respectively, as due to the air density gradient, the water vapor density gradient, and the temperature gradient.

As a typical and simple numerical example, we may select the values of "ρ," "T," and "α" for $h = 0$, that is at the earth's surface. We have then

$$\frac{\partial\rho}{\partial h_0} = -1.85 \times 10^{-4}, \qquad \alpha_0 = 1.372,$$
$$\rho_0 = 1.224 \times 10^{-3}, \qquad \frac{\partial T}{\partial h_0} = -6.19,$$
$$\frac{\partial\alpha}{\partial h_0} = -0.405, \qquad T_0 = 288,$$
$$M = 28.6,$$

and hence

$$R = \frac{917.5}{390 + 262 - 12.6} = 14350 \text{ miles},$$

where the three terms in the denominator are: the air, water vapor, and temperature gradient terms, respectively. It is evident that the existence and distribution of the small amount of water vapor present (1.37 per cent), adds very greatly to the effectiveness of the air itself as a refractive medium.

Acoustical Instruments *

By E. C. WENTE

Previous to the development of amplifiers most of the instruments used in acoustical research depended for their operation upon purely mechanical principles. This paper includes a brief survey of such of these instruments as are still of interest in connection with the investigation of technical or research problems in acoustics, but it deals primarily with the more recent electrical devices used in the study of air-borne sound waves.

The limitations and fields of application of various electrical instruments, including microphones, particularly adapted to definite types of acoustic measurements, are discussed.

MEASUREMENTS in acoustics may be said to date from the fifth century B.C., when Pythagoras observed that the lengths of strings giving the fifth, the fourth and the octave had the ratios 6 : 4 : 3, but no further really significant quantitative acoustic measurements were reported until the 17th centry when the frequencies of vibration of the notes in the musical scale were determined by Mersenne.[1] The first systematic treatise on experimental acoustics was published by Chladni [2] whose work on the vibration of plates and diaphragms is well known. With respect to the development of present day acoustical instruments the most outstanding contribution of the last century was the application of diaphragms for receiving sound waves by Scott and Koenig. Such diaphragms not only are used in most of these instruments, but also form an important element in two notable inventions of the last century, the telephone and the phonograph.

One of the chief functions of an acoustic diaphragm is to translate the extremely small pressures of sound waves into comparatively large corresponding forces, but a diaphragm cannot deliver more power to a system than it absorbs from the sound field. Telephony over comparatively long distances was made possible by the invention of the carbon microphone, an instrument which is capable of translating the small powers of acoustic diaphragms into relatively much larger electrical powers. This microphone, while of great commercial utility, was, for a number of reasons, unsuited for most quantitative acoustic measurements. Practically all shackles were removed from

* Presented before Acoustical Society of America, December, 1935. Published in *Jour. Acous. Soc. Amer.*, July, 1935.

[1] *Harmonie Universelle* (1636).

[2] *Die Akustik* (1802).

the designer of acoustical instruments some twenty years ago through the invention of the vacuum tube telephone amplifier. Previously his chief concern lay in making a device sufficiently sensitive to give a measurable response; now sensitivity became of secondary importance, and attention could be focused on the design of instruments which should be capable of performing their function without distortion.

As a result of this invention, instruments depending upon an amplifier for their utility have come so to dominate the field of acoustic measurements that we might easily be led to disregard all others. A number of such other instruments have, however, in recent years been brought to a high state of development which are peculiarly suited either for the calibration of other devices, or for the study of certain special problems. This paper attempts to give a brief critical survey of the various types of acoustical instruments which at the present time are finding applications in technical fields and in acoustical research studies.

General Principles

The Rayleigh Disc

That under certain conditions a torque is exerted on a thin disc suspended by a fine fibre in a stream of air was first observed by the late Lord Rayleigh,[3] who recognized in this phenomenon a means for measuring the intensity of sound.

The following quantitative relationship between the torque and the stream velocity was derived by W. Koenig:[4]

$$\tau = \frac{4}{3}\rho_0 r^3 u^2 \sin 2\theta,$$

where ρ_0 is the mean density of the medium, r the radius of the disc, u the stream velocity, and θ the angle between the undisturbed stream and the normal to the disc. The assumptions underlying the derivation of this formula are that the fluid is incompressible, that the disc is an infinitely thin ellipsoid and that there are no forces due to viscosity or to discontinuities of flow at the edges of the disc; i.e., the velocities are derivable from a potential. In view of these assumptions how far may we rely on the above formula in applying it to a suspended plane flat disc as commonly used for measuring the particle velocity of a sound wave, where none of these assumptions are strictly fulfilled? In most acoustical problems it is perfectly safe to assume a potential field as the forces due to viscosity and eddies are of the

[3] *Phil. Mag.* **14**, 186 (1882).
[4] *Wied. Ann.* **43**, 43 (1891).

second order. With respect to the torque on the Rayleigh disc it is not so obvious that such forces may be neglected as the potential torque is itself of the second order. The effects of discontinuities in the flow at the edges and of viscosity have not been determined theoretically. To what extent these are negligible can be found out only by experiment. Strictly speaking we therefore cannot regard the Rayleigh disc as an absolute means for determining sound intensities, as is often implied. A number of experiments have been carried out to determine the accuracy of the Koenig formula. Koenig himself, in a subsequent paper,[5] made an estimate of the effect of the discontinuous flow at the edges and reported measurements of the torque exerted on a flat disc when placed in a steady stream of air of known velocity. As a result of these studies he came to the conclusion that the application of his simplified formula to the Rayleigh disc did not provide a reliable and simple method for measuring the absolute value of sound intensity. He felt that the effect of viscosity would probably also have to be taken into account. However, Koenig's experiments were made under difficult conditions and it is possible that his measurements were affected by eddies in the air stream.

Greater confidence in the accuracy of Koenig's formula is derived from the experiments of Zernow.[6] Zernow experimented with both thin true ellipsoids and flat discs; these were placed in a box attached to one prong of a tuning fork, the motion of which was observed microscopically. The tuning fork was driven at a frequency of 92 c.p.s. and the relation between the amplitude of motion and the resulting deflection of the disc was determined. The values so found for the ellipsoids agreed remarkably closely with those computed by the formula. For the discs the agreement was within about 10 per cent. On the basis of the values so found Zernow proposed an empirical correction factor which reduces to unity for infinitely thin discs. Barnes and West,[7] using thinner discs, made measurements similar to those of Zernow. They found almost perfect agreement between the experimental and the theoretical values. They were able to show also, by measurements made at audio frequencies with discs of different diameters, that the torque varied as the cube of the diameter, provided first, that the diameters did not exceed 1/5 wave-length, and second, that the discs were sufficiently rigid to be free from resonant vibrations at the measuring frequency. Mallet and Dutton [8] found that the torque was proportional to the square of the velocity up to

[5] *Wied. Ann.* **50**, 639 (1893).
[6] *Ann. d. Physik* **26**, 79 (1908).
[7] *Jour. I.E.E.* **65**, 871 (1927).
[8] *Jour. I.E.E.* **63**, 502 (1925).

velocities of 5 cm. per second. We might, however, expect that a torque resulting from discontinuous stream flow at the edges would be governed by a similar law.

No direct tests have been reported on the accuracy of the coefficient in Koenig's formula above 92 c.p.s., at which Zernow's measurements were made. However, sound intensities determined by the Rayleigh disc through the application of Koenig's formula have been found to be in good agreement with those determined with a microphone calibrated by other independent means. All the tests of the formula have been made with plane or spherical sound waves of moderate intensity. This fact should be borne in mind in order to guard against the use of the device under conditions where the formula may not be applicable. Such may be the case, for instance, where the sound intensity is very high. Measurements in non-uniform sound fields recently made by Kotowski [9] showed quite anomalous effects; in some cases the deflection was even in a direction opposite to that expected.

One great disadvantage of the Rayleigh disc method of measuring sound intensity, as ordinarily applied, is the fact that the disc will deflect under the action of a steady air stream. As the stream velocities in a sound wave are in any case quite small, circulating air currents may easily produce comparable deflections unless the instrument is well shielded therefrom. Under carefully controlled conditions measurements can be accurately made at sound intensities corresponding to pressures as low as one bar.

The effect of circulating air currents is greatly reduced in the method of measurement with the Rayleigh disc adopted by Sivian.[10] In this method the intensity of the sound to be measured is modulated at the source at a frequency of about 0.4 cycle per second. The disc with its suspension is proportioned so that its natural frequency is equal to this modulating frequency. The disc will then oscillate under the action of the modulated sound wave at an amplitude proportional to the square of the velocity. As circulating air currents generally have components lying below the modulating frequency they will have but little effect on the amplitude of the oscillations of the disc.

Determination of Intensity from Static Pressure Measurements

Another purely mechanical means for measuring sound intensity in absolute terms is based upon the fact that when radiant energy falls on a reflecting surface a static pressure is exerted on this surface, which in the case of sound is equal to [11] $((\gamma + 1)/2)I/c$, where I is the

[9] *E.N.T.* **9**, 404 (1932).
[10] *Phil. Mag.* **5**, 615 (1928).
[11] Lord Rayleigh, *Phil. Mag.* **10**, 365 (1905).

intensity and c is the velocity of sound. A disc which just clears the opening in a plane baffle wall is attached to one arm of a torsion balance. From the deflection of the balance when sound falls at perpendicular incidence on the disc the radiation pressure and hence the intensity of the sound may be determined. This method has been successfully used in experiments with supersonic waves. At these high frequencies the diameter of the disc may be made a large fraction of a wavelength and the baffle may be omitted. At audio frequencies the necessity of using a baffle is a distinct handicap to this method.

Since the relation between pressure and condensation in air is not strictly linear a sound wave will, under certain circumstances, produce a change in static pressure. For a plane wave this has been shown by Thuras, Jenkins and O'Neil [12] to be equal to $-((\gamma + 1)/4) \times I/c$, where I is the sound intensity. Eichenwald [13] has suggested that a measurement of this pressure should provide a means for determining the absolute value of the sound intensity. Such increments in static pressure can, however, exist only when equalization by air flow to regions of normal pressure is precluded, a condition not easily established in practice.

Acoustic Valve

An extremely simple device for measuring sound intensities was devised by Kundt.[14] One end of a tube, which is placed in the sound field, is terminated by a valve which is so delicate that it will close during the negative and open during the positive half of the pressure cycle of the sound wave. The other end of the tube is terminated by a manometer. With perfect operation of the valve the sound wave will force air into the tube until the pressure indicated by the manometer is approximately equal to the maximum pressure in the sound wave. Recently Eisenhour and Tyzzer [15] have developed a sound meter operating on this principle. It is provided with an ingenious type of sensitive manometer with which the pressures are indicated on a dial. It has a fairly uniform sensitivity up to 2,000 c.p.s. The construction of the valve used in this meter is not disclosed in the literature. However, Ribbentrop [16] recently has described a similar sound meter in which the valve consists of the wing of a house-fly placed over an opening. It is stated that the instrument is capable of giving reliable measurements for sound pressures above 70 bars.

[12] *Jour. Acous. Soc. Amer.*, January, 1935.
[13] *Rend. Sem. Mat. e Fisico d. Milano*, Vol. 6 (1932).
[14] *Ann. d. Physik* **134**, 568 (1868).
[15] *Jour. Franklin Inst.* **208**, 397 (1929).
[16] *Zs. f. Tech. Phys.* **13**, 396 (1932).

Measurement of Periodic Changes in Density

As the optical index of refraction of an elastic medium depends upon the density, it is possible to measure sound by letting one of the paths of the light beams of an interferometer pass through the sound field while the other is shielded therefrom. The interference fringes of the interferometer will be displaced periodically in synchronism with the periodic variations in density of the wave. This method was first used by Boltzmann and Toepler [17] who in this manner observed the rather large variations in density within a sounding organ pipe. This method has the advantage that the measurements are independent of frequency but it is not very sensitive and at best is rather cumbersome. An interesting modification [18] of this method has recently been applied in measurements of high-frequency sound waves in liquids. At these high frequencies the wave-lengths are so small that the spatially periodic variations of the density of the medium can act as a diffraction grating for light waves. This phenomenon has provided a neat means of picturing the propagation of high-frequency sound waves in liquids.[79]

Instruments Employing Diaphragms and Optical Magnification

In the phonautograph of Scott (1857) a circular diaphragm is actuated by sound waves and the motion is recorded on a moving strip of smoked paper by a stylus attached to the center of the diaphragm. The recorded amplitudes are no greater than the actual amplitudes of motion of the diaphragm which, except at the resonance frequency or for very intense sounds, are so small that they cannot be accurately determined from the record. Small motions can be observed and recorded if the stylus is replaced by an optical lever. This arrangement in various forms has been used in the past by a number of investigators. It reached its highest state of development in the well-known phonodeik of D. C. Miller.[20] In this instrument a horn is used for increasing the sound pressure acting on the diaphragm, the motion of which is magnified in some forms of the instrument by as much as 40,000 times. By refinements in mechanical design and construction a remarkably uniform sensitivity was achieved.

Microphones

The instruments discussed so far operate without the benefit of electric-current amplifiers. The important role that these amplifiers

[17] *Pogg. Ann.* **141**, 321 (1870).

[18] Debye and Sears, *Proc. Nat. Acad. Sci.* **18**, 409 (1932). Lucas and Biquard, *Jour. de Physique et le Radium* **3**, 464 (1932).

[19] R. Baer and E. Meyer, *Phys. Zeits.* **34**, 393 (1935).

[20] *Science of Musical Sounds*, The Macmillan Company (1922).

have played in recent developments of acoustic instruments has already been indicated. To apply such amplifying means we must first of all have a device to convert sound power into electrical power. By far the most important instrument of this class is the microphone, which is a device that translates sound into corresponding electrical currents. When a medium is traversed by a sound wave it undergoes periodic variations in pressure, density, temperature and particle velocity. A device which translates any one of these variations into corresponding electrical currents may be classified as a microphone.

The great utility of the carbon microphone rests upon the fact that it in itself functions as an amplifier, i.e., the electrical power generated is greater than that absorbed from the actuating sound wave. The carbon microphone, however, has not been widely used for acoustic measurements, lacking the requisite stability and constancy. After amplifiers became available high sensitivity was no longer so important. It became possible to develop microphones in which high sensitivity was a subordinate property but which were stable and constant and relatively free from distortion.

The sensitivity of a microphone as a function of the frequency can usually not be easily determined from its physical constants. It must, therefore, be calibrated to be useful for general acoustic measurements. Such calibrations are commonly made in terms either of the voltage generated per unit of pressure acting on the instrument, or of the voltage per unit of the pressure obtaining in a plane progressive sound wave before the microphone is placed in the sound field. The former is referred to as a pressure and the latter as a free field calibration. Very complete discussions of the various methods of effecting such calibrations have been given by L. J. Sivian [21] and by S. Ballantine.[22] Unless the dimensions are small compared with the wave-length the microphone will diffract the sound waves and the pressure on the diaphragm will not be the same as that of the undisturbed sound field; for example, at normal incidence and at frequencies for which the wave-length is small compared with the diameter of the microphone the pressure will be doubled. The diffraction effect exhibits itself, particularly in a variation in the response-frequency characteristic with angle of incidence of the sound wave, generally in not an easily predetermined manner. If the form of the instrument is that of a sphere it is possible to determine this variation with angle of incidence theoretically. Ballantine [23] and also Oliver [24] have, there-

[21] *Bell Sys. Tech. Jour.* **X**, 96 (1931).
[22] *Jour. Acous. Soc. Amer.* **3**, 329 (1932).
[23] *Phys. Rev.* **32**, 988 (1928).
[24] *Jour. Sci. Inst.* **7**, 113 (1930).

fore, worked with instruments of this form. In any type of microphone diffraction effects can be entirely eliminated only by making the dimensions small compared with the wave-length.

The calibration of a microphone for a particular sound field may be carried out by measuring the undisturbed field with a device which is small compared with the wave-length and then noting the response of the instrument when placed in this field. This kind of calibration, when made in a nearly plane progressive wave, is referred to as a free field calibration. For the standard measuring instrument a Rayleigh disc is commonly used. This calibration is then applicable only for cases where we have substantially this type of sound field, i.e., when the microphone is at some distance from the source and all the sound is received by direct transmission. Where this condition is not fulfilled, the free field calibration is no true indication of the performance; for instance, when an instrument is used as a close talking microphone our experience indicates that in some cases at least an instrument having a flat characteristic, as obtained by a pressure calibration, delivers a voltage having frequency components of more nearly the same relative intensity as that in the voice when no microphone is near the mouth than does a microphone having a flat characteristic as given by a free field calibration. To eliminate diffraction effects a number of investigators have constructed microphones of small size, to some of which reference will be made in subsequent sections. Where it is necessary to make measurements with an extremely small instrument, such as in the exploration of the sound field within conduits and horns, the most satisfactory method of procedure is to use a small tube leading to a chamber closed over the diaphragm of a larger microphone.[25] The disadvantage of this arrangement is the fact that the loss in pressure through such tubes increases rapidly with frequency, so that at high frequencies it is necessary to work with high sound intensities or use uncomfortably high gain amplifiers. In working with single frequencies a great advantage in ease of measurement can be gained by the use of band-pass filters.

Pressure Microphones

Although microphones may conceivably be designed to translate directly the periodic variations of pressure, temperature, density, or particle velocity of a sound wave into corresponding electrical voltages, it is convenient to divide them into two classes: pressure microphones and velocity microphones, since the first three of the above characteristics of sound waves are proportional in any type of sound field.

[25] Sell, *Wiss. Ver. d. Siemens-Konz.* 2, 353 (1922).

Condenser Microphone

One of the first so-called high-quality microphones developed for use with amplifiers was of the condenser type. This is in principle one of the simplest of all microphones. It consists essentially only of two parallel insulated plates, one of them fixed and the other movable under the action of the alternating pressure of the sound wave. When these plates are connected in series with a resistance and a battery an alternating current will flow in this circuit in accordance with the variations in capacitance between the two plates. The resulting potential variations across the resistance are impressed on the grid of a vacuum tube.

A different method of using the condenser microphone has been described by Riegger.[26] The microphone is made a part of the capacitance element of a high-frequency electric oscillator. The frequency of the oscillations is thus modulated in accordance with the sound pressure acting on the diaphragm. If the modulated current is transmitted through a circuit, the transmission of which varies linearly with the frequency, in series with a linear rectifier, the output current of the rectifier will correspond to the sound pressure.

The condenser microphone as commonly used is of a size such that at the higher acoustic frequencies it will distort the sound field. The pressure and free field calibrations begin to diverge from each other at about 1000 c.p.s. To eliminate this distorting effect a number of investigators [27] have developed miniature condenser microphones for laboratory use. Generally such instruments have been designed at a sacrifice in sensitivity and uniformity of response. The small size microphone developed by Harrison and Flanders, however, has a remarkably flat response frequency characteristic and a sensitivity comparable with that of the larger instrument. Still smaller condenser microphones have been constructed but at a sacrifice in sensitivity.

At this point it may be of interest to give an example which illustrates the great advantage that the vacuum tube amplifier has given us in the design of sound measuring instruments. With an amplifier having a uniform amplification from 50 to 10,000 cycles, it is possible to measure, under favorable conditions, voltages as low as 1 microvolt. The amplitude of motion of the diaphragm of a common form of condenser transmitter delivering this voltage is about 10^{-11} cm., or about 1/1000 of an Angstrom. This illustrates the extremely small amount of motion that has to be imparted to the moving element of the

[26] *Wiss. Ver. Siemens-Konz.* 3, 2, 67–100 (1924).

[27] K. Hall, *Jour. Acous. Soc. Amer.* 4, 83 (1932). Harrison and Flanders, *Bell. Sys. Tech. Jour.* XI, 451 (1932).

measuring instrument. Not even with an optical interferometer could we hope to evaluate displacements so small.

Moving Coil Microphone

The condenser microphone has inherently a high electrical impedance, so high in fact that any attempt to connect the microphone to an amplifier by leads of appreciable length results in a loss of voltage. To avoid this loss an amplifier of at least one stage has generally been placed in close connection with the microphone. However, since the input impedance of a vacuum tube is also high, the microphone can be connected to it without the use of an impedance transformer, a distinct advantage at the time when transformers of good frequency characteristic were not available. During the last few years, through the development of new magnetic materials and advances in design, it has been possible to build transformers having a substantially uniform response over the whole acoustic frequency range. This development has made it possible to design microphones operating on electromagnetic principles, which have a good response-frequency characteristic and a greater sensitivity than the condenser microphone. They have an important advantage over the condenser microphone in that, because of their relatively low and constant impedance, they may be connected to the amplifier by a relatively long cable without appreciable loss. One such instrument [28] is shown diagrammatically in Fig. 1.

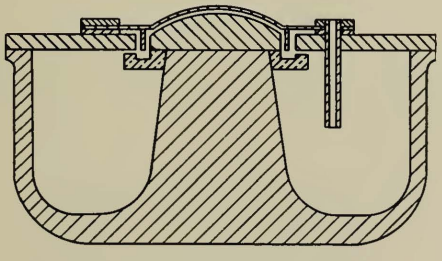


Fig. 1—Moving coil microphone.

The diaphragm has attached to it a coil which lies within a radial magnetic field. As the voltage generated by an axial motion of the coil is proportional to the velocity, if the same voltage is to be generated at all frequencies under a given sound pressure, the impedance of the moving element must be independent of frequency. This type of impedance characteristic over a wide frequency range is obtained by properly proportioned air chambers and resistances in back of the

[28] *Jour. Acous. Soc. Amer.* **3**, 44 (1931).

diaphragm. This microphone, when provided with a coil having an electrical resistance of 20 ohms, will generate 10^{-4} volts per bar of sound pressure. The smallest voltage that can be measured at the terminals of a resistance is limited by the voltage due to thermal agitation of the electrons,[29] which under normal conditions and for a frequency band of 15,000 c.p.s. is equal to 7×10^{-8} volts for a resistance of 20 ohms. Hence the smallest pressure that it is possible to measure with this microphone is about 7×10^{-4} bars. However, over a narrow band of frequencies, or at a single frequency, measurements may be made down to still lower pressures if the circuit is provided with a band-pass filter. The sensitivity of this instrument is higher than that of any other microphone of comparable frequency range at present available. In evaluating some of the other microphone principles we shall, therefore, use its sensitivity as a reference, without meaning to imply that sensitivity is the sole criterion of the merit of a microphone. There is also an upper limit to the sound intensities that may be measured with this instrument. This is governed by the maximum amplitude of excursion that the diaphragm can make without the generation of appreciable harmonics. The upper and lower limits at the various frequencies are shown by the curves in Fig. 2.

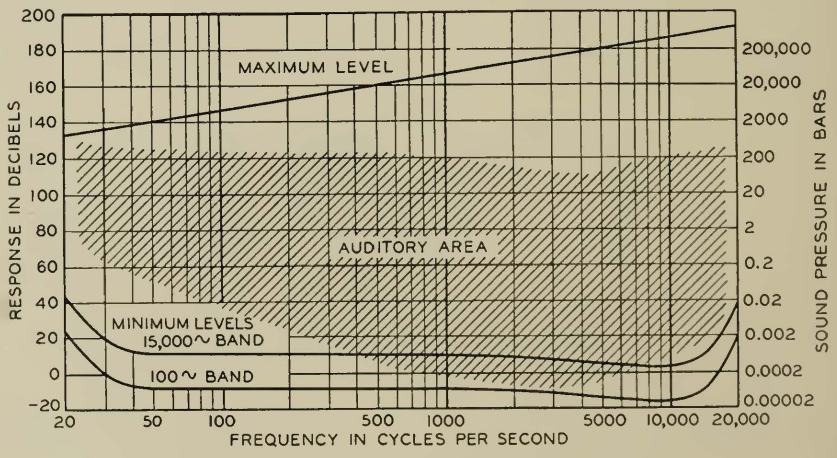


Fig. 2—Operating range of moving coil microphone.

The upper limit is taken as the pressure at which the higher harmonic components of the voltage are equal to 3 per cent of the fundamental. The lower limit represents the pressure at which the signal voltage is just equal to the voltage of thermal agitation. For comparison the

[29] J. B. Johnson, *Phys. Rev.* **32**, 97 (1928).

corresponding auditory range is indicated by the cross-sectioned area. It will be noted that when operated at its full frequency range even this relatively sensitive microphone is incapable of translating practically sound of intensities as low as the ear can hear.

This instrument, which in a similar form is used as a commercial microphone, is several inches in diameter and so is not without effect on the sound field. Where a certain amount of operating range and sensitivity may be sacrificed, as in many acoustic measurements, it is possible to construct this instrument in a much smaller form.

Capillary and Magnetostriction Microphones

Besides the electrostatic and electromagnetic methods of translating the mechanical pressures of a sound wave into corresponding electrical potentials, there are other electromechanical phenomena which may be applied for the purpose. Outstanding among these are the capillary electrometer, magnetostriction, and piezoelectric action.

When a potential is applied at the interface between an electrolyte and mercury the surface tension is changed. If the mercury is in a capillary tube the change in surface tension will result in a change in position of the mercury; conversely when a force tending to move the surface is applied, there will be a resulting change in potential across the interface. This phenomenon has been applied in the design of microphones. One form of construction of such an instrument is shown in Fig. 3, taken from a paper by Latour.[30] The instrument

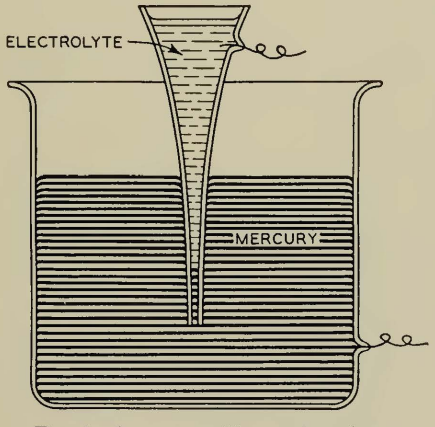


Fig. 3—Latour capillary microphone.

appears to have been used but little up to the present time and there seems to be very little in the literature regarding its performance.

[30] *Compt. Rend.* **186**, 223 (1928).

Magnetostriction also has found little application in microphones for air-borne waves at audio frequencies, although this principle has been applied with notable success in both the generation and detection of ultra-audio waves by G. W. Pierce and in the generation of audio frequency waves of high intensity in liquids.[31]

Piezoelectric Microphones

The application of piezoelectric action in the construction of acoustic microphones was first made by A. M. Nicolson,[32] who used Rochelle salt as the active material. Rochelle salt is unique in that its piezoelectric constant is about a thousand times as great as that of any other crystal. It has, however, several characteristics which would appear to render it unsuitable for use as a measuring microphone. It is mechanically fragile and its piezoelectric activity, under normal conditions, varies greatly with temperature, falling to a very low value for temperatures above 23° C. R. D. Schulwas-Sorokin [33] found, however, that by the application of a static stress the temperature coefficient could under certain conditions be greatly reduced and the activity extended to higher temperatures. C. B. Sawyer [34] found that if two thin slabs are cut and cemented together in such a way that one of the slabs will expand and the other contract when a potential is applied between the interface and the two outer surfaces, variations of activity with temperature are reduced to a low value. Presumably stresses are set up in the slabs by temperature variations which reduce the temperature coefficient of activity in accordance with the experiments of R. D. Schulwas-Sorokin. Sawyer has utilized these so-called bimorphic slabs in the construction of microphones. Single elements can be constructed of sufficiently small dimensions to avoid diffraction of the sound. In order to obtain microphones of greater practical efficiency a number of elements may be used in combination. If these elements are mounted symmetrically the translating efficiency will be the same in all directions about the axis of symmetry, as is the case for any microphone having an axis of symmetry. The amount of variation in respect to other directions depends upon the relation between the dimensions and the wave-length. According to the published data the sensitivity of a multiple element microphone of this type is about 25 db below that of a moving coil instrument.[35]

[31] Gaines, *Physics* 3, 209 (1932).
[32] *Trans. A. I. E. E.* 38, 1315 (1919).
[33] *Zs. f. Physik* 73, 9–10, 700 (1932).
[34] *Proc. I. R. E.* 19, 2020 (1931).
[35] A. L. Williams, *Jour. S. M. P. E.* 23, 196 (1934).

Of the more common piezoelectric crystals tourmaline possesses a characteristic which renders it peculiarly suitable for the absolute measurement of sound intensities at all audio frequencies, in that it may be so cut into slabs that a potential difference will be developed between its lateral surfaces when it is subjected to a purely hydrostatic pressure. Because of this characteristic Sir J. J. Thomson [36] suggested its use for measuring pressures in gun barrels. Such a slab of tourmaline, having dimensions small compared with the wavelength, except for its low sensitivity, is the ideal microphone. Tourmaline is mechanically strong and its activity is practically constant under all atmospheric conditions. Resonant frequencies in the slab lie far out of the range of audio frequencies so that the response at all frequencies is the same and is easily determined from static or low-frequency measurements. Unfortunately the sensitivity of such a device is low, some 70 db below that of a moving coil microphone. In spite of this low sensitivity it can be used for calibrating other microphones if sound waves of rather high intensities are used and if the electrical circuit is provided with a band-pass element transmitting only frequencies in the immediate neighborhood of the measuring frequency. A measuring system of this character, unlike the Rayleigh disc, is not subject to disturbances from circulating air currents.

Thermometric Microphones

As the pressure variations in a sound wave are accompanied by corresponding variations in temperature, corresponding electrical currents will be generated by a resistance thermometer or a thermocouple when placed in the sound field. The temperature variations are of the order of 0.0001° C. per bar acoustic pressure. The use of a resistance thermometer (for measuring these periodic temperature variations) was first investigated by Heindlhofer,[37] and more recently by Friese and Waetzmann,[38] who found that at a frequency of 1000 c.p.s. a wire 0.0004 cm. in diameter will undergo temperature variations equal to about 0.15 of the variations in the surrounding medium. To derive an alternating electric current from the periodic resistance variations that follow the temperature variations, a direct current must be passed through the wire. The heat generated by this current, unless it is kept down to an extremely small value, will set up convection currents around the wire and so greatly complicate the operation.

The thermocouple is entirely free from this objection, but is not readily constructed so as to have a heat capacity as small as the Wollas-

[36] *Engineering* **107**, 543 (1919).
[37] *Ann. d. Physik* **37**, 247 (1912); **45**, 259 (1914).
[38] *Zs. f. Physik* **29**, 110 (1925); **31**, 50 (1925); **34**, 131 (1925).

ton wire. Recently A. E. Johnson [39] has been able to make thermocouples with exceedingly small heat capacities with which measurements have been made up to 5,000 cycles and it is stated that they are usable up to several hundred thousand cycles. They are so small that they do not alter the sound field by diffraction and are free from resonance effects inherent in most instruments depending upon mechanical movement. As compared with other types, thermocouple microphones have a low sensitivity, at least 100 db below that of the moving coil microphone, according to the data given by Johnson.

Velocity Microphones

All the preceding types of microphones depend ultimately for their operation upon pressure variations in the sound wave. As the two primary characteristics of sound are pressure variations and alternating flow of the air particles, it is possible also to design microphones which generate voltages in accordance with the velocity of the air particles.

Hot Wire Microphone

One form of microphone of this character depends upon the change in resistance of a heated fine wire resulting from changes in temperature produced by the transverse flow of air. A microphone operating on this principle was first devised by Tucker [40] and used extensively during the war for locating enemy artillery. In order to increase the sensitivity and reduce distortion a steady stream of gas should be passed across the wire. An application of this principle to the construction of a microphone is shown in Fig. 4. Maximum response is

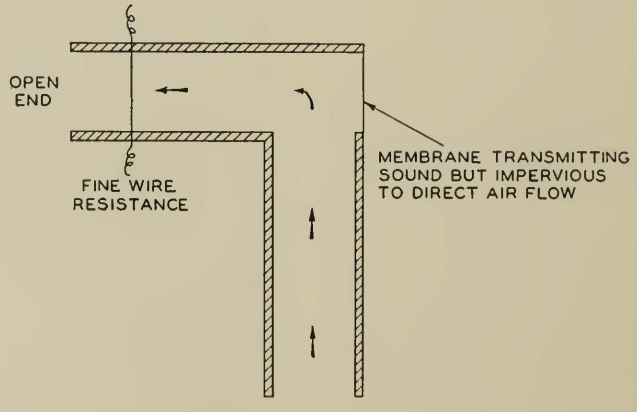


Fig. 4—Hot wire velocity microphone.

[39] *Phys. Rev.* **45**, 645 (1934).
[40] *Phil. Trans.* **221**, 389 (1921).

obtained when the direction of the sound wave coincides with the direction of the steady stream. At a given frequency the resistance variation is nearly proportional to the product of the steady stream velocity, the particle velocity of the sound wave and the cosine of the included angle. Since velocity in contrast with pressure is a vector quantity, a velocity microphone will respond selectively to sound coming from certain directions even at low frequencies. This characteristic is of considerable advantage in certain types of measurements, for it is often possible to so place and orient the instrument that its response is a minimum for an interfering or disturbing sound and a maximum for the sound to be measured. An illustration of such an application in sound measurement or pick-up is shown in Fig. 5

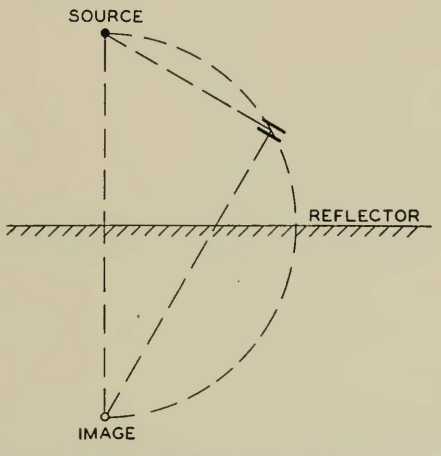


Fig. 5—Method of reducing effect of interfering waves with velocity microphone.

where the instrument is so placed that it will receive the sound directly from the source, but is insensitive to the sound reflected from the floor or neighboring wall. Other examples of acoustic measurements, where benefit is derived from the directional characteristics of the velocity microphone, are discussed in a recent paper by Wolff and Massa.[41]

There is one other important difference in the performance of velocity and pressure microphones. In a plane progressive wave particle velocity and pressure are strictly proportional at all frequencies. For a spherical sound wave, the radius of curvature of which is small compared with a wave-length, this is no longer true.

[41] *Jour. Acous. Soc. Amer.* IV, 217 (1933).

If we have a simple source of constant strength $A \cos kct$, the pressure at a distance r is given by

$$\frac{A\rho f}{2r} \sin k(ct - r),$$

where ρ is the density, c the velocity of sound, and k is equal to ω/c. The particle velocity is given by

$$\frac{Af}{2cr}\left[1 + \left(\frac{\lambda}{2\pi r}\right)^2\right]^{1/2} \sin [k(ct - r) + \psi],$$

where ψ is a function of λ and r. It will be seen that the pressure varies inversely with the distance for all frequencies, while the relationship between velocity and distance involves the wave-length, or frequency. In Fig. 6 are given some response-frequency characteristics

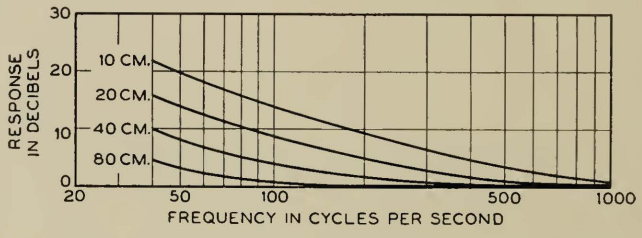


Fig. 6—Response of velocity microphone as a function of distance from source.

for several distances from the source of a velocity microphone having a uniform characteristic for plane waves. The change in characteristic as the instrument is brought close to the source is very marked. Some care is therefore required in interpreting the results of measurements made with this type of instrument. On the other hand, a pressure microphone, except in so far as diffraction may modify the sound field, will exhibit the same form of response-frequency characteristic at all distances from the source, so the wave form of the voltage generated by a pressure microphone will be the same for all positions in the free sound field of a simple source. This difference in characteristics of the two types of instruments is easily observed by comparing reproduced speech when the microphones are first placed near and then at some distance from the speaker's mouth.

Ribbon Microphone

A form of microphone, which has been extensively used in recent years, is the ribbon microphone. Essentially it consists of a very thin

strip of aluminum with circuit terminals at its two ends. This ribbon is placed in a magnetic field so that the lines of force lie in the plane of the ribbon and perpendicular to its long dimension, as shown in Fig. 7. Motion of the ribbon set up by sound waves will then generate a potential between its terminals. This type of microphone construction was first suggested by Reinganum.[42] It was developed into a practical form by Gerlach [43] and Schottky.[44] They apparently preferred to shield one side of the ribbon so that the instrument operated as a pressure microphone. H. F. Olson,[45] recognized the greater simplicity of the instrument in construction and in operation if both sides of the ribbon were freely exposed to the air. Constructed in this

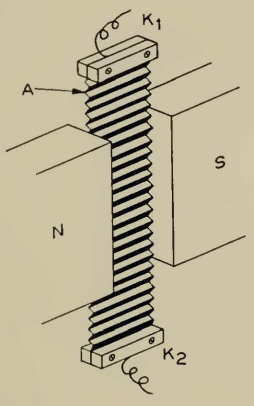


Fig. 7—Ribbon microphone.

way the instrument is virtually a velocity microphone at least at low frequencies. At the higher frequencies the ribbon with its surrounding structure almost completely shields the rear from sound reaching the front of the ribbon at perpendicular incidence. Under these conditions the instrument operates substantially as a pressure microphone, but even at these frequencies sound reaching the instrument from a direction parallel to the plane of the ribbon is without effect. Curves published by Olson on the directional characteristics of this microphone show that in a plane perpendicular to the axis of the ribbon the variation of response with direction follows approximately a cosine law, where the angle is measured from a line drawn normal to the plane of the ribbon. The relationship is more complicated over a plane passing through the axis of the ribbon.

[42] *Phys. Zs.* **11**, 460 (1910).
[43] *Phys. Zs.* **25**, 675 (1924); *Wiss. Ver. Siemens-Konz* **3**, 139 (1923).
[44] *Phys. Zs.* **25**, 672 (1924).
[45] *Jour. Acous. Soc. Amer.* **3**, 56 (1931).

A microphone that responds to the velocity of the air particles has equal sensitivity for sound waves traveling in opposite directions. Weinberger, Olson and Massa [46] have modified the ribbon microphone so that its response is not the same for the positive as for the negative direction of propagation of the sound wave. So modified the microphone has a greater response for sound waves coming towards one side of the ribbon than for those coming towards the other side. In a plane wave the magnitude of the pressure and the velocity are proportional. For waves traveling in a positive direction the two are in phase and for waves traveling in the negative direction they are in opposite phase. If, then, a pressure and a velocity microphone of equal sensitivity are connected in series the resultant voltage will be double for sound of normal incidence coming from one direction and equal to zero for sound coming from the opposite direction. Weinberger, Olson and Massa placed an appropriate acoustic impedance over a part of the ribbon so as to give this part the characteristics of a pressure microphone, while the other part of the ribbon was left free so as to function as a velocity microphone. The voltage at the ends of the ribbon is then proportional to the vector sum of the pressure and the velocity in the sound wave. In this way a pressure and velocity microphone combination is obtained in one instrument. It is insensitive to sound falling at perpendicular incidence on one side of the ribbon but not to sound propagated in the plane of the ribbon, as in the case of a velocity ribbon microphone.

Electrical Instruments of Particular Interest In Acoustical Studies

So far our discussion has been restricted mainly to microphones and instruments used in their calibration. There have, of course, in recent years been developed many other devices especially adapted for the quantitative study of particular acoustic problems, but a rather extensive discussion of the microphone has been given because it is an adjunct in almost all of these other instruments. In great part acoustic measurements are today made by first translating sound into a corresponding amplified electric current. The results of measurement or analysis of this current may then be referred back to the sound if the characteristics of the translating device are known. The type of analyzer or measuring instrument applied to the electrical circuit depends then altogether upon the kind of information that is desired. Strictly speaking we should classify these not as acoustical but as electrical instruments. In fact, every kind of electrical instru-

[46] *Jour. Acous. Soc. Amer.* **5**, 139 (1934).

ment may at times find an application in the study of acoustical problems. We must, therefore, necessarily restrict ourselves to a discussion of the kind of instruments which can give types of information of general acoustical interest.

The Oscillograph

If we wish to obtain a complete picture of the sound wave the microphone amplifier output is connected to an oscillograph, which is a device for translating a time pattern of the electric current into a corresponding space pattern. If an undistorted pattern is to be obtained, not only must the various harmonic components of the current and the recorded wave have the same relative amplitudes, but their phase relationships must be preserved. One form of instrument very closely satisfying these conditions up to about 10,000 cycles is a Curtis string oscillograph,[47] which is a modified form of the Einthoven galvanometer. The arrangement of this instrument is shown diagrammatically in Fig. 8. It records the wave form optically on photographic paper, which is automatically developed and fixed within a fraction of a minute after exposure.

In certain types of problems one of the various recording devices employed in the production of sound pictures may be used advantageously. These instruments have, in general, not been designed to be free from phase distortion but the records are in a form suitable for reproduction so they lend themselves particularly to the study of the subjective aspects of sound.

When the sound wave to be studied is steady, the wave form can be conveniently observed or photographed by means of a cathode ray oscillograph. These instruments are now to be had in convenient form. When used with an automatic sweep circuit, as suggested by Bedell and Reich,[48] the wave form of any steady state current is shown as a stationary pattern on a screen. These oscillographs are generally free from both frequency and phase distortion up to the highest audio frequencies.

Harmonic Analyzers for Steady Currents

If we wish to study the composition of a sound wave in terms of its harmonic components we may, of course, analyze the oscillographic records by means of any one of the well known methods of harmonic analysis, but this is at best a laborious process. Also, it is usually difficult to read an oscillogram with sufficient accuracy to determine the magnitude of any component that is much smaller than that of the

[47] *Bell Sys. Tech. Jour.* **XII**, p. 76.
[48] *Science* **63**, 619 (1926).

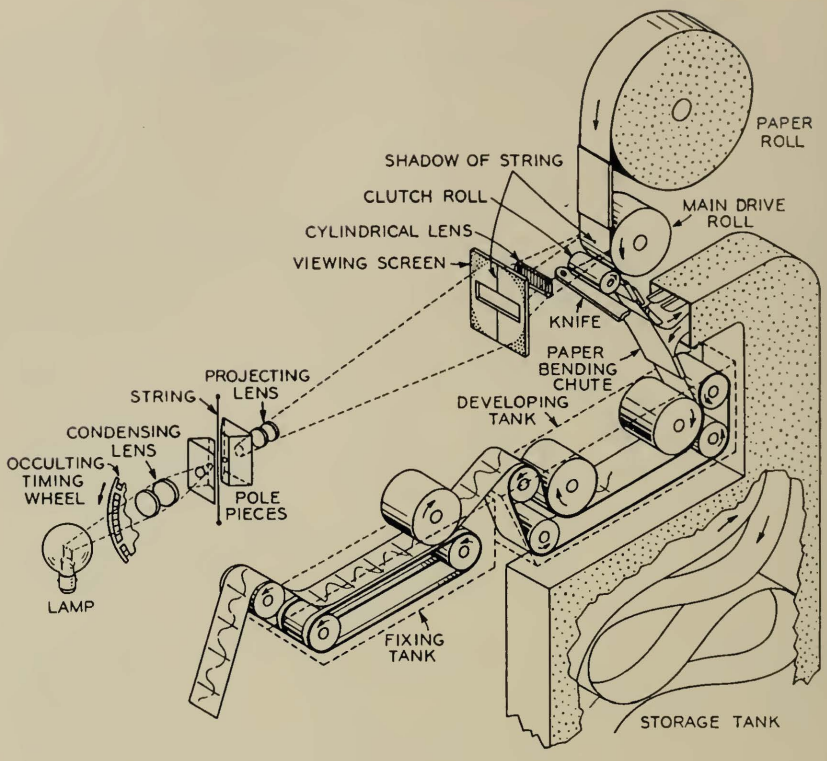


Fig. 8—Diagrammatic arrangement of Curtis oscillograph.

maximum component. When the conditions are such that a current can be steadily maintained, the analysis can be made much more conveniently and generally over a wider range of amplitudes by means of one of a number of recently developed types of current analyzers. These analyzers, although the mode of operation may differ in the various types that are available, have the common characteristic that they transmit only a narrow band of frequencies with the position of the band adjustable along the frequency scale. For analyzing a current wave the mid-frequency of the transmission band is shifted along the whole frequency range and the magnitude of the current in each frequency region is recorded or noted on a meter. This type of analyzer may also be used to get a statistical distribution of the power with respect to frequency when the sound is not periodic, as in certain kinds of noise.

High-Speed Analyzers

For the rather detailed study of sounds whose wave form varies with time, such as those of music or speech, an instrument is needed which

shall indicate the variations with time in the frequencies and amplitudes of all the harmonic components. Analyzers of the type just described indicate at a given instant the amplitude of only one component. In order to follow variations in all the components it would be necessary to sweep the frequency of transmission of the analyzing circuits rapidly back and forth over the frequency range of interest. The selective element of the analyzer, however, possesses a finite time constant; that is, when the selective circuit is set at a given transmission frequency a finite time is required for the transmitted current to reach a certain fraction of its steady state value and, similarly, a finite time is required for the current to decay to a certain fraction of this value when the transmission frequency is changed. This time constant depends to some extent upon the shape of the transmission versus frequency characteristic of the analyzing circuit but, in general, it bears an inverse relation to the selectivity. It is therefore not possible with an analyzer of this type, having a single variable selective element, to perform a rapid analysis without sacrificing resolution. This difficulty can, however, be circumvented if the analyzer is provided with a large number of fixed selective elements which are continuously operative. To build up the large number of required circuits from electrical elements would be extremely costly and would result in a bulky piece of apparatus. A compact form of analyzer having a large number of fixed selective mechanical elements has recently been described by C. N. Hickman.[49] This device has a series of tuned reeds, all driven electromagnetically at the same time by the current to be analyzed. The reeds are tuned so that their resonant frequencies differ progressively by equal pitch intervals. One hundred and twenty reeds are used to cover the range from 50 to 3,200 cycles. The deflection of each reed is made visible by the projection on a screen of a spot of light reflected from a mirror attached to the reed. The strength of each component in the current may thus be observed simultaneously on the screen or, if desired, the deflections may be recorded photographically.

A different and ingenious approach to this problem has been made by E. Meyer [50] in a recently described instrument. By methods well known in communications engineering the frequency of each component in the current to be analyzed is increased by an equal amount. A special high-frequency loud speaker translates the resultant currents into sound waves which are now all of very short wave-length. These waves are reflected from a concave grating made up of a large number

[49] *Jour. Acous. Soc. Amer.* **6**, 108 (1934).
[50] *Zeits. für Tech. Phys.* **12**, 630 (1934).

of equally spaced rods. The component waves are brought to a focus at different points along a focal surface analogous to the dispersion of light waves by an optical grating. A high-frequency microphone is moved back and forth along the focal plane through an amplitude large enough to cover one order of the spectra. This microphone is connected to an appropriate meter which records optically the intensity at various parts of the spectrum, which have a 1:1 correspondence with the component frequencies in the original current.

Measurement of Pitch

For acoustical studies, where it is of no particular importance to know the wave form but where interest lies in the variation of pitch with time, as in the study of the vibrato in musical tones, or in the inflections of the speaking voice, several types of instruments have been devised. Perhaps of these the most widely known is the tonoscope developed by C. E. Seashore [51] and his associates, which operates on the stroboscopic principle. This instrument has rows of uniformly spaced dots on a rotating cylinder, the number of dots increasing in successive rows. A neon light is made to flicker in synchronism with the fundamental of the tone under investigation. The particular row which under the light appears stationary gives the pitch of the tone at any instant. By the aid of a suitable camera the time variations of pitch may be recorded photographically, giving a so-called strobophotograph.

A frequency recorder operating on a different principle has been described by Hunt.[52] By a special circuit arrangement, employing gas-filled discharge tubes in combination with a spark recorder, the pitch of a tone can be recorded on paper. The scale is linear up to 8,000 cycles. This instrument is capable of following changes in pitch at a high rate of speed.

High-Speed Level Recorder

In some important types of sound measurements we are not interested in a detailed analysis of the sound wave but merely in the variation with time of the average level of the sound, as in the measurement of the rate of decay in a room or the flow of energy in speech, music, or noise. In some cases this average is preferably taken over long and in others over short time intervals. For long time averages, a thermocouple or rectifier and an ammeter may be used, but for short time averages an instrument is required which can follow changes in intensity at a higher rate of speed. Frequently also the range of inten-

[51] *Jour. Acous. Soc. Amer.* **2**, 77 (1930).
[52] *Rev. Sci. Inst.* **6**, 43 (1935).

sities over which we desire to make measurements of this character is very wide. Reverberation measurements are preferably made over a range of at least 60 db and the level range of orchestral music covers about 75 db. Several instruments designed for such purposes have been described recently.[53] In the instrument described by Wente, Bedell and Swartzel the level is recorded by a stylus on waxed paper. The recorder can be adjusted to give either a short or a long time average. At the higher speeds it is capable of following changes in intensity at the rate of 840 db per second and fluctuations in intensity of about 100 per second. The instrument may be adjusted so that the full scale covers a range of 30, 60 or 90 db.

Loudness Measurements

The preceding discussion was restricted to the purely objective or physical aspects of sound. In certain types of acoustical problems, as in the study of noise, we are, however, interested in subjective characteristics, but we do not yet have instruments which respond to an acoustic stimulus in the way the brain does through the ear. In fact

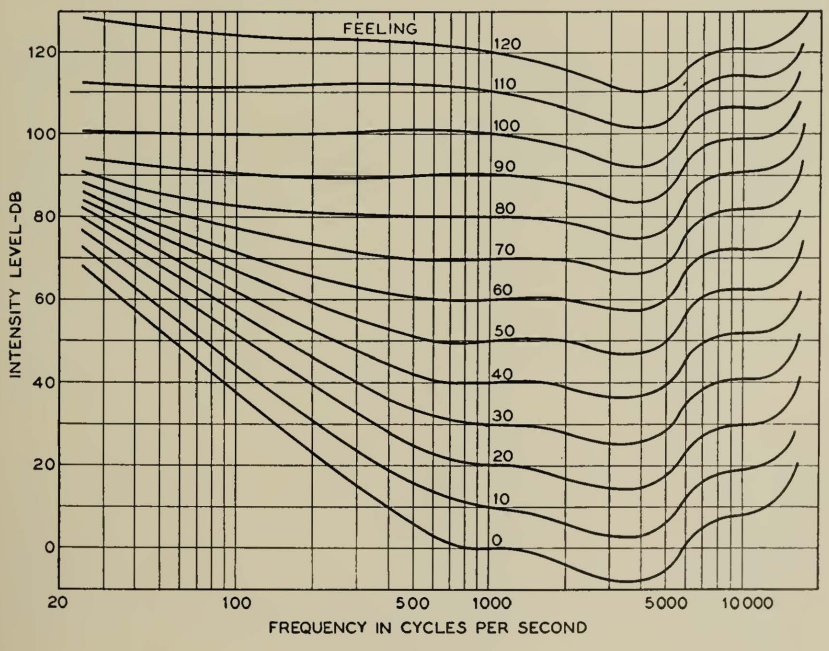


Fig. 9—Auditory chart.

[53] Hunt, *Jour. Acous. Soc. Amer.* **6**, 54 (1934). Wente, Bedell and Swartzel, *Jour. Acous. Soc. Amer.*, January, 1935.

we are not yet completely clear as to the relationship between sensation and stimulus, although Fletcher and Munson [54] have developed formulae whereby the loudness level of a steady sound can in most cases be computed from the intensity level of its components. For single pure tones the relationship between sensation and stimulus has been extensively explored with the results which are indicated in the auditory chart shown in Fig. 9, as given by Fletcher and Munson. The various curves give the intensity level of pure tones of equal loudness. This chart gives some idea of the complexity of the relationship between loudness and stimulus. The threshold of audibility of course varies widely with frequency and the relationship between sensation level and intensity level is not the same at the various frequencies and levels; for instance, at a loudness level of 40 db above threshold, a change of 5 db in the stimulus at 100 cycles produces the same change in sensation as a change of 10 db at 1,000 cycles. Fechner's law does not hold strictly over a wide range of intensities at any of the audible frequencies. The difficulty of devising an instrument which would have similar characteristics is apparent. "Sound level meters" have, however, been designed which give a reading which is approximately proportional to the subjective intensity of the sound. These meters are generally so designed that they have a frequency characteristic corresponding to the auditory curve at about the level of the noise being measured. They have proved themselves extremely useful although from our knowledge of hearing phenomena we might expect large variations in the actual loudness of sounds of different character, even if a noise meter of the above type should show them to be equal.

[54] *Jour. Acous. Soc. Am.* **5**, 82 (1933).

Thermionic Electron Emission *

By J. A. BECKER

Introduction

THERE have appeared in the *Reviews of Modern Physics* two excellent summaries on thermionic emission, one by Compton and Langmuir [1] and one by Dushman.[2] Compton and Langmuir, while dealing primarily with discharge in gases, also discussed many phases of thermionic emission. Dushman's article is a comprehensive review on thermionics. He faithfully reflects whatever viewpoints and experiments appear in the literature. Besides reviewing the work that has been performed since 1930, the present article will be an attempt to review in a critical manner some of the matters which in the preceding reviews were left undecided. On the other hand, no attempt will be made to give a complete presentation of all the views appearing in the literature. As to the close connection between thermionic and adsorption phenomena, this will be dealt with in an article now in preparation.

Recently there have been published two comprehensive books on thermionics. One is in English by A. L. Reimann.[3] The other is Vol. IV of Müller-Pouillets *Lehrbuch der Physik* [4] edited by A. Eucken, with contributions by A. Eucken, R. Suhrmann, L. Nordheim and others. The topics which are fully covered in these two books and in the book [5] by W. Schottky and H. Rothe, *Physik der Glühelektroden* will not be covered in detail in the present article. Since photoelectric phenomena are closely associated with thermionics, it is well to refer also to Linford's [6] review on the external photoelectric effect and the book by Hughes and DuBridge [7] on photoelectric phenomena.

Empirical and Theoretical Richardson Formulæ

One topic on which considerable confusion has existed goes to the very root of thermionic emission. It is the interpretation that is to be put on the slope and intercept of a Richardson line and how the slope and intercept are related to certain quantities in theoretical formulæ. Empirically it is found that the thermionic emission current density, i, is related to the temperature, T, by the Richardson formula

$$i = A_n T^n \exp(-b_n/T), \tag{1a}$$

* Published in *Reviews of Modern Physics*, April, 1935.

or its equivalent

$$\log_{10} i = \log_{10} A_n + n \log_{10} T - (b_n/2.3T). \tag{1b}$$

A_n and b_n are constants characteristic of the surface. Their value depends on the value assigned to n. From such experiments it is impossible to decide whether n should equal 0 or 4 or any value between these. There are good theoretical reasons, which are given below, why $n = 2$. In that case

$$i = AT^2 \exp(-b/T), \tag{2a}$$

or

$$\log i - 2 \log T = \log A - b/2.3T. \tag{2b}$$

If $\log i - 2 \log T$ is plotted *versus* $1/T$, a straight line is usually obtained. Call this line a Richardson line. Its slope is $-b/2.3$, and its y intercept is $\log A$. Since we shall have numerous occasions to refer to the slope and intercept of a Richardson line, we shall find it convenient to refer to them by their equivalents $-b/2.3$ and $\log A$, respectively. On those rare occasions when the Richardson plot yields a curved line, we can draw a tangent at any point on the curve. Equation (2) will then represent the equation for this tangent; $-b/2.3$ and $\log A$ will depend on the particular point at which the tangent is drawn, so that b and A will depend on T.

The Thermodynamic Equation

The slope and intercept of experimental Richardson plots are to be correlated with certain quantities in one or the other of two theoretical equations. The first of these * is based on the first and second law of thermodynamics and the assumption that the electron vapor acts like a perfect gas.† The equation is:

$$\log i_T = \log i_{T'} + \log [(1 - r)/(1 - r')] + \tfrac{1}{2} \log T' - \tfrac{1}{2} \log T + (1/2.3) \int_{T'}^{T} (L_p/RT^2) dT, \tag{3}$$

in which T' is any fixed temperature in the experimental temperature range; r and r' are the electron reflection coefficients at T and T', respectively; L_p is the heat of vaporization per g. mole of electrons at constant pressure; R is the gas constant per g. mole.

Thermodynamics cannot tell us how L_p varies with T and until we know this we cannot perform the integration indicated. By consider-

* For a recent critical derivation see Becker and Brattain.[8]

† This assumption is subsequently justified by experiment.

ing the mechanism by which the electrons evaporate from the metal, we can arrive at some conclusions regarding the temperature dependence of L_p. Since in the derivation of equation (3) it was assumed that the electron vapor acts like a perfect gas, it follows that when 1 g. mole of electrons is vaporized at constant pressure an amount of work RT must be done against the external pressure and an amount of heat $(3/2)RT$ must be provided to furnish the known mean kinetic energy of the vaporized electrons. It then becomes desirable to define a new quantity h by the equation,

$$h = (L_p/R) - (5/2)T. \tag{4}$$

Since h plays an important role in the final formula, it will be convenient to give it the name "heat function." * The product kh, where k is Boltzmann's constant, represents the average heat of vaporization per electron less $(5/2)kT$. Substituting equation (4) in equation (3),

$$\log i = \log i_{T'} + \log [(1 - r)/(1 - r')] - 2 \log T' + 2 \log T + (1/2.3) \int_{T'}^{T} (h/T^2) dT. \tag{5}$$

Thermodynamics alone cannot tell us how the heat function h varies with T and we cannot perform the indicated integration until this is known. However, we can deduce an important theorem even without performing the integration: *If the experimental value of log i − 2 log T is plotted versus $1/T$, the slope of the tangent at any value of T is $-h/2.3$.*†

Hence for those surfaces for which the Richardson lines are straight, h is independent of T in the experimental range. For these surfaces, equation (5) reduces to

$$\log i = \log i_{T'}/(1 - r')(T')^2 + h/2.3T' + \log (1 - r) + 2 \log T - h/2.3T = \log I(1 - r) + 2 \log T - h/2.3T, \tag{6}$$

where

$$\log H = \log i_{T'}/(1 - r')(T')^2 + h/2.3T'.$$

Log $I(1 - r)$ is the intercept of the Richardson line on the y axis.

An alternative derivation of the thermodynamic emission equation uses the absolute zero of temperature as the lower limit in the various integrals. In this way Bridgman derives the equation,‡

$$i = U\alpha(1 - r)T^2 \exp. [- L_0/kT + \varphi(T)], \tag{6a}$$

* This is of course not the heat function used in thermodynamics. The heat function defined here has the dimensions of temperature. It is often given in volts $V = kh/e$. Later h will also be used for Planck's constant but we believe no confusion will arise.

† For the proof see Becker and Brattain.[8] In the proof it is assumed that dr/dT is zero or very small. This assumption is justifiable.

‡ See Eq. IV, 33 on page 99 of his book named in References.[9]

in which U is a universal constant equal to $2\pi Gk^2me/h^3 = 120$ amps./cm.^{2} °K.^{2}; h (Planck's constant), m, e and k have the customary significance; G is the statistical weight which is equal to 2 for electrons; $\alpha = \exp(S_{0\rho}/k)$; $S_{0\rho}$ is the entropy per atom of a metal whose surface has a charge density ρ at $T = 0$; L_0 is the heat of vaporization per electron at constant pressure at $T = 0$;

$$\varphi(T) = \frac{1}{k}\int_0^T \frac{(C_{p\rho} - C_{pm})}{T}\,dT - \frac{1}{kT}\int_0^T (C_{p\rho} - C_{pm})dT;$$

$C_{p\rho}$ is the specific heat per atom at constant pressure when the metal surface has a charge density ρ while C_{pm} is the specific heat for the uncharged metal. In the derivation it is assumed that the entropy of the uncharged metal at $T = 0$ is zero in accordance with the third law of thermodynamics; it is also assumed that the electron vapor acts like a perfect gas. The value of U follows from the value of the entropy constant of a perfect gas deduced from quantum statistics. Up to the present time neither theory nor experiment has yielded numerical values for α or $\varphi(T)$. If, however, it is assumed that $\alpha = 1$, $\varphi(T) = 0$ and $r = 0$ then equation (6a) reduces to

$$i = UT^2 \exp.(-L_0/kT), \tag{6b}$$

which is the equation derived by Dushman [10] in 1923. It predicts that all Richardson lines should have the same intercept on the y axis, namely, log U. Since this prediction is not fulfilled by experiment it would appear that the assumptions made in obtaining equation (6b) are not valid. It may be well to point out also that adsorbed particles on the surface probably contribute additional terms to the expression for the specific heats and entropy at absolute zero. These have not been taken into account.

We are now in a position to show why the exponent of T in equation (1a) should be 2. To do this we consider h in equation (5) or L_p in equation (3). The heat of vaporization L_p is defined as the heat energy that must be added to the system in order to evaporate one g. mole of electrons at constant pressure. We have seen that $(5/2)RT$ ergs must be added to account for the specific heat of the vaporized electrons and work done against the external pressure. The remainder, Rh, which includes all other energies can be put equal to $\bar{P} - \bar{K} - T(d\bar{P}/dT)$ where $\bar{P}$ is the increase in potential energy of the electrons, and $\bar{K}$ is the mean kinetic energy which the electrons had in the metal. $\bar{P}$ includes work done against the image force or any other electrical forces. So little is known about the exact nature of $\bar{P}$

that there is little point in examining the temperature dependence of the quantity $\bar{P} - Td\bar{P}/dT$ more closely. On the other hand, the quantity $\bar{K}$ and its variation with temperature does depend on the particular assumption that is made with regard to the energy distribution of the free electrons in the metal. In particular $\bar{K}$ is very nearly independent of T if the electrons in the metal have kinetic energies given by the Fermi-Dirac function.* That this is the correct distribution function is quite well established by the numerous successes which this theory has had in explaining experimental facts in connection with metals.[11–16] For the Fermi-Dirac distribution $\bar{K}$ is practically a constant term in the expression for the heat function h. There is then no reason for changing the form of equation (6) which contains the term 2 log T. This is equivalent to an exponent of 2 in equation (1a).

The case was somewhat different before the advent of the quantum theory. The electrons in the metal were then assumed to act like a perfect gas. Hence the energy $\bar{K}$ was taken to be $(3/2)RT$. It was thus natural to subtract this from the $(5/2)RT$ for the electron vapor. In this way one is led to an expression for log i similar to equation (6), but instead of 2 log T there now appears $\frac{1}{2}$ log T. So that the exponent of T in equation (1a) was taken to be $\frac{1}{2}$.

It is well to note that on the basis of this thermodynamic argument, there is no good reason why the heat function should be independent of T and why the Richardson lines should be straight. Experiment shows, however, that for nearly all surfaces which are not close to their melting point, the heat function is independent of T to within experimental error. In the neighborhood of the melting point, the heat function varies with T. It should also be noted that thermodynamics does not predict that all Richardson lines should have a common intercept on the y axis. This prediction which is true only for special classes of surfaces has been made on the basis of a statistical theory which we will now discuss.

The Statistical Equations

a. Classical treatment. If we knew the velocity distribution and density of the electrons inside a metal at various temperatures and the difference in potential energy between an electron at rest inside and outside the metal, it would be a comparatively easy task to determine statistically how many electrons could escape from a square centimeter of surface in one second. It was at first assumed that the electrons inside the metal acted like a perfect gas; the velocity distribution is

* This function will be discussed later in this paper.

given by Maxwell's law

$$n(u, v, w)dudvdw = \frac{nm^{\frac{3}{2}}}{(2\pi kT)^{\frac{3}{2}}} \exp. \left[\frac{-m(u^2 + v^2 + w^2)}{2kT} \right] dudvdw, \quad (7)$$

where u, v, w are the velocities in the x, y, z directions respectively; $n(u, v, w)$ is the number of electrons per cm.^{3} having u values in the range (u, du), i.e., between u and $u + du$, v values in the range (v, dv), and w values in the range (w, dw); n is the total number of electrons per cm.^{3}; m is the electron mass; and k is Boltzmann's constant. The number of electrons having u components of velocities in the range (u, du) is obtained by integrating equation (7) with respect to v and w from $-\infty$ to $+\infty$.

$$n(u)du = n(m/2\pi kT)^{\frac{1}{2}} \exp. (-mu^2/2kT)du. \quad (8)$$

The total number of particles striking a cm.^{2} of surface per second is given by

$$N_t = \int_0^\infty un(u)du = n(kT/2\pi m)^{\frac{1}{2}}. \quad (9)$$

But only those electrons whose values of u exceeds $u_0 = (2p/m)^{\frac{1}{2}}$ will escape from the surface. The quantity p, called the work function, represents the potential energy of the electron outside the metal; it is the work that must be done to take an electron at rest in the metal and transport it across the surface to a distance at which the surface forces are negligible. The total number, N, of electrons which can escape from one cm.^{2} of surface in one second is then

$$\begin{aligned} N &= \int_{u_0}^\infty un(u)dn \\ &= \int_{u_0}^\infty n(m/2\pi kT)^{\frac{1}{2}} u \exp. (-mu^2/2kT)du \\ &= n(kT/2\pi m)^{\frac{1}{2}} \exp. (-p/kT). \end{aligned} \quad (10)$$

The emission current in amperes per cm.^{2} is

$$i = Ne = A'T^{\frac{1}{2}} \exp. (-p/kT) \quad (11a)$$

or

$$\log i - \tfrac{1}{2} \log T = \log A' - p/2.3kT, \quad (11b)$$

where

$$A' = ne(k/2\pi m)^{\frac{1}{2}}, \quad (12)$$

e = charge on the electrons in coulombs = 1.59×10^{-19}.

If p is independent of T and if $\log i - \frac{1}{2}\log T$ is plotted *versus* $1/T$, the slope $= -p/2.3k$ and the y intercept $= \log A'$ or $\log ne(k/2\pi m)^{\frac{1}{2}}$. For clean tungsten it is found by experiment that the current density can be represented by

$$i = 2.06 \times 10^7 T^{\frac{1}{2}} \exp.\ (-55{,}300/T).$$

From this it follows that $ne(k/2\pi m)^{\frac{1}{2}} = 2.06 \times 10^7$ and that $n = 8.4 \times 10^{20}$ electrons/cm.^{3} This is to be compared with 635×10^{20} atoms/cm.^{3} So that if we postulate one "free electron" for every 75 atoms, we can account for the observed thermionic emission classically.

Such a concentration of free electrons may be considered to be in quite good accord with the first of two possible deductions from experiments on specific heats. From these it follows that: (1) Either the number of free electrons must be small compared to the number of atoms and the mean kinetic energy per electron is $(3/2)kT$; or else, (2) the number of free electrons is of the order of the number of atoms but the kinetic energy increase per degree rise in temperature is much smaller for electrons than it is for atoms. The correlation of experiment and classical theory in the case of the optical properties, electrical conductivity, thermoelectricity, Thomson and Peltier effects lead to certain inconsistencies. These disappear when theories based on the Fermi-Dirac distribution are used for these effects and it is postulated that the number of free electrons in metals is of the same order as the number of atoms. The classical theory for Richardson's equation thus leads to values of n which are incompatible with values deduced from these effects. The newer theory has also made progress in explaining ferromagnetism. It is thus a better basis for a statistical theory of electron emission. Such a theory was developed by Sommerfeld [11] and Nordheim.[16]

b. Quantum-mechanical treatment. The Fermi-Dirac theory gives the velocity distribution as

$$u(u, v, w)dudvdw = \frac{Gm^3}{h^3} \times \frac{1}{M^{-1} \exp.\ [m(u^2 + v^2 + w^2)/2kT] + 1} dudvdw, \quad (13)$$

G is the statistical weight; for electrons its value is 2. h is Planck's constant. The quantity M is so adjusted that the integral of $n(u, v, w)$ gives the total number of electrons/cm.^{3} This integration is so difficult that no relatively simple and exact expression for M can be found.

However, in two limiting cases good approximations have been obtained.

In the first case M is so small a quantity that

$$M^{-1} \exp. \left[m(u^2 + v^2 + w^2)/2kT\right] \gg 1.$$

It then follows that

$$M = nh^3m^{\frac{3}{2}}/G(2\pi kT)^{\frac{3}{2}} \tag{14}$$

and that equation (13) is the same as equation (7) for the classical treatment.

In the second case M is a large quantity and the 1 in the denominator of equation (13) cannot be neglected. Sommerfeld [11] has shown that in this case

$$M = \exp. (K/kT), \tag{15}$$

where

$$K = \frac{h^2(3n/4\pi G)^{\frac{2}{3}}}{2m}\left[1 - \frac{(2\pi mkT)^2}{12h^4}\left(\frac{3n}{4\pi G}\right)^{-4/3} + \cdots\right]. \tag{16}$$

The second term in the brackets is usually a very small numerical quantity and can nearly always be neglected.

If we assume that n, the number of electrons/cm.^{3}, is equal to the number of atoms/cm.^{3} in a metal or a small factor times this number, we can compute M for case 1 by equation (14) or for case 2 by equation (15). In either case M turns out to be a large quantity. Hence for metals the second case is applicable while the first case is not. Hence

$$n(u, v, w)dudvdw = \frac{Gm^3}{h^3} \times \frac{1}{\exp.\left[\dfrac{\frac{1}{2}m(u^2 + v^2 + w^2) - K}{kT}\right] + 1} dudvdw, \tag{17}$$

since

$$M^{-1} = \exp. (-K/kT).$$

Integrating this from $-\infty$ to $+\infty$ with respect to v and w Nordheim [14] has shown that the number of electrons per cm.^{3} having velocities in the range (u, du), i.e., between u and $u + du$, is

$$n(u)du = \frac{2\pi Gm^2kT}{h^3} \ln\left[1 + \exp.\left(\frac{K - \frac{1}{2}mu^2}{kT}\right)\right] du. \tag{18}$$

The number of electrons striking a surface normal to the u direction per cm.^{2} per sec. and having velocities in the range (u, du) is given by

$N(u)du = un(u)du$; hence

$$N(u)du = \frac{2\pi Gm^2kT}{h^3} u \ln\left[1 + \exp.\left(\frac{K - \frac{1}{2}mu^2}{kT}\right)\right] du. \quad (19)$$

Now only those having velocities greater than u_c will be able to cross the surface and escape where u_c is given by $\frac{1}{2}mu_c^2 = P_m$. P_m is the difference in potential energy between an electron at rest inside and outside the metal. Now P_m is about 3/2 times as large as K and therefore $\frac{1}{2}mu_c^2 > 1.5K$. Also for the values of T encountered in thermionic experiments kT is small compared with $(\frac{1}{2}mu_c^2 - K)$. Therefore, for values of $u > u_c$, exp. $[(K - \frac{1}{2}mu^2)/kT]$ is a very small quantity and

$$\ln\left[1 + \exp.\left((K - \tfrac{1}{2}mu^2)/kT\right)\right] = \exp.\left[(K - \tfrac{1}{2}mu^2)/kT\right]$$

to a good approximation. This follows, since

$$\ln(1 + \Delta) = (\Delta - \tfrac{1}{2}\Delta^2 + \tfrac{1}{3}\Delta^3 - \tfrac{1}{4}\Delta^4 + \cdots)$$

provided $\Delta^2 < 1$. Hence, for $u > u_c$ and the temperatures encountered in thermionic emission

$$N(u)du = \frac{2\pi Gm^2kT}{h^3} u \exp.\left(\frac{K - \frac{1}{2}mu^2}{kT}\right) du. \quad (20)$$

The number that cross the surface per cm.^{2} per second is given by

$$N = \int_{u_c}^{\infty} N(u)du = \frac{2\pi Gm^2kT}{h^3} \int_{u_c}^{\infty} u \exp.\left(\frac{K - \frac{1}{2}mu^2}{kT}\right) du$$

$$= \frac{2\pi Gmk^2T^2}{h^3} \exp.\left(\frac{K - \frac{1}{2}mu_c^2}{kT}\right)$$

$$= \frac{2\pi Gmk^2}{h^3} T^2 \exp.\left(-\frac{P_m - K}{kT}\right). \quad (21)$$

Finally

$$i = Ne = (2\pi Gmek^2/h^3)T^2 \exp.\left[-(P_m - K)/kT\right]$$

$$= UT^2 \exp.(-W/kT) = UT^2 \exp.(-w/T)$$

$$= UT^2 \exp.(-\varphi e/kT). \quad (22)$$

where

$$U = 2\pi Gmek^2/h^3 \quad (23)$$

and

$$P_m - K = W = kw = \varphi e. \quad (24)$$

If i is expressed in amperes per cm.^{2}, the value of U is 120 amperes/cm.^{2} °K.^{2} The quantities W, w, or φ are called the work function; the difference between them is merely one of units. φ is expressed in volts, w in degrees Kelvin, and P_m, K and W in ergs. P_m is called the outer work function and K, the inner work function. Oftentimes it is convenient to refer to P_m, K and W as if they were expressed in volts.

c. Treatment in terms of energies. For many purposes it is convenient to have expressions for the distribution in energies instead of in velocities. We can then express these energies in equivalent volts and obtain numerical values which are more familiar. Let

$$E = (m/2)(u^2 + v^2 + w^2); \qquad E_n = (m/2)u^2,$$

" the normal component of the energy "; $V_n = E_n/e$; $n(E)dE =$ the number of electrons per cm.^{3} having energies in the range (E, dE); similarly for $n(E_n)dE_n$; $N(V_n)dV_n$ is the number striking a cm.^{2} of surface per second having normal component of energies in the range (V_n, dV_n).

Then

$$n(E)dE = \frac{2\pi G}{h^3}(2m)^{\frac{3}{2}} \frac{E^{\frac{1}{2}}}{1 + \exp.\,[(E - K)/kT]} dE, \tag{25}$$

$$n(E_n)dE_n = \frac{2^{\frac{1}{2}}\pi G m^{\frac{3}{2}} kT}{h^3} \frac{1}{E_n^{\frac{1}{2}}} \ln\left[1 + \exp.\frac{(K - E_n)}{(kT)}\right] dE_n, \tag{26}$$

$$N(E_n)dE_n = \frac{2\pi G m k T}{h^3} \ln\left[1 + \exp.\left(\frac{K - E_n}{kT}\right)\right] dE_n, \tag{27}$$

$$N(V_n)dV_n = \frac{2\pi G e m k T}{h^3} \ln\left[1 + \exp.\left(\frac{K - V_n e}{kT}\right)\right] dV_n. \tag{28}$$

Equations (25) and (26) are readily derived from equation (18); while equations (27) and (28) follow from equation (19). It is also instructive to compare equation (28) with the corresponding equation which is based on classical statistics, namely,

$$N(V_n)dV_n = n(e^2/2\pi mkT)^{\frac{1}{2}} \exp.\,(-V_n e/kT)dV_n. \tag{29}$$

This is readily derived from equation (8).

d. Comparison between classical and quantum-mechanical treatment. Comparison between equations (28) and (29) is best brought out by a graph such as Fig. 1 which shows log $N(V_n)$ *versus* V_n for the two cases. It is to be remembered that $N(V_n)dV_n$ is the number of electrons in the metal which strike 1 cm.^{2} of surface per second whose

energy components normal to the surface are in the range (V_n, dV_n) volts. It has been customary to plot $N(V_n)$ *versus* V_n for equation (28). At $T = 0$, $N(V_n)$ decreases linearly with V_n from a value of

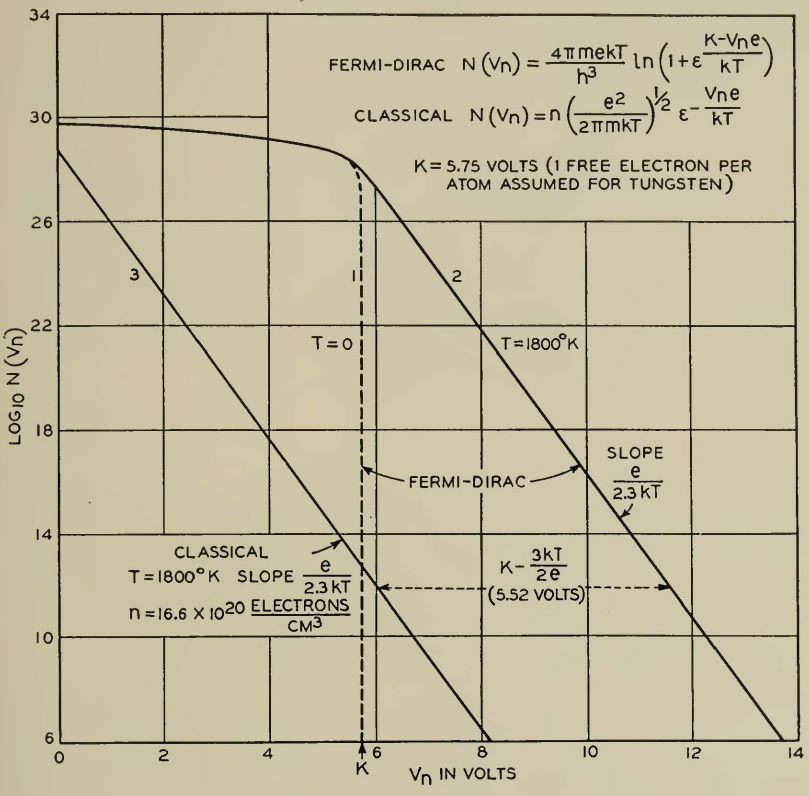


Fig. 1—Classical and Fermi-Dirac distributions.

$2\pi Gemk/h^3$ when $V_n = 0$, to zero when $V_n = K/e$; for $V_n > K/e$, $N(V_n) = 0$. For $T > 0$, the function is much the same except in the neighborhood of $V_n = K/e$ and for $V_n > K/e$; the curve is here everywhere higher than the curve for $T = 0$ and decreases exponentially. Since only those electrons can escape for which $V_n \geq P_m > (3/2)K$, we are primarily interested in the exponential portion of the curve. It is therefore more advantageous to plot log $N(V_n)$ rather than $N(V_n)$.

In Fig. 1 curves 1 and 2 are for equation (28) at $T = 0$ and $T = 1800°$ K., respectively; while curve 3 is for the classical case or equation (29). For curves 1 and 2, the value of K/e has been taken as 5.75 volts which is the value appropriate for tungsten assuming one

free electron per atom. For curve 3 the value of n has been so chosen that this curve is shifted with respect to curve 2 by $K/e - 3kT/2e$ or 5.52 volts. The value of n which does this is 16.6×10^{20}/cm.^{3} To account for the observed emission from tungsten we have previously deduced a value $\frac{1}{2}$ as great or 8.4×10^{20}/cm.^{3} The factor of 2 is due to the fact that the intercept of the observed Richardson plot for tungsten corresponds to 60 amp./cm.^{2} °K^2 while the theoretical intercept corresponds to 120 amp./cm.^{2} °K.^{2}

At first sight it might appear that the shift between curves 2 and 3 should be K/e rather than $K/e - 3kT/2e$. The additional term is accounted for by comparing the classical or $T^{\frac{1}{2}}$ equation (11a) with the quantum-mechanical or T^2 equation (22). It is well known that the experimental results can be made to fit either the $T^{\frac{1}{2}}$ or the T^2 equation and that the constants in the two equations are related by

$$W(\text{or } P_m - K) = p - (3/2)kT, \tag{30}$$

and

$$A = A'/e^{\frac{1}{2}}T^{\frac{1}{2}}. \tag{31}$$

From equation (30) it follows that the classical work function p is larger than the quantum-mechanical work function W or $P_m - K$ by $(3/2)kT$ and that to obtain the same emission from the two distributions the curves must be shifted by $K/e - 3kT/2e$.

The Temperature Dependence of the Work Function

Thus far little has been said about the temperature dependence of the work function. While there is no good theoretical reason for expecting a large temperature dependence, there is also no good reason to expect that the work function is accurately independent of T. Experiments on contact potential and photoelectric effect indicate that there is indeed a small temperature effect.* In investigating the effect of the temperature dependence we shall limit ourselves to the quantum-mechanical equations. However, a similar treatment would be applicable to the classical or $T^{\frac{1}{2}}$ equation.

If in equation (22), w or its equivalents W or φ are independent of T, then the slope of a Richardson line is $-w/2.3$ or $-W/2.3k$ or $-\varphi e/2.3k$; the intercept is $\log U$. So that

$$b = w = W/k = \varphi e/k \quad \text{and} \quad \log A = \log U. \tag{32}$$

If the work function varies linearly with temperature,

$$w = w_0 + \alpha T \quad \text{or} \quad W = W_0 + \alpha kT$$

* For a detailed discussion see reference 8.

or

$$\varphi = \varphi_0 + \alpha(k/e)T; \tag{33}$$

where $\alpha = dw/dt$ is a constant independent of T; its units are degrees per degree. The slope of a Richardson line is now $-w_0/2.3$ so that

$$b = w_0 = W_0/k = \varphi_0 e/k$$

while

$$\log A = \log U - \alpha/2.3. \tag{34}$$

Since the slope is constant, the Richardson line is straight. This line is determined either by the empirical constants A and b or by the values of w and dw/dT in theoretical equations.

If w is a general function of T, the Richardson line will be curved. If a tangent is drawn at a point corresponding to any temperature, the slope of the tangent is $-(1/2.3)(w - Tdw/dT)$ and its intercept is $\log U - (1/2.3)dw/dT$; w and dw/dT are to be taken at the point of tangency. Hence

$$b = w - Tdw/dT$$

and

$$\log A = \log U - (1/2.3)dw/dT. \tag{35}$$

In a previous section it was shown that the slope of the Richardson line is always equal to $-h/2.3$. Hence

$$-h/2.3 = -(1/2.3)(w - Tdw/dT)$$

or

$$h = w - T(dw/dT). \tag{36}$$

This important equation gives the relation between the heat function and the work function. It is similar in form to the relation between the total energy E and the free energy F, *viz.*,

$$E = F - T(dF/dT). \tag{37}$$

The distinction between the heat function h and the work function w is strikingly brought out in Fig. 2. The slope of the Richardson line is $-h/2.3$, while the slope of a straight line connecting any point on the Richardson line with the intercept $\log U(1 - r)$ is $-w/2.3$.

The theory that the work function is indeed a function of temperature has been championed in recent times by R. Suhrmann and his collaborators. A good account of this work can be found in Volume 4 of Müller-Pouillets *Lehrbuch der Physik*. One method by which Suhrmann has shown the temperature dependence of the work function is

that of the complete photoelectric emission. The surface to be investigated is illuminated by light from a source whose temperature is varied. It is found that the resulting photo-current obeys a Richardson law and the slope of the Richardson line is taken as the work

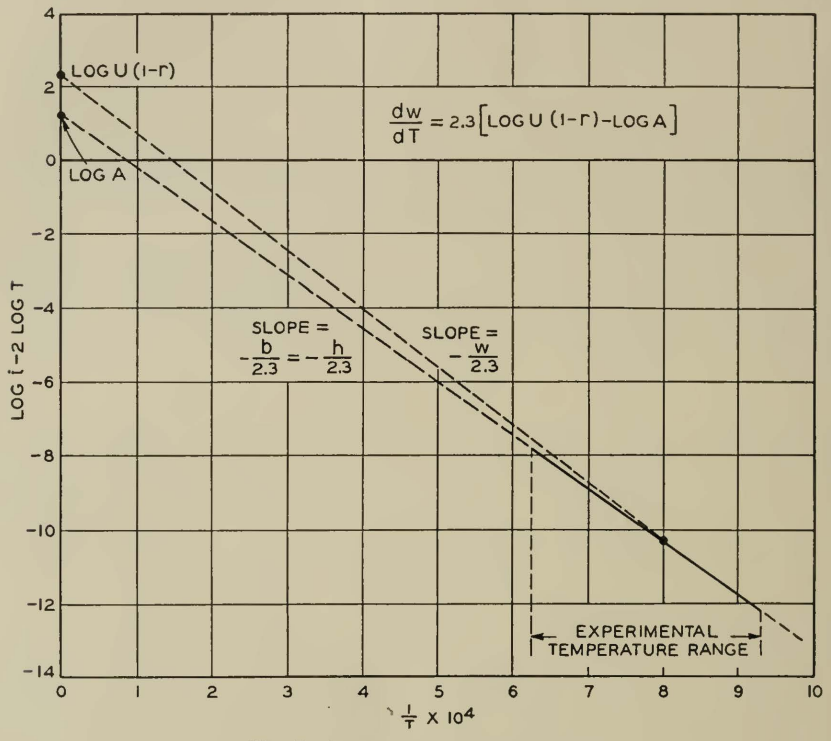


Fig. 2—Typical Richardson plot.

function. The temperature of the cathode is now altered and a new Richardson line is obtained. It is found that the slope has changed. It may be worth while to analyze critically the theory of this experiment to ascertain whether the slope is related to the work function or to the heat function.

On True and Apparent Surface Areas

One other point in the correlation between experiment and theory is to be noted. In the empirical equation, i is the current per cm.^{2} of apparent surface; while in the theoretical equation, i is the current per cm.^{2} of an ideal or true surface. The real surface in thermionic experiments is not smooth; it usually consists of a large number of

etch facets which are oriented at various angles with respect to a mean surface plane. The appearance of the surface is to be compared with an airplane view of a city whose gabled roofs have various designs and various angles. The size and shape of the etch facets depend on the material of the cathode, the crystal size, the orientation of the crystal with respect to the mean surface, the degree of heat treatment and presumably some unknown factors. Theoretically it is possible to deduce values of S, the ratio of the true surface area to the apparent surface area, for certain simple cases. Thus, Tonks [17] has computed the following average values of S: For cubic facets or 100 planes, 1.500; for dodecahedral facets or 110 planes, 1.225; for 100 and 110 planes, 1.129. Some of the assumptions on which these values are based are: (1) The surface is covered with pyramids whose sides are crystal planes; (2) the orientation of crystal axes with respect to the surface is random; (3) for a given type of etch plane or planes, the facets occur in such a way as to give a minimum surface area. No one has made a thorough investigation to test these assumptions by experiment. Some microscopic pictures of etched surfaces which I have seen showed truncated pyramids in contrast with the first assumption; they also showed sub-facets, thus violating the third assumption.

Values of S have been obtained from experiments on adsorption of gases on solid and liquid surfaces. Particularly significant experiments are those of Bowden and Rideal [18] on the adsorption of hydrogen ions deposited on metal surfaces by electrolysis of a solution of sulphuric acid. The potential of these surfaces was determined against a calomel electrode. They found that when the electrolytic current exceeded a minimum value, the surface potential increased linearly with the quantity of electricity until it reached a new steady value. For a mercury surface as well as for a thin film of platinum on mercury the potential increased by one volt for 6×10^{-6} coulomb/cm.^{2} The direction of the potential change and its amount are such as to be expected if hydrogen ions are adsorbed on the surface. For surfaces other than mercury the charge per cm.^{2} required to change the potential by one volt was S times 6×10^{-6}. They obtained the following values for S: smooth platinum, 2.0; platinum black, 2000; sandpapered nickel, 10; oxidized and reduced nickel, 50. They interpret this S as the ratio of the true area to apparent area. Their values are considerably greater than those expected from Tonks' theoretical calculations.

As a result of this it is my opinion that a considerable amount of careful work must be done before reliable values of S are obtained for

thermionic cathodes. For the present it would seem best to consider S as an unknown whose value lies somewhere between 1 and 10 for rough surfaces such as those on oxide coated filaments, and between 1 and 2 for relatively smooth surfaces such as tungsten. The exact value will no doubt depend on the exact treatment of each surface.

Fortunately the uncertainty of our knowledge of S does not seriously affect our correlation made above. It is necessary to divide values of i and A in the empirical equations by S to reduce to the basis of true surface area before comparing them with theoretical equations. The observed values of i and A should thus be reduced by 25 to 50 per cent for smooth surfaces and by larger values for rough surfaces. Thus in the case of surfaces, such as tungsten, molybdenum and tantalum, for which A has the value 60, the true A should be between about 30 and 45 as compared with a theoretical value of 120. Since the deviations from 120 are due to a temperature dependence of the work function, it means that we must postulate a somewhat larger value of α in equation (34) than otherwise.

On the Reflection Coefficient

There is still another topic that enters into the correlation of experiment and theory, namely the reflection coefficient. Thus far we have assumed that every electron whose normal component of velocity exceeded a certain value escaped while those having less than this value failed to escape. On the classical viewpoint this assumption is justified but on the quantum-mechanical viewpoint there is a finite probability that the electron considered as a wave will be reflected at the surface even though its velocity is such that it could escape; also a wave electron has a finite probability of passing through a potential peak when classically its velocity is not large enough to permit it to pass over the top of the peak. Consequently we should include an average transmission coefficient $\bar{D}$ in the theoretical emission formula. $\bar{D} = 1 - \bar{r}$ where $\bar{r}$ is the average reflection coefficient. Equation (22) then becomes

$$i = U(1 - \bar{r})T^2 \exp. (- w/T). \tag{38}$$

A number of writers [2, 3] have attempted to explain the deviations between A and U by postulating such values of $\bar{r}$ that $A = U(1 - \bar{r})$. This explanation is possible only for cases for which $A < U$ since $0 < \bar{r} < 1$. Even when $A < U$ the numerical values turn out to be such that the difference between A and U cannot be accounted for by computed probable values of $\bar{r}$. These values of $\bar{r}$ are determined chiefly by the shape of the curve giving the work an electron must do

to get to various distances from the surface. Only when this work-distance curve is postulated to have a high sharp peak within a few atom diameters from the surface, is it possible to deduce values of $\bar{r}$ which are appreciable. Now we have good reasons [8] for believing that no such peaks exist, and that the maximum of the work distance curve occurs at relatively large distances from the surface in a region where the forces on the electron are given by the well-known image law. For the latter type of curve, the computed value of $\bar{r}$ is less than 0.07 which is negligibly small. Nordheim, who first pointed out that the transmission coefficient might differ from unity says: "However, the exact computation taking into account the image force which must necessarily be considered, has shown that such a rounded-off potential curve yields a value of $\bar{D}$ which differs inappreciably from 1.0." * A more complete case showing that the values of the reflection coefficient are negligibly small is given by Becker and Brattain.[8]

The Effect of Accelerating Fields and Retarding Potentials

Thus far we have considered how the emission current and the work function depend on the emitting surface and its temperature; we have implicitly assumed that the current was "saturated" or that every electron which escaped from the surface was collected by the anode. It is, however, well known that the emission current depends also on the applied fields and the applied potentials. In considering the effects of these fields and potentials we shall incidentally obtain an insight into the nature of some of the forces responsible for the work function.

For simplicity consider a large plane cathode and parallel to it a large plane anode. If the temperature of the cathode is high enough to emit a small but appreciable current, $\log i$ will vary with the potential applied to the anode in the manner shown in Fig. 3. In drawing curve 1 in this figure three more simplifying assumptions have been made; namely (1) that the contact potential between cathode and anode is zero; (2) that all portions of the cathode and anode have the same work function, and (3) that space charge effects are negligible. The effect of these assumptions will be considered later.

The curve in Fig. 3 naturally divides itself into two portions: the part to the left of $V_a = 0$ corresponds to retarding potentials while the part to the right of 0 corresponds to accelerating potentials. In the latter region the current is said to be "saturated" although strictly speaking the current is never saturated but increases indef-

* See Section by Nordheim in Müller-Pouillets *Lehrbuch der Physik*,[4] Vol. IV, "Elektrizitat und Magnetismus," Part IV, p. 294. See also footnote 2 on p. 290.

initely as V increases. Obviously, the effect of V on the current is quite different in the two regions and these two regions require different explanations.

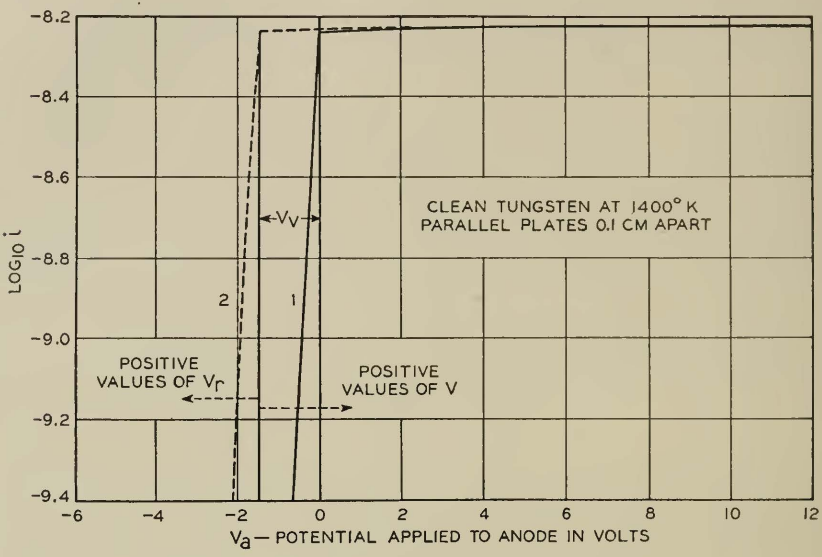


Fig. 3—Retarding potential curves for parallel plates.

Retarding Potentials

Consider first the retarding potential region in which log i decreases linearly with $-V_a$, the applied potential V. The explanation is to be found in extending the theories which were used to derive the Richardson equation. In that derivation it was implicitly assumed that the only forces which the escaping electron had to overcome were the cathode surface forces, and that any electron which escaped from the cathode would reach the anode. If a retarding potential V_r* is applied to the anode then only those electrons whose normal component of velocity u exceeds a value u_a can reach the anode; where u_a is given by

$$mu_a^2/2 = (\varphi_c + V_r)e \tag{39}$$

in the classical case, or

$$mu_a^2/2 = P_m + eV_r \tag{40}$$

in the quantum-mechanical case. Figure 4, curve 1, illustrates the

* In discussing retarding potentials it is convenient to consider retarding potentials as positive even though the anode potential is negative, so that $V_r = -V$.

potential energy of an electron at various distances between the cathode and anode when the anode is V_r volts negative to the cathode. It is tentatively assumed that the anode work function φ_a is the same as the cathode work function φ_c; this is another way of saying that

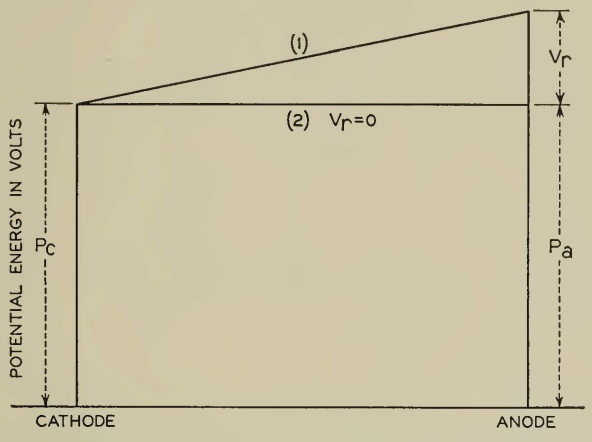


Fig. 4—Potential distribution between parallel plates; $P_a = P_c$.

the contact potential is zero. When $V_r = 0$ nearly all the space between the cathode and anode is field free as shown in curve 2; only in the immediate neighborhood of the cathode or the anode is the electron subjected to any forces. When a retarding potential is applied the electrons must have sufficient energy to pass over the maximum in curve 1, Fig. 4, in order to reach the anode.

To determine the number of electrons that can reach the anode we integrate equation (10) or (20), from $u = u_a$ to $u = \infty$ where u_a is given by equation (39) or (40), respectively. Whether we use the classical or the quantum-mechanical statistics we arrive at the same result.

$$i = Ne = i_0 \exp.\ (-\ V_r e/kT), \tag{41a}$$

or

$$\log i = \log i_0 - (e/2.3kT)V_r, \tag{41b}$$

where $i_0 = i$ when $V_r = 0$. The slope of the straight line in Fig. 3 should thus be $e/2.3kT$.

If φ_a and φ_c are not equal, the field between anode and cathode will not be zero when the applied potential is zero; a contact potential or Volta potential V_V will exist between a point just outside the cathode and a point just outside the anode. To produce zero field a potential

must be applied which neutralizes the Volta potential. The true potential V between anode and cathode is the sum of the applied potential V_a and V_V or

$$V = V_a + V_V. \tag{42}$$

Since

$$V_V = \varphi_c - \varphi_a, \tag{43}$$

$$V = V_a + \varphi_c - \varphi_a. \tag{44}$$

Since

$$V_r = -V,$$

$$V_r = -V_a - (\varphi_c - \varphi_a) = -V_a + \varphi_a - \varphi_c. \tag{45}$$

V_r is the true value of the retarding potential and these values of V_r are to be used in equations (39), (40) and (41). V and V_r are measured from the break point in Fig. 3. Figure 5 illustrates the case

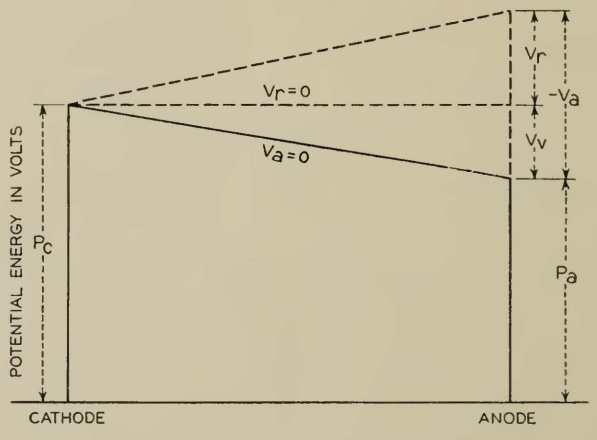


Fig. 5—Potential distribution between parallel plates; $P_a < P_c$.

when $\varphi_c > \varphi_a$. For $V_a = 0$, V is positive and equal to $\varphi_c - \varphi_a$. To produce zero field V_a must be negative and equal to $\varphi_c - \varphi_a$. The dashed line gives the potential energy distribution for a somewhat larger negative applied potential.

When the contact potential is not zero, the break point in the log i *versus* V_a curve will occur when the field is zero or when

$$V_a = -(\varphi_c - \varphi_a).$$

This is illustrated in Fig. 3 by the dashed line for a case for which $\varphi_c > \varphi_a$.

Usually thermionic experiments are not performed with plane parallel cathodes and anodes but with a small cylindrical cathode concentric with a cylindrical anode. In the cylindrical case, the normal or radial component of velocity is not the only one which determines whether the electron will reach the anode. Schottky [19] derived the following formula for this case on the assumption that the emitted electrons leave the filament with a velocity distribution given by Maxwell's law (equation (7)) for a temperature T. As we have seen above both the classical and the Fermi-Dirac theory predict this distribution for the electrons which escape from the filament. This formula replaces equation (41).

$$i = i_0(2/\pi^{\frac{1}{2}}) \left[[V_r e/kT]^{\frac{1}{2}} \exp.\ (-\ V_r e/kT) + \int_{(V_r e/kT)^{\frac{1}{2}}}^{\infty} \exp.\ (-x^2)dx \right]. \quad (46)$$

It is assumed that the diameter of the cathode is small compared to the diameter of the anode, and that the current is not limited by space charge. Table I gives values of $\log_{10}(i_0/i)$ for values of $V_r e/kT$ taken from an article by Germer.[20]

TABLE I

VALUES OF $\log_{10} i_0/i$ FOR VARIOUS VALUES OF $V_r e/kT$ (GERMER) [20]

$V_r e/kT$	1	2	3	4	5	6
$\log_{10}(i_0/i)$	0.2423	0.5827	0.9523	1.3371	1.7312	2.1318
	7	8	9	10	11	12
	2.5369	2.9455	3.3567	3.7698	4.185	4.6024
	14	16	18	20	25	
	5.4398	6.2812	7.1245	7.9714	10.0978	

Figure 6 shows various ideal plots of log i *versus* V for cathodes of clean tungsten and thoriated tungsten. It is assumed that the anode is clean tungsten. Curves 1, 2 and 3 are for a clean tungsten cathode at temperatures of 1400, 1550 and 1700° K., respectively. Curves 4 and 5 are for a thoriated tungsten cathode at 1400° K. activated to such an extent that the work function is 4.03 and 3.53 volts, respectively. The dashed lines indicate the currents for a plane cathode and parallel anode.

Curves 1, 4 and 5 illustrate an important theorem which follows from the analysis on contact potential given in connection with Figs. 3, 4 and 5. The theorem is: The current collected by an anode is independent of the work function of the cathode provided that the

cathodes are in the same position and have the same temperatures and that the retarding potential is sufficiently great. This theorem was verified experimentally by Davisson.[21]

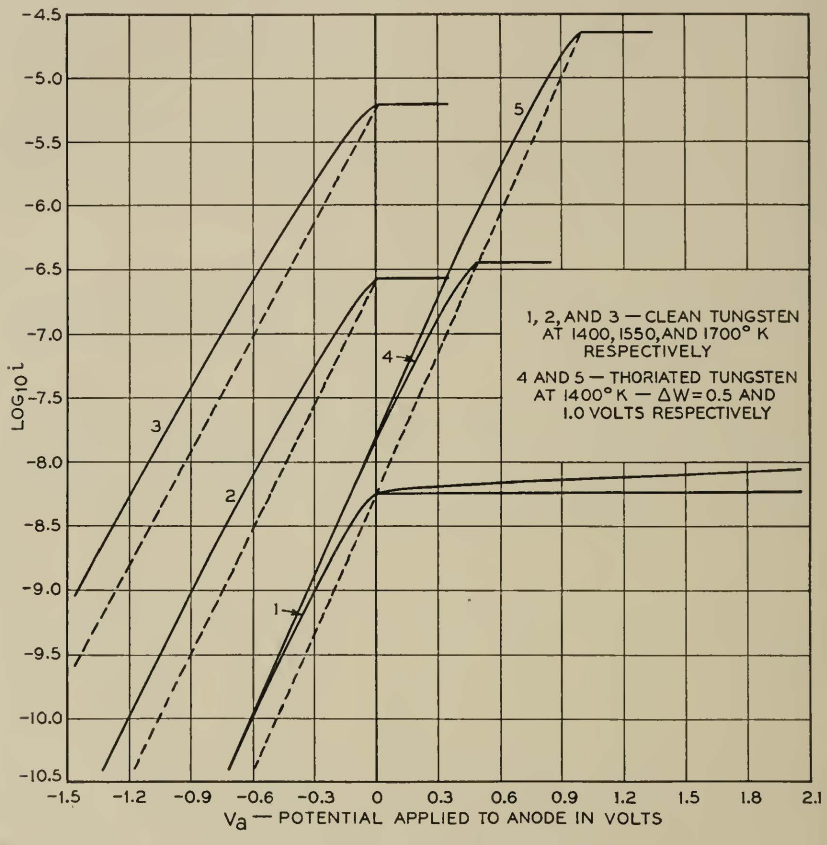


Fig. 6—Retarding potential curves for cylindrical electrodes.

The curves shown in Fig. 6 are for ideal conditions. Experimental conditions frequently differ from ideal conditions in at least five respects: (1) The various portions of the cathode are not at the same potential; (2) the work function of the cathode is non-uniform; (3) the work function of the anode is non-uniform; (4) the temperature of the cathode is non-uniform; and (5) the current is limited by space charge rather than by the applied potential. The last two conditions are discussed in other treatises. The first condition is usually due to the fact that the cathode is a filament and is heated by passing a current through it. As a result the observed curve is the sum of a series of

elementary curves, each one of which is shifted along the V axis by the amount of the potential drop along the filament. Such a sum curve consists of a straight line having the correct slope at sufficiently great retarding potentials; the sharpness of the break point in the curve is, however, destroyed and the slope of the curve for small retarding potentials is decreased, thus simulating the ideal curve for a higher temperature. The best way to obviate this difficulty is to work with equipotential cathodes which are heated indirectly. This makes the construction of the tube more difficult and has been used only by Demski.[22] Most of the work has been done on filaments which were heated intermittently by means of a mechanical or electrical commutator.

In this way Germer,[20] Demski and others have shown that the distribution of thermionically emitted electrons is Maxwellian and corresponds to a temperature which is equal to the temperature of the cathode to within less than 5 per cent. Germer worked with tungsten for a series of temperatures between 1440 and 2475° K. Demski worked with tungsten and with oxide-coated filaments. He used a mechanical and an electrical commutator and also worked with equipotential cathodes. Nottingham [21] and others have reported that for thoriated tungsten and oxide-coated filaments the temperature computed from the shape of the log i *versus* V curve for small retarding potentials was about 1.5 times the temperature of the cathode. Nottingham explains this as due to a sharp peak in the potential distance curve through which a part of the wave electrons can penetrate. In my opinion it is much more likely that these observations are due to non-uniformities in the work function of the cathode and the anode.

If the work function of the cathode is non-uniform, the observed curve should result from the summing up of the currents for a series of curves somewhat similar to curves 1, 4 and 5 in Fig. 6. The sum curve will have the correct slope at sufficiently great retarding potentials; but at low values of V_r the slope should be too small corresponding to too high a temperature. The break point will be less sharp.

If the work function of the anode is non-uniform, the elements of the sum curve will consist of a series of ideal curves shifted parallel to the V axis. The sum curve will again yield correct temperatures at large values of V_r but too high temperatures at small values of V_r. That the work function of cathodes is usually non-uniform will be shown in the next section. It is to be expected that the anode work function will also be non-uniform since the anode is more difficult to heat treat than the cathode. However, when one takes into account the effect of these non-uniformities, it is seen that the experiments

abundantly confirm the theory that the distribution of velocities of thermions is that given by Maxwell's law for an ideal gas.

Accelerating Fields

As illustrated in Fig. 3, when positive potentials are applied to the anode, log i increases continuously; but the rate of increase becomes progressively less so that the current is almost independent of the anode potential. For many purposes one can safely say that the current is saturated; for some purposes, however, it is very important to consider this lack of saturation. More specifically a consideration of this effect gives us direct evidence of some of the forces which are responsible for the work function. Thus, as the electron escapes from the surface, it must overcome certain forces which tend to pull it back. The electrical fields responsible for these forces presumably decrease with the distance from the surface. Call them surface fields F_s. When a positive potential is applied to the anode, a field F_a is produced near the surface of the cathode which tends to help the electrons escape. The value of the field depends on the dimensions of the cathode and anode. This applied field neutralizes the surface field at some distance z from the surface; call this distance the critical distance z_c. If an electron can reach the critical distance, it will escape, since beyond this distance the sum of the applied and surface fields pulls the electron toward the anode. Obviously the critical distance moves closer toward the cathode as the applied field is increased.

A more quantitative concept is obtained by considering the effect of the applied field on the potential energy-distance curve similar to Fig. 4. Now, however, we will be concerned more particularly with regions close to the cathode, so that we will greatly enlarge the distance scale. Figure 7, curve 1, shows such a curve when the true field between cathode and anode is zero. The true field F is the algebraic sum of the applied field F_a and the field produced by the contact potential. Frequently it is convenient to use the term " applied field " in the sense of " true field," i.e., including the contact potential field. An applied field decreases the potential energy of the electron as shown in curve 2. The net potential energy is shown in curve 3.

The maximum height in curves 1 or 3 represents the work function φ in the classical theory or the quantity P_m/e in the quantum theory. In the latter case, since $\varphi e = P_m - K$ from equation (24) and since K does not depend on the applied field,

$$\Delta\varphi = \Delta P_m/e \tag{47}$$

or the decrease in the work function due to an applied field is equal to the decrease in the maximum of the potential energy-distance curve. Since P depends on F, the true field (applied + contact potential field),

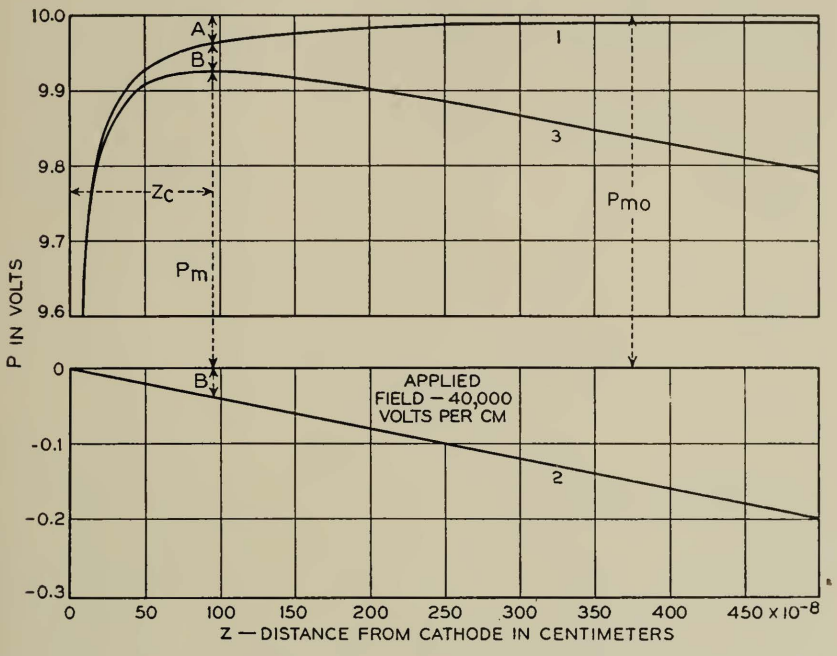


Fig. 7—Potential energy *versus* distance from cathode surface for image force and applied fields of zero or 40,000 volts/cm.

it will be convenient to designate values for curve 1 for which $F = 0$ by the subscript 0. Then

$$P = P_0 - Fez, \tag{48}$$

and

$$dP/dz = dP_0/dz - Fe. \tag{49}$$

At the maximum in the P *versus* z curve, $dP/dz = 0$ and $z = z_c$. Hence

$$(dP_0/dz)\,|_{z_c} = Fe. \tag{50}$$

We require an expression for ΔP_m, the decrease in P_m due to the field F. From Fig. 7 it is clear that

$$\Delta P_m = \text{distance } A + \text{distance } B = P_{m0} - P_0|_{z_c} + Fez_c \tag{51}$$

$$d(\Delta P_m)/dF = -\,dP_0/dF|_{z=z_c} + Fe(dz_c/dF) + ez_c. \tag{52}$$

Now from equation (50)

$$dP_0/dF|_{z=z_c} = dP_0/dz|_{z_c} dz_c/dF = Fedz_c/dF. \tag{53}$$

Hence from equations (52) and (53)

$$d(\Delta P_m)/dF = ez_c. \tag{54}$$

Combining this with equation (47) we obtain

$$d(\Delta\varphi)/dF = z_c. \tag{55}$$

Now from

$$\log i = \log U - 2 \log T - (\varphi - \Delta\varphi)e/2.3kT = \log i_0 + \Delta\varphi e/2.3kT \tag{56}$$

we obtain

$$d \log i/dF = (d(\Delta\varphi)/dF)e/2.3kT. \tag{57}$$

Combining this with equation (55) we obtain

$$d \log i/dF = (e/2.3kT)z_c. \tag{58}$$

This equation which was first derived by Becker and Mueller [24] allows us to obtain numerical values for z_c from the slope of the experimental log i *versus* F curve. At z_c the surface field F_s is equal to the applied field F. Hence a plot of F_s *versus* z can be obtained, and by integrating this from z to ∞, values of $P_{m0} - P_0$ can be obtained for various values of z greater than some minimum value corresponding to the largest value of F.

A particular case of a surface field, namely, that given by the image law, is especially significant. In this case $F_s = e/4z^2$ and it can be shown that the distances A and B in Fig. 7 are equal. At the critical distance $F = F_s$ and $F = e/4z_c^2$ or

$$z_c = (e/4F)^{\frac{1}{2}}. \tag{59}$$

By substitution in equation (55) and integration from 0 to F it follows that

$$\Delta\varphi = (eF)^{\frac{1}{2}}. \tag{60}$$

Substituting this in equation (56) yields

$$\log i = \log i_0 + (e^{\frac{1}{2}}/2.3kT)\sqrt{F}, = \log i_0 + (1.91/T)\sqrt{F}. \tag{61}$$

This equation, which was first derived by Schottky [35] and is called the Schottky equation or law, predicts that a plot of log i *versus* $\sqrt{F}$ should yield a straight line whose slope is $e^{\frac{1}{2}}/2.3kT$ or $1.91/T$.

Experimental log i *versus* $\sqrt{F}$ plots are found to be straight and to have approximately the right slope for sufficiently high applied fields. At low fields, the line is curved and the experimental slopes are greater than the predicted values. These deviations from Schottky's law are slight in the case of clean surfaces but become quite pronounced for composite surfaces such as thorium on tungsten or cæsium on tungsten. We shall show below that these deviations can be ascribed to non-uniformities in the work function for different regions of the cathode surface. The prediction that the slope should vary as $1/T$ has been verified by Dushman's experiments.[10]

In so far as Schottky's law is verified by experiment, we can conclude that the escaping electron must in certain regions overcome the forces due to its own image and no other forces. Thus for clean surfaces the electron is acted on only by its image force from about 10^{-7} to about 50×10^{-7} cm. from the surface; for composite surfaces this region will depend on the size and degree of the non-uniformities; for a particular surface of thorium on tungsten the image law held from 6×10^{-7} to about 20×10^{-7} cm. When the critical distance is very small, the emission is modified because of sharp points on the surface and because of "intense field" emission.[24] When the critical distance is larger than about 100×10^{-7} or 1×10^{-5} cm. there are apparently other fields superimposed on the image field. These are larger than the image field at these distances and thus cause deviations from the Schottky law. As we shall see later these fields are due to non-uniformities on the surface. From all this we can conclude that an appreciable part of the work function is due to the image force and to other surface fields.

Table II shows values of $\Delta\varphi$, z_c, log i/i_0 and i/i_0 if the surface field is given by the image law.

TABLE II

VALUES OF $\Delta\varphi$, z_c, LOG i/i_0 AND i/i_0 IF THE SURFACE FIELD IS GIVEN BY THE IMAGE LAW

F, volts/cm	0	100	1000	10,000	40,000
$\sqrt{F}$	0	10	31.6	100	200
$\Delta\varphi$, volts	0	0.0038	0.0120	0.0378	0.0755
z_c cm	∞	1.89×10^{-5}	5.98×10^{-6}	1.89×10^{-6}	9.45×10^{-7}
log i/i_0, $T = 1000°$ K	0	0.0191	0.0604	0.191	0.382
i/i_0 "	1.000	1.045	1.149	1.553	2.410
log i/i_0, $T = 2000°$ K	0	0.0096	0.0302	0.096	0.191
i/i_0 "	1.000	1.022	1.072	1.25	1.55

The Use of the Term "Effective Work Function"

There has been a tendency to restrict the term work function to zero field and to use "effective work function" for accelerating

fields.[1, 2, 23] In my opinion this tendency is to be deplored since it is unnecessary and places too much emphasis on zero field. Richardson's equation, and the theories underlying it are just as applicable for accelerating fields as they are to zero field. Since the work function depends on T as well as F, it would be just as logical to coin a new name for the work function at any temperature other than $T = 0$. It seems to me more desirable to retain " work function " in its general sense and to recognize that it may depend on temperature and on the accelerating field. The work function or more precisely the quantity P_m/e would be defined as the work required to take an electron at rest inside the metal to a point at distance z_c from the surface, where z_c is the distance at which the accelerating field is equal to the surface field.

The Effect of Non-uniform Work Functions: Patch Theory

We shall now consider how the emission is altered if the cathode work function is non-uniform. Here again we shall find it necessary to consider the effect of such non-uniformities on the P *vs.* z curves, i.e., on the curves for the potential energy of the electron *versus* distance from the surface. For the present we shall not consider the causes for the mechanism which is responsible for the non-uniformities. We shall assume that the surface work functions are non-uniform. As a consequence, local fields must exist between the various regions having different work functions. The effect of these fields on the log i *vs.* F curve will depend on the size, shape and degree of the non-uniformities.

The Simple Condenser Analog

Consider a simple case: The cathode is uniform except in a circular region of radius R which is covered with a positive charge density σ, a short distance l above the surface. There is induced at a distance l below the surface the image charge density $-\sigma$. These two sheets of charge act like a finite circular condenser. The field between the condenser plates will be $4\pi\sigma$ e.s.u. or $300 \times 4\pi\sigma$ volts/cm. if σ is expressed in e.s.u. If the zero of potential is taken at the surface of the metal or at the center of the condenser, the potential just outside the outer sheet of charge will be $300 \times 4\pi\sigma l$. If the sheet of charge were infinite in extent or if R were several times the distance from cathode to anode, then the field outside the condenser would be zero, and the work function of the patch for electrons would be reduced by $300 \times 4\pi\sigma l$ or by $300 \times 2\pi M$; where $M = 2\sigma l$ the moment per cm.^{2} of surface. Actually there is a field outside the finite condenser which tends to pull an electron back to the surface. The integral of this

field out to infinity or a distance large compared to R is just sufficient to reduce the potential to zero again. Hence when the applied field is zero so that z_c is very large, the work function over the condenser or patch is not reduced at all. Calculations show that if a small accelerating field is applied, the work function is reduced more than it would have been if there had been no condenser. For a sufficiently large applied accelerating field, z_c moves so close to the surface that $z_c \ll R$. At this distance the potential at z_c due to the sheets of charge will not

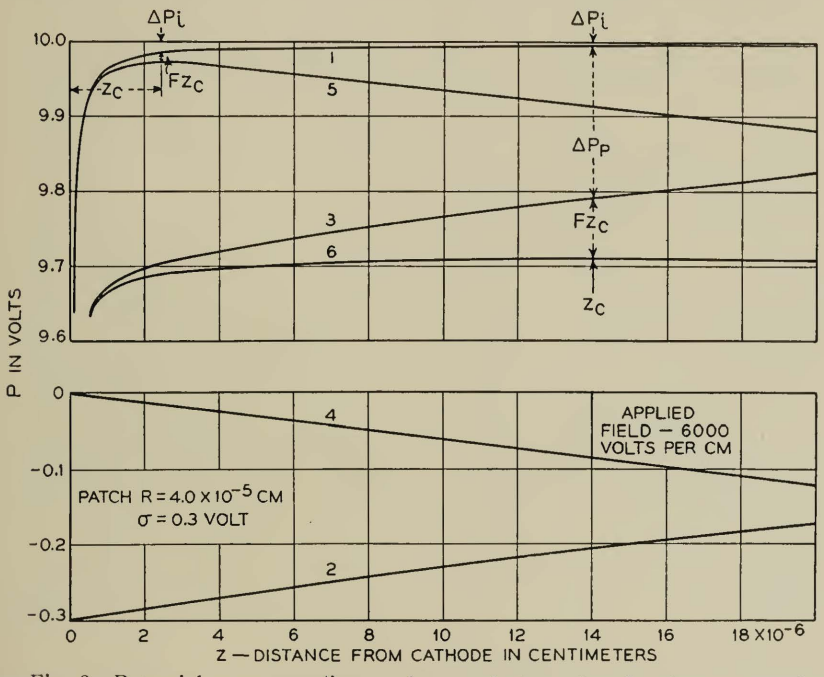


Fig. 8—Potential energy *vs.* distance from cathode surface for electrons moving against image field plus field due to a uniform circular patch for applied fields of zero or 6000 volts/cm.

differ greatly from the potential just outside the condenser. The work function will now be $300 \times 4\pi\sigma l$ less than it would have been without the patch. Hence the extra reduction of the work function due to the patch is zero at zero accelerating field and increases with this field up to its limiting value $300 \times 4\pi\sigma l$.

To treat this case in more detail consider the curves in Fig. 8. Curve 1 is a plot of P_i/e in volts *vs.* z in cm. P_i is the potential energy due to the image force and is given by

$$P_i/e = P_\infty/e - 300e/4z, \tag{62}$$

where P_∞ is the potential energy when $z = \infty$ and $F = 0$; $e = 4.774 \times 10^{-10}$. Curve 2 is the potential energy P_p due to the patch or condenser along a line normal to the surface at the center of the condenser. R has been taken as 4×10^{-5} cm. The equation of this curve is

$$P_p = 1200\pi\sigma l[z/(z^2 + R^2)^{\frac{1}{2}}]. \tag{63}$$

In the derivation of this formula it has been assumed that either $R \gg l$ or else that $z \gg l$; for our case the first of these assumptions will always be fulfilled so that the formula is applicable when z is equal to or larger than l; it is not applicable for z less than l.

Curve 3 is the algebraic sum of P values for curves 1 and 2. It represents the potential energy along the central normal due to the image and patch fields. Curve 4 represents the potential energy due to an applied field of 6000 volts/cm. Curve 5 is the sum of curves 1 and 4; curve 6 that of 3 and 4.

The effect of the applied field is to reduce the critical distance z_c and the work function. A given applied field will reduce z_c more for a clean surface than for one with the patch; but the converse is true for the work function. The reduction in the work function is equal to the reduction in the value of P_m/e. This consists of three parts as indicated in the figure for curve 6. $\Delta P_i/e$ is the decrease in P due to the image forces from $z = z_c$ to $z = \infty$; $\Delta P_p/e$ is the decrease due to the patch field from $z = z_c$ to $z = \infty$; Fz_c is the decrease in P due to the applied field from $z = 0$ to $z = z_c$. These quantities can be evaluated after one has determined the value of z_c as follows:

The peak or maximum in curve 6 occurs at a value of $z = z_c$ at which

$$dP/dz = dP_i/dz + dP_p/dz = Fe. \tag{64}$$

From equation (62) $dP_i/dz = 3.58 \times 10^{-8}/z^2$,

and from equation (63)

$$\frac{dP_p}{dz} = -(1200\pi\sigma l)e\left[-\frac{1}{(z^2 + R^2)^{\frac{1}{2}}} + \frac{z^2}{(z^2 + R^2)^{\frac{3}{2}}}\right]$$

$$= 1200\pi\sigma l e \frac{R^2}{(z^2 + R^2)^{\frac{3}{2}}}$$

so that

$$F = \frac{3.58 \times 10^{-8}}{z^2} + 1200\pi\sigma l\left[\frac{R^2}{(z^2 + R^2)^{\frac{3}{2}}}\right]. \tag{65}$$

From this equation F is plotted for various values of z. For any value of F a value of z can be read off. This value of z will be the critical

distance z_c. To obtain $\Delta\varphi$, z_c is substituted in the equation,

$$e\Delta\varphi = \Delta P_i + \Delta P_p + Fz_c = P_\infty - \frac{3.58 \times 10^{-8}}{z_c} + 1200\pi\sigma l\left(1 - \frac{z_c}{(z^2 + R^2)^{\frac{1}{2}}} + Fz_c\right). \quad (66)$$

This $\Delta\varphi$ is the decrease in the work function for a region near the center of the patch. For other regions on the patch $\Delta\varphi$ will be smaller; for regions on the uncovered portion of the surface $\Delta\varphi$ will be still smaller until at large distances from the patch $\Delta\varphi$ will correspond to the $\Delta\varphi$ appropriate for the image law.

To obtain the effect of the patch on the log i *vs.* F or log i *vs.* $\sqrt{F}$ curve it is necessary to divide the entire surface into small regions, compute $\Delta\varphi$ for each and substitute these in equation (56); the values of i_0 in this equation are the same for all regions of equal area since at large distances P_∞ has the same value over all regions. The values of i are then added up for all regions and log i is plotted *vs.* $\sqrt{F}$. Since this process is very tedious, and since in most thermionic experiments one is not likely to deal with a single patch, it is not worth while to make such an exact computation. It is, however, instructive to make some further computations based on simplifying assumptions.

Suppose we assume: (1) That for all regions on the patch, $\Delta\varphi$ has the same value as for the central region, and (2) that the current from the patch is large compared to the current from the uncovered portions of the surface. These assumptions approximate the true conditions for some cases and the errors due to the first assumption tend to balance out those due to the second. If " a " is the area of the patch then

$$\log ai = \log a + \log U + 2\log T - \varphi e/2.3kT + (\Delta\varphi)e/2.3kT = \log ai_0 + \Delta\varphi e/2.3kT, \quad (67)$$

where ai_0 is the current from the patch area when $F = 0$.

Hence
$$\log i/i_0 = (\Delta\varphi e/2.3kT). \quad (68)$$

Values of $\Delta\varphi e$ obtained from equation (66) are substituted in equation (68) and log i/i_0 is plotted as a function of $\sqrt{F}$. Figures 9 and 10 show such plots.

Figure 9 shows the effect of varying the radius R of the patch while the charge density σ is kept constant. The value of σ is so chosen that

$1200\pi\sigma l$ is equal to 0.3 volt. It will be convenient to treat σ as if it were expressed in volts, i.e., as if σ stood for $1200\pi\sigma l$. If the patch were very large σ in volts would be the decrease in the work function due to the patch. It is to be noted that a typical curve starts along a

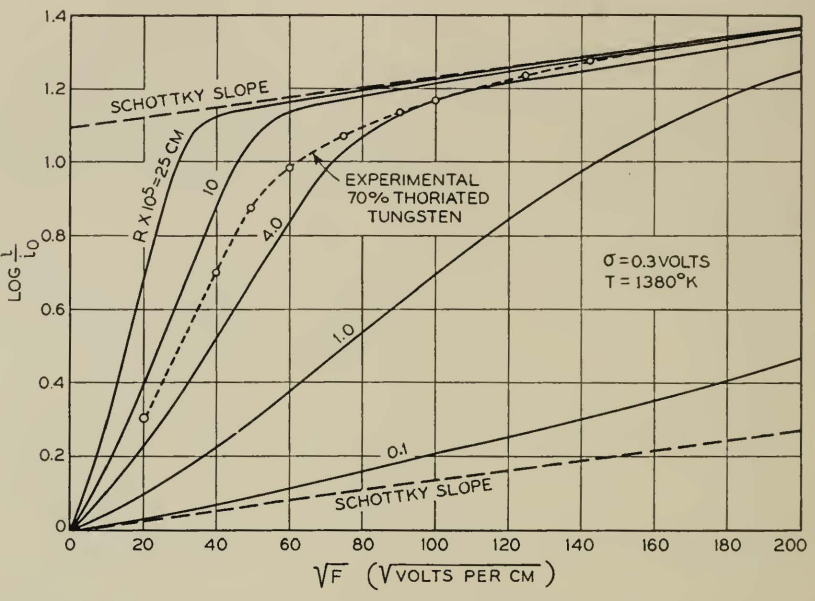


Fig. 9—Variation of emission current with applied field for circular patches of various sizes; T and σ constant. Comparison with experimental curve for thoriated tungsten.

line having the Schottky slope; but soon it rises at a much more rapid rate and continues until it almost reaches an upper line having the Schottky slope; then it bends rather sharply and approaches this line asymptotically. For the larger patches, the curve starts to rise at very small values of $\sqrt{F}$ and it is very steep. For smaller values of R, the curve follows the lower Schottky line for an appreciable distance and its slope never attains very large values; also the place at which it bends toward the upper Schottky line moves to large values of $\sqrt{F}$. Note also that as long as σ is constant all curves are bounded by the same two Schottky lines.

Figure 10 shows the effect of varying σ while R is kept constant. The distance between an upper Schottky line and the lower Schottky line is directly proportional to σ. In fact this shift is given by

$$\Delta \log i/i_0 = \sigma e/2.3kT. \tag{69}$$

It is also apparent from Fig. 10 that increasing σ results in a steeper curve and in an increase in the value of $\sqrt{F}$ at which the curve bends toward the upper Schottky line.

Actually, of course, the observed current will be composed of the current from the uniform part as well as that from the patch. The

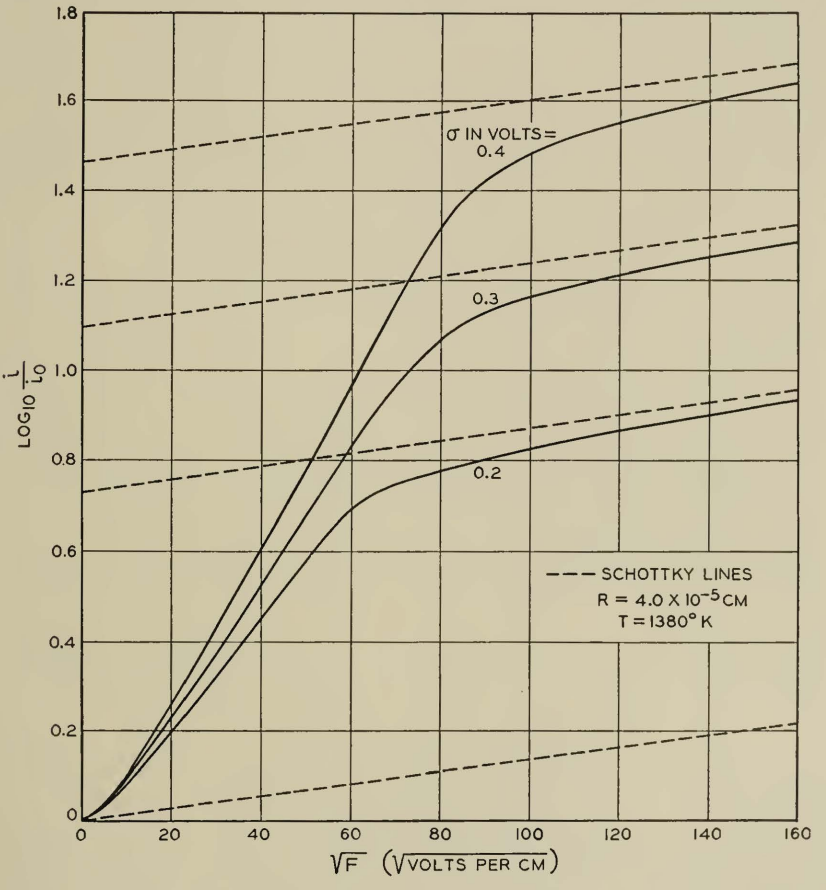


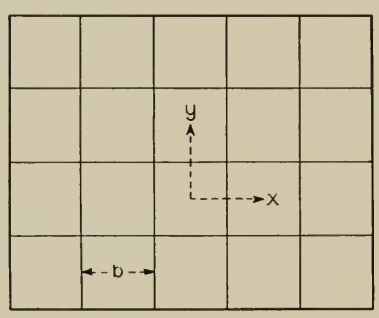
Fig. 10—Variation of emission current with applied field for circular patches; σ variable, T and R constant.

amount by which the patch current influences the total current will depend on the ratio of the patch area to the total area.

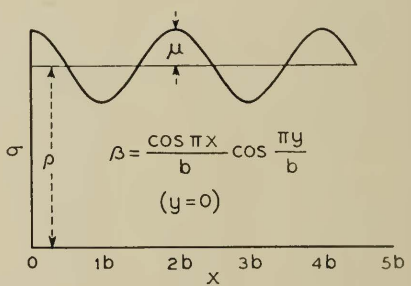
The Hill and Valley Checkerboard

Instead of being covered with a single patch, the surface in the case of most thermionic cathodes consists of numerous patches of varying

sizes and varying work functions above and below some mean value. To treat this case would obviously require very complex expressions. We can simplify the problem without departing too far from actual conditions by postulating a surface which is divided up into a large number of squares arranged in a checkerboard fashion. We might suppose that all black squares have the same σ and all white squares are bare or else have a smaller σ. It turns out, however, that the formulas and the computations are much simpler if we suppose σ is largest at the center of each black square and is least at the center of each white square; between the centers σ is given by a cosine law. In other words on the black squares we have a hill of charge while on the white squares we have a valley of charge. It will be found that such a charge distribution predicts changes in emission with applied fields, which agree rather well with experiment if the size of the squares is comparable to the crystal size and the difference in contact potential between the hills and valleys corresponds to several tenths of a volt.



(A)



(B)

C	B	C	C'	B'	C'
B	A	B	B'	A'	B'
C	B	C	C'	B'	C'

SUB CHECKERS

(C)

Fig. 11—A. Checkerboard array. B. Charge distribution for hill and valley checkerboard. C. Subdivision of checkers.

To represent such a charge distribution, choose the origin of coordinates at the center of a covered square; let x be measured parallel to one edge of the squares while y is measured perpendicular to this edge as indicated in Fig. 11A. Let the length of each square be b.

Then the surface charge density σ is given by

$$\sigma = \rho + \mu \cos(\pi x/b) \cos(\pi y/b) = \rho + \mu\beta, \qquad (70)$$

in which ρ is the mean value of σ, $\rho + \mu$ is the maximum value of σ, $\rho - \mu$ is the minimum σ, and $\beta = \cos(\pi x/b) \cos(\pi y/b)$; β has values between $+1$ and -1. It readily follows that along the edges of the squares $\beta = 0$ and $\sigma = \rho$. Figure 11B shows σ as a function of x when $y = 0$, b, $2b$ or nb.

The great advantage of this particular charge distribution is that we can represent the potential due to this charge and its image *at any point* above the surface by means of a comparatively simple formula, *viz.*,

$$P_\sigma/e = -300 \times 4\pi l[\rho + \mu\beta \exp(-\sqrt{2}\pi z/b)], \qquad (71)^*$$

P_σ is the potential energy of an electron due to the charge distribution at a point which is z cm. above the surface over a region at which the charge density is σ. The charge distribution is located in a plane which is l cm. above the surface. This charge distribution induces a corresponding negative charge distribution at $z = -l$, i.e., l cm. below the surface. ρ and μ are in e.s.u. of charge per cm.^{2} Sometimes it will be convenient to treat ρ and μ as if they were expressed in volts, i.e., as if ρ and μ stood for $1200\pi\rho l$ or $1200\pi\mu l$, respectively. The total potential energy of an electron at z cm. from the surface is given by P in

$$\begin{aligned} P &= P_i + P_\sigma - Fez \\ &= P_\infty - 300e/4z - 1200\pi l[\rho + \mu\beta \exp(-z\sqrt{2}\pi/b)] - Fez, \end{aligned} \qquad (72)$$

where $P_i = P_\infty - 300e/4z$ and $e = 4.774 \times 10^{-10}$.

In Fig. 12, curve 1 shows P_i *vs.* z; curve 2 shows P_σ for various values of β; the curve for $\beta = +1$ is for the normal taken at the center of a hill; $\beta = -1$ is for the center of a valley; $\beta = 0$ is for the edge of the squares; all other curves must lie between those for $\beta = +1$ and $\beta = -1$. Curves 3 show $P_i + P_\sigma$ for $\beta = 1$, 0 and -1.

For all values of β between 1 and 0 the curves have the same maximum value which occurs when $z_c = \infty$. The value of this maximum is $P_\infty - 1200\pi l\rho$. This means that for all points of a hill checker the work function is reduced by the same amount, namely, $1200\pi l\rho$;

* For the derivation of this and several other formulas I am indebted to Professor V. Rojansky now at Union College, who worked with me on this problem in the summer of 1930.

and this amount is the same as would occur if the charge density of the hill and valley checkers were uniformly distributed over the entire surface. On the other hand for β between 0 and -1, i.e., for points

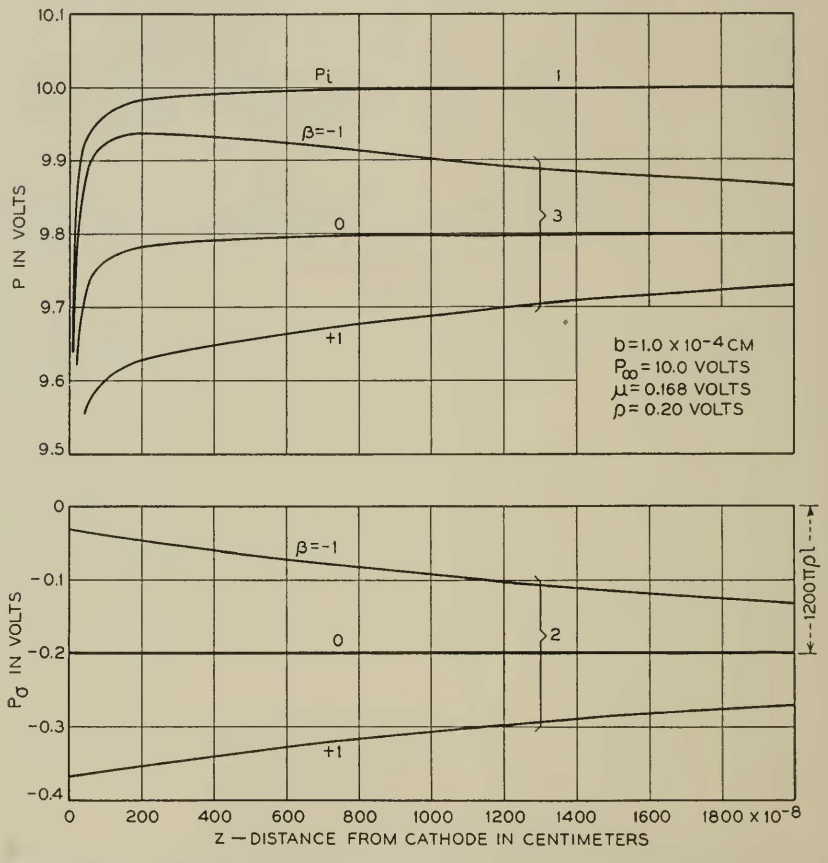


Fig. 12—Potential energy *vs.* distance from the surface above the center of the different subcheckers of the hill and valley checkerboard; zero applied field.

on a valley checker, the value of z_c depends on the particular point chosen. .The reduction in the work function is less than $1200\pi l\rho$ and varies from point to point.

The next step is to ascertain the effect of an accelerating field on these P *vs.* z curves. Figure 13 shows this for $\beta = 1, \frac{1}{2}, \frac{1}{4}, -\frac{1}{4}, -\frac{1}{2}$ and -1, respectively, for $F = 5000$ volts/cm. The effect of the field is to decrease z_c and P_m. From Figs. 12 and 13 it follows that the decrease in z_c and in P_m due to F varies with β and is much larger for

points above a hill checker than for points above a valley checker. Values of z_c and P_m are shown in Fig. 14.

Figure 14A shows z_c at various values of x for $y = 0$; it also shows z_c at various values of x for $y = b/3$. Figure 14B shows P_m for these

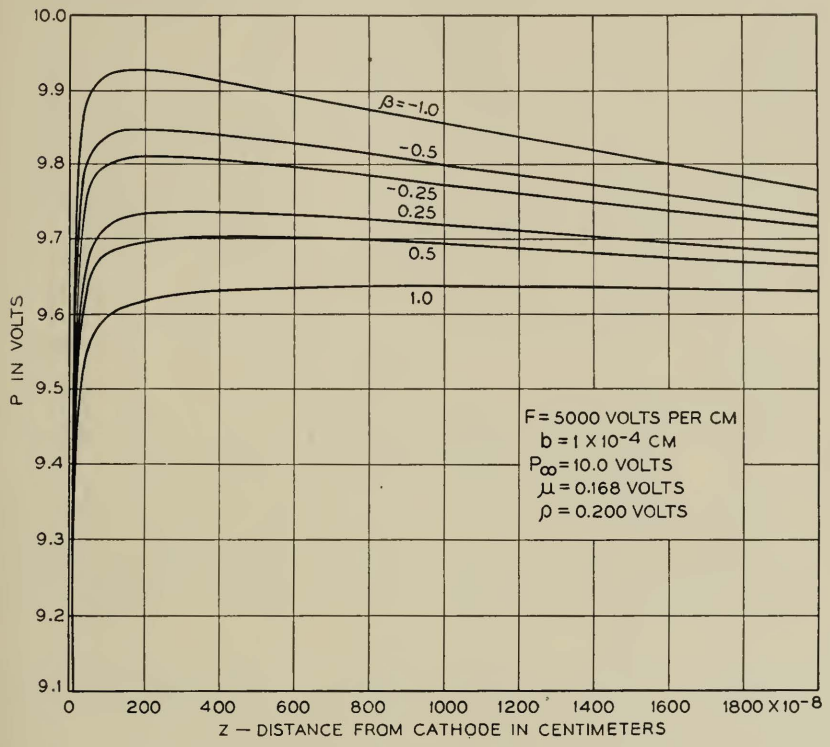


Fig. 13—Potential energy *vs.* distance from the surface above the center of the different subcheckers of hill and valley checkerboard; applied field of 5000 volts/cm.

same values of x and y. Since $P_m - K = \varphi e$, it is clear from these figures that different portions of the surface will have different work functions, or stated more precisely, the energy an electron must have to cross the critical surface (loci of the values of z_c) depends upon where it crosses the critical surface. This in turn means that the chance that a given electron will escape depends not only on its normal component of velocity but also on the place at which it leaves the surface and on the angle its path makes with the surface.

To compute accurately the emission current is a very difficult task. It would appear that the following procedure should give a good approximation to the true current. Divide a "hill" square and a

neighboring "valley" square into nine subsquares each, as indicated in Fig. 11C. It is apparent that the B squares are all alike; similarly the C, B' and C' squares are alike. Determine the β for the center of each subsquare. For the A, B, C, A', B' and C' subsquares the values of β

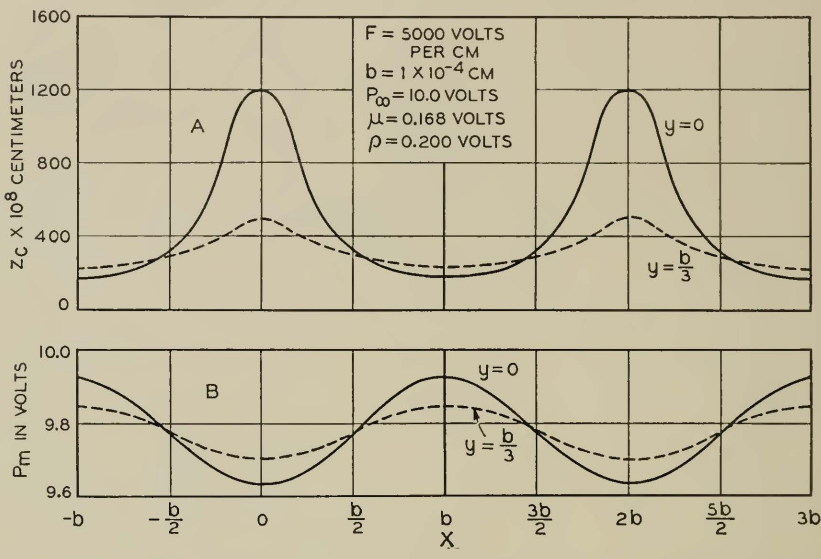


Fig. 14—A. Distance to critical plane for various points of hill and valley checkerboard. B. Potential energy at critical plane for various points of hill and valley checkerboard.

are, respectively, 1, $\frac{1}{2}$, $\frac{1}{4}$, -1, $-\frac{1}{2}$ and $-\frac{1}{4}$. Then compute the P *vs.* z curve for the normals at the center of each subsquare. These are the curves shown in Fig. 13. Next compute the current for each subsquare assuming this to be the same as it would be for an equal area of a large surface having the same work function as that for the center of the subsquare. This is equivalent to assuming that the effect of the velocity components in the x and y directions average out. Do this for various values of F. At each F, add up the currents for all 18 subsquares and multiply this by $1/2b^2$, the number of pairs of squares in a cm.^{2} This will give the current per cm.^{2} of surface for various values of F.

Figure 15 shows the values of the current for the various subsquares; more precisely it shows $\log i/i_{ao}$ *vs.* $\sqrt{F}$ where i_{ao} is the current that would be obtained from an equal area at zero field for a surface covered with a charge density ρ which is the average value of the charge density for the entire surface. i_{ao} is equal to the current from the hill or active subsquares at zero field, but the current from the valley

subsquares is usually less than i_{ao}. The reason for this is clear from an inspection of the curves in Fig. 12. Figure 15 also shows log i/i_{ao} for the average current for 18 subsquares in a pair of hill and valley squares; more precisely, i/i_{ao} is the sum of the current for the 18 subsquares divided by the current that would be obtained if the charge density

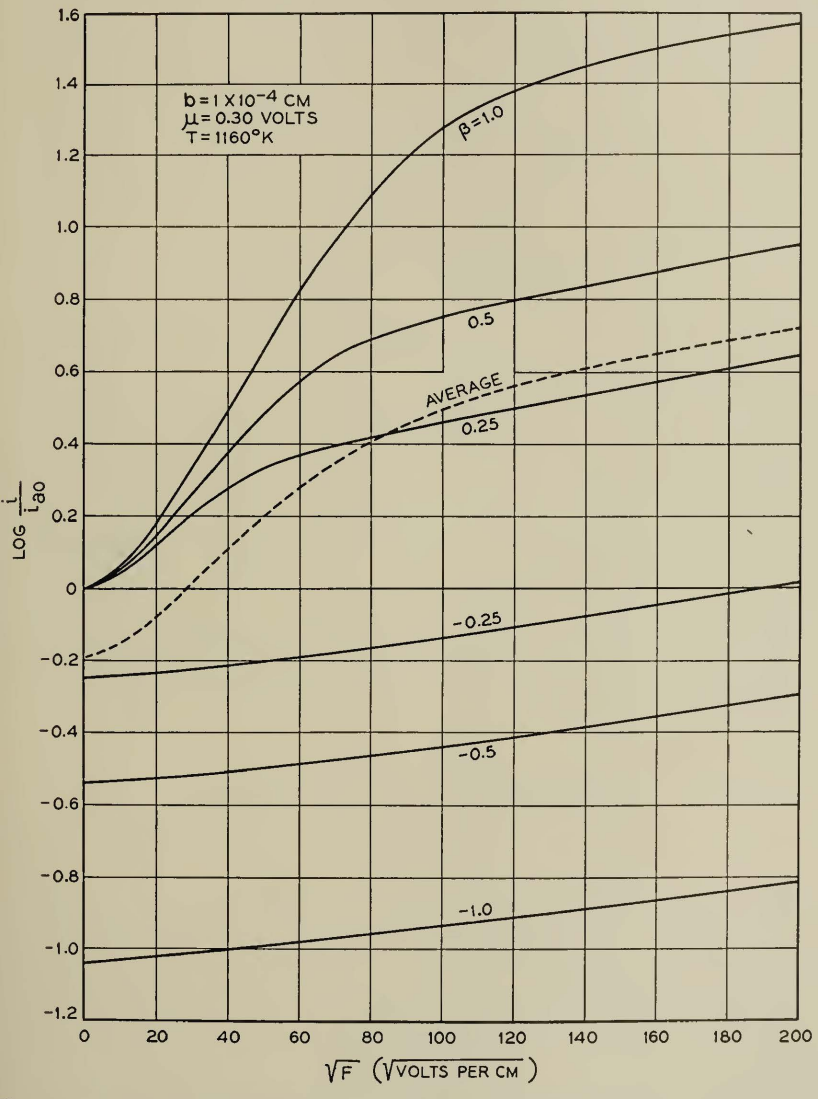


Fig. 15—Variation of emission current with applied field from different types of subcheckers for hill and valley checkerboard; also average curve.

were uniform and the applied field were zero. As the field is increased the *relative* contribution to the sum current from the central hill square becomes larger and larger, while that from the valley squares becomes less and less. At very large fields the curve for any subsquare approaches a straight line whose slope is the Schottky slope; hence the sum curve also approaches a straight line having this same slope. The area which is now contributing most of the current is, however, considerably less than the entire area. Roughly speaking, one might say that at low fields something more than half the area is "effective" in emitting electrons; as the field increases the "effective" area decreases; at large fields and high values of μ, 50 per cent of the total current comes from about 5 per cent of the total surface; one might say that the "effective" area is approximately twice as large as this or 10 per cent. The values of the "effective" areas depend on the value of μ: as μ increases the "effective" area decreases.

Such average curves as the one shown in Fig. 15 depend on three variables, b, μ and T. This dependence is illustrated in Figs. 16, 17 and 18; in each case two of the variables are kept constant. Figure 16 shows $\log i/i_{ao}$ *vs.* $\sqrt{F}$ for three values of b. This curve is similar to Fig. 9 for a single circular patch. All the curves still approach a Schottky line at high values of $\sqrt{F}$ but because of the averaging process they do not start out from a common value when $F = 0$ and the initial slope is not equal to the Schottky slope. In both figures as the size of the patch decreases, the curves get less steep and the place at which the curves bend over toward the upper Schottky line moves to higher values of F. Note also that beyond this bend, the curves are approximately straight but only approximately and that the values are still somewhat below the theoretical Schottky line.

This theoretical line has an intercept given by

$$\log i_S/i_{ao} = \log (1/18)[\exp \mu e/kT + \exp - \mu e/kT + 4(\exp \mu e/2kT + \exp - \mu e/2kT) + 4(\exp \mu e/4kT + \exp - \mu e/4kT)]. \quad (73)$$

This equation which defines i_S is based on a subdivision of the hill and valley checker into 18 subcheckers. The exponentials contain the value of β for each type of subchecker, in this case 1, $\frac{1}{2}$ and $\frac{1}{4}$. If each checker were divided into a larger number of subcheckers the number of terms would be increased, but fortunately the value of $\log i_S/i_{ao}$ would not be greatly affected. This is particularly true as long as $\mu e/kT$ is less than 5. Since $e/kT \sim 0.1$ this means that values of $\log i_S/i_{ao}$ are essentially correct for μ less than 0.5 volt. We have plotted $\log i_S/i_{ao}$

vs. $\mu e/kT$ for 9 subcheckers and for 25 subcheckers. For $\mu e/kT = 5$ the former curve is only 4 per cent higher than the latter; for $\mu e/kT = 10$ the difference is about 6 per cent; for $\mu e/kT$ less than 5 the difference is negligible. This makes us feel that our average curves which are

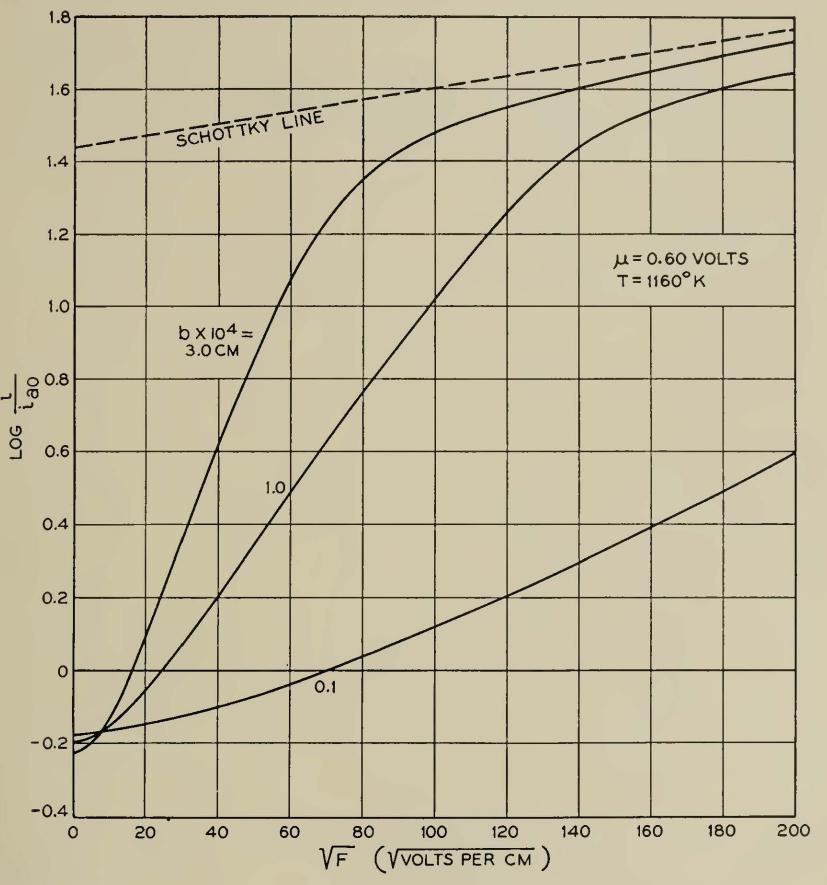


Fig. 16—Variation of average emission current with applied field for hill and valley checkerboard; b variable, T and μ fixed.

based on 18 subcheckers are essentially the same as would be obtained for a much larger number of subcheckers. Equation (73) is analogous to equation (69). Note that it depends on μ but not on b.

Figure 16 might have shown log i/i_S instead of log i/i_{ao}. To convert it to this coordinate it is merely necessary to reduce all values of the ordinate by log i_S/i_{ao} or by 1.44. The advantages of this ordinate will become apparent when we compare theoretical and experimental

curves. For the latter it is quite easy to obtain log i_S but not so easy to determine log i_{ao}.

In Fig. 17 we have thus shown log i/i_S *vs.* $\sqrt{F}$ for constant b and T but varying μ. Had log i/i_{ao} been plotted the curves would have been

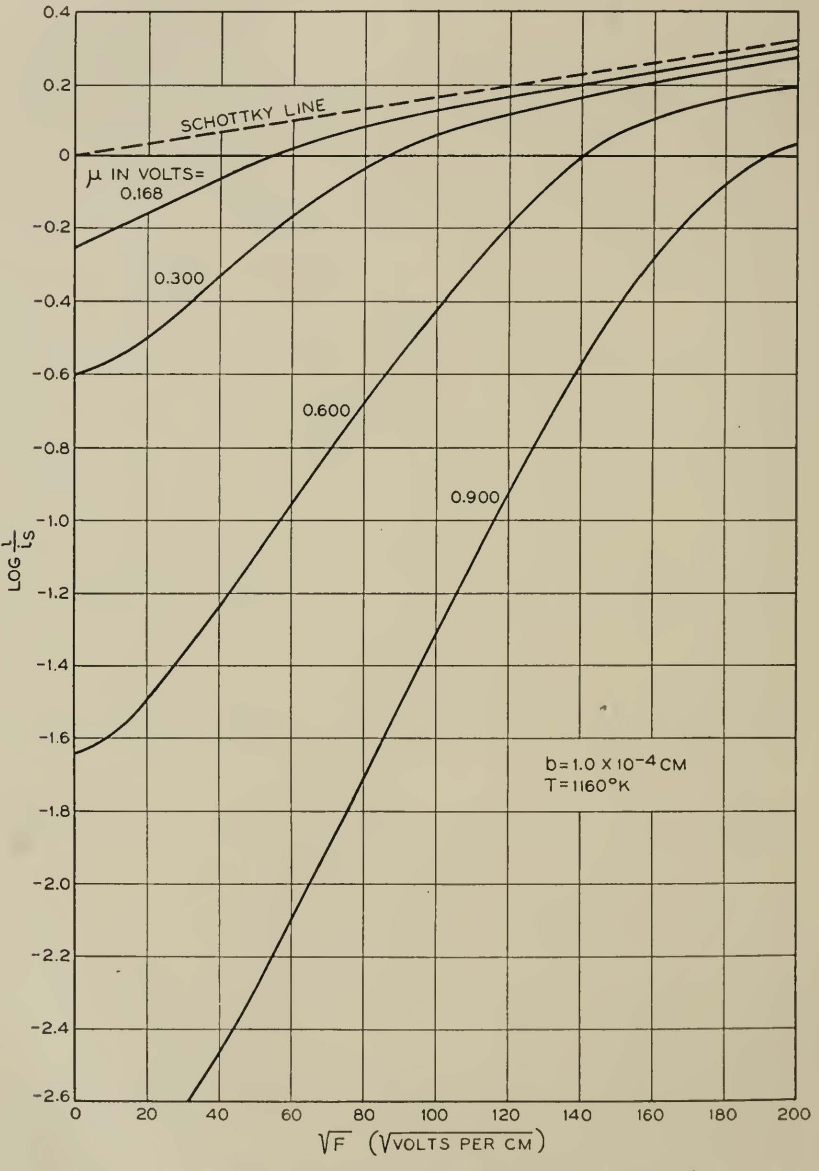


Fig. 17—Variation of average emission current with applied field for hill and valley checkerboard; μ variable, T and b fixed.

close together at $F = 0$ but would have approached upper Schottky lines whose position varied greatly. As it is, all curves approach the same upper Schottky line but "fan out" toward lower values of $\sqrt{F}$. Note that the curves do not cross over as they do in Fig. 16; note also that as μ increases the curves become steeper and the bend toward the Schottky line occurs at larger values of $\sqrt{F}$.

At first sight it might appear that by plotting $T \log i/i_S$ it would be possible to eliminate T as a parameter; while this is true for any one subchecker for which β is a constant, it is not true for the average curves. This is illustrated in Fig. 18 which shows $(kT/e) \log i/i_S$ *vs.*

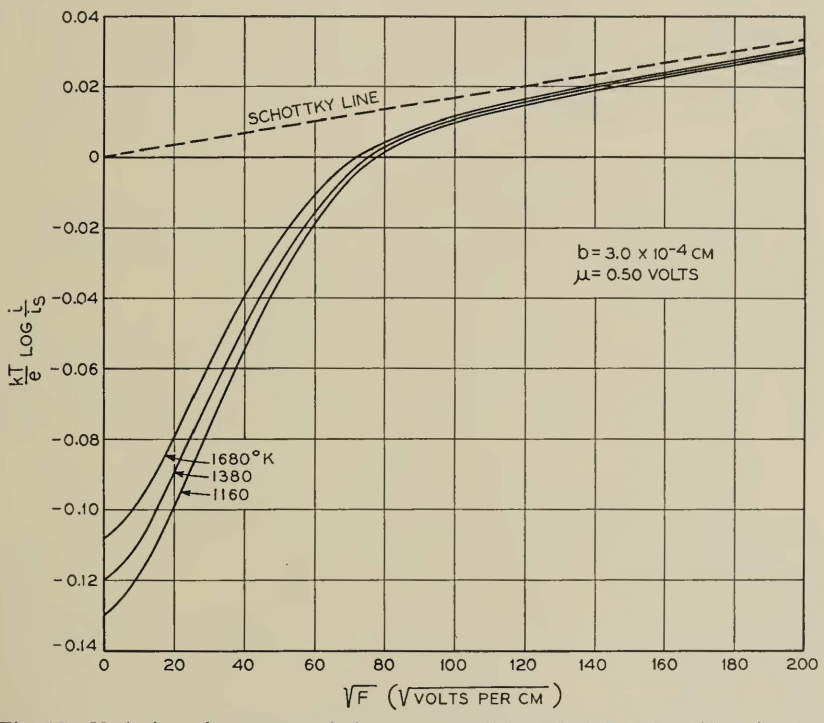


Fig. 18—Variation of average emission current with applied field for hill and valley checkerboard; T variable, b and μ fixed.

$\sqrt{F}$ for constant b and μ and varying T. Note that the departure from the Schottky law is more pronounced for the low temperatures.

Comparison Between Theory and Experiment

It has been found possible to choose values of b and μ such that the calculated average curve fits a given experimental curve over its entire range. The agreement is not perfect; but a perfect fit is not to be

expected when one considers that in an experimental filament the patches are not all of the same size and neighboring patches do not have the same differences in work function. One example is illustrated in Fig. 19 which shows (kT/e) log i/i_{ao} *vs.* $\sqrt{F}$ for a thoriated tungsten

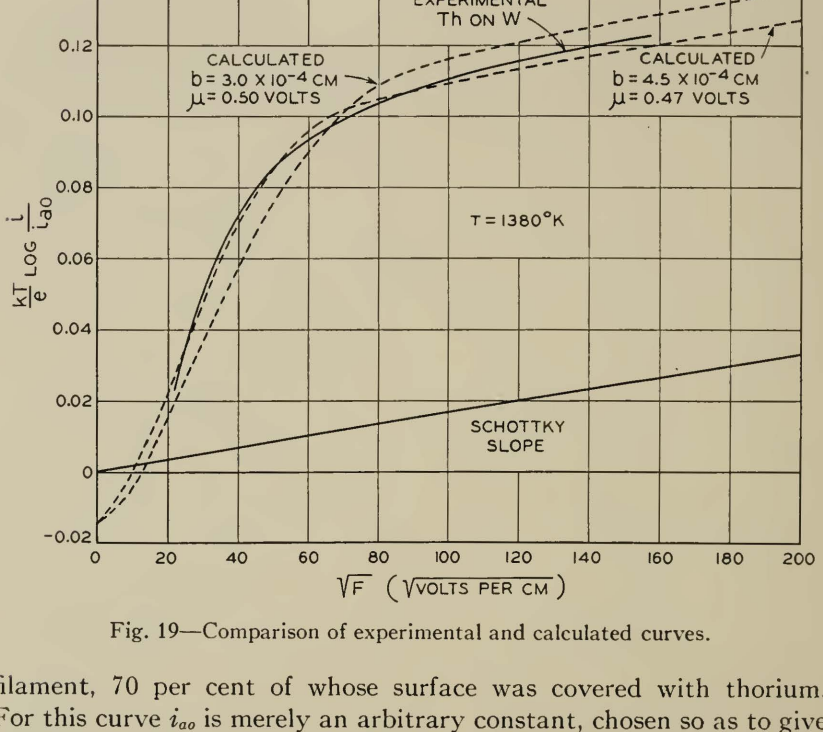


Fig. 19—Comparison of experimental and calculated curves.

filament, 70 per cent of whose surface was covered with thorium. For this curve i_{ao} is merely an arbitrary constant, chosen so as to give the best fit with the computed curves. By comparing the experimental curve with calculated curves in Figs. 16 and 17, we decided to try $b = 3.0 \times 10^{-4}$ cm. and $\mu = 0.50$ volt.* The average curve for these values is shown in the figure. It appeared that b was too small and μ was too large and a new curve was computed and plotted assuming $b = 4.5 \times 10^{-4}$ cm. and $\mu = 0.47$ volt. The agreement is better. Probably a slightly better fit could be obtained with $b = 4.0 \times 10^{-4}$ cm. and $\mu = 0.48$ volt. For other thoriated tungsten filaments we have found values of b from 1×10^{-4} to 1×10^{-3} and μ values from 0.25 to 0.48 volt.

* In choosing these values some allowance had to be made for the fact that the experimental curve was for $T = 1380°$ K., while the calculated curves were for $T = 1160°$ K.

An examination of a number of thoriated tungsten filaments with a microscope showed that the diameter of the tungsten crystals was of the same order as the values of b given above, namely 10^{-4} to 10^{-3} cm. These filaments had been given the customary heat treatment in a vacuum at temperatures near 2800° K. for times measured in minutes and at temperatures near 2100° K. for many hours. It was natural, therefore, to form the hypothesis that different crystals have different adsorptive properties and that consequently different crystals in the same filament should be covered with varying amounts of thorium. Different crystals will then have different work functions. The values of μ found by the above analysis are consistent with this hypothesis since 2μ, which is the difference in work function between a hill checker and a valley checker, is always considerably less than 2.0 volts which is the maximum difference in work function between clean tungsten and thorium on tungsten.

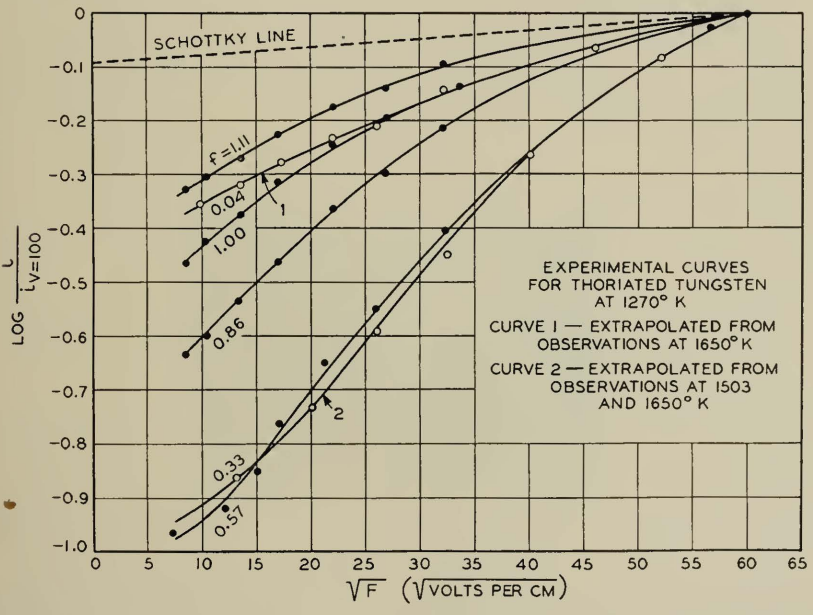


Fig. 20—Experimental log i *vs.* $\sqrt{F}$ curves for thoriated tungsten for various values of f at 1270° K. f = fraction of surface covered with thorium.

If this hypothesis is true, then as a thoriated tungsten filament is activated, the curves for the various stages of activation should all correspond to approximately the same value of b. Figure 20 shows such a family of experimental curves taken by W. H. Brattain of these laboratories. Curves 3, 4, 5 and 6 were taken at T = 1270° K.; curve

1 is extrapolated from data at $T = 1650°$ K.; curve 2 is extrapolated from data at 1503 and 1650° K. The extrapolations were made by extending Richardson lines; curves 1 and 2 are, of course, not quite as certain as data taken at 1270° K. The currents at any V or F vary by large factors as f, the fraction of the surface covered with thorium varies. In order to compare the curves more effectively log $i/i_{V=100}$ has been plotted. This is equivalent to shifting the log i curves until they pass through a common point at $V = 100$ or $\sqrt{F} = 60.5$. A comparison of this family of curves with the computed curves in Figs. 16 and 17 shows a great similarity with Fig. 17 but not with Fig. 16. In Fig. 17, b was constant while μ was varied. From this similarity it follows that the experimental curves in Fig. 20 are consistent with an approximately constant value of b but varying μ.

From the position and shapes of the curves in Fig. 20, we estimate that the value of b or the crystal size is about 4×10^{-4} cm. This value is probably too large for curve 1 and too small for curve 6 but unless a complete analysis were made, it is not desirable to discuss small variations in b. It is apparent from the figure that μ changes with f. We have estimated the following values of μ, the second significant figure being in doubt:

TABLE III

VALUES OF f, THE FRACTION OF SURFACE COVERED WITH THORIUM AND VALUES OF μ IN VOLTS (EQ. 71) μ IN VOLTS = $1200\pi\mu l$.

f =	0.04	0.33	0.57	0.86	1.0	1.11
μ(volts) =	0.23	0.44	0.45	0.36	0.28	0.23

These values of μ are reasonable. Furthermore, the way in which μ varies with f is to be expected from the shape of the work function *vs.* f curve which will be discussed later under adsorption.

A particularly interesting test of the patch theory is furnished by Taylor and Langmuir's [25] electron emission from cesium on tungsten because in this case the crystal size of the tungsten is known. In Figs. 11, 12 and 13 of their article they give log i *vs.* V or $\sqrt{V}$ curves. Since the diameter of the filament is given as 2 mils, it is possible to convert values of V to values of F and to obtain log i *vs.* $\sqrt{F}$ curves. We have done this for the curve for $\theta = 0.60$ and have then analyzed it on the basis of the hill and valley theory. This analysis gave $b = 0.8 \times 10^{-3}$ cm. and $\mu = 0.20$ volt. The article states * "the average grain size in these filaments was about one-fifth the diameter of the wire." So that the average grain size was about 1×10^{-3} cm. which is about the same as the calculated value of b.

* Taylor and Langmuir, reference 25, page 431.

Another striking confirmation of the hypothesis that various crystals of a filament have different work functions and thus emit electrons with greatly varying intensities, is given by pictures of such filaments obtained by means of electron optics.* Figure 21 shows an electron

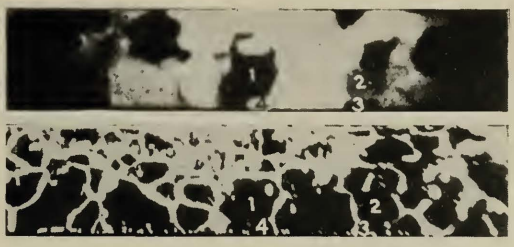


Fig. 21—Electron-micrograph (above) and photo-micrograph (below) of platinum ribbon.

and photo-micrograph † of a portion of a platinum filament. Corresponding crystals have been labeled by 1, 2, 3 and 4. The reader can find more cases of correspondence. Of course, a perfect correspondence is not to be expected since two neighboring crystals may have the same reflection properties for light while the electron emissivities differ and *vice versa*.

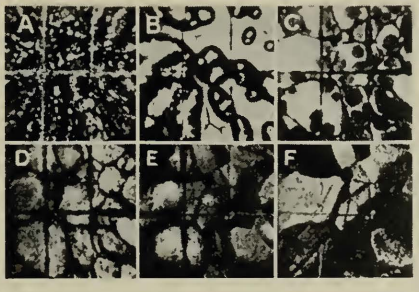


Fig. 22‡—Electron-optical pictures at various stages of heat treatment for a nickel surface coated with oxide.

Figure 22 shows a series of electron-optical pictures taken by W. Knecht [26] of part of a nickel surface covered with BaO and SrO. The

* For an interesting and instructive account of the technique and applications, see the book: *Geometrische Elektronenoptik* by E. Brüche and O. Scherzer, published by J. Springer (Berlin, 1934).

† I am gratefully indebted to Dr. C. J. Davisson and Mr. C. J. Calbick of these laboratories for this figure.

‡ We take this opportunity to thank the publishers of the Annalen der Physik for permission to reproduce this figure from Knecht's article in Annalen der Physik 20, 180 (1934).

original magnification is 27-fold. Several lines were scratched on the surface. The filament was mounted in a vacuum tube and activated. Picture "a" shows the electron emission from the various granules of the oxide. At this stage the cathode had the characteristic appearance of ordinary oxide coated filaments. The filament was then flashed at a comparatively high temperature for successive intervals of time. After each interval another electron picture was taken. As a result of this treatment the oxide evaporated so that the filament had a metallic appearance. However, there is good reason for believing that metallic barium had been alloyed with the nickel, and the emission was much greater than that from clean nickel. The pictures show that when the oxide has disappeared, different areas emit electrons with greatly different intensities. The shapes and sizes of these areas are strikingly similar to those obtained in ordinary optical pictures. In fact Knecht states that the pattern is that of the nickel crystals.

Numerous other pictures similar to Fig. 22 can be found in the book by Brüche and Scherzer referred to above. Some of these indicate that there is a fine structure non-uniformity inside of a single crystal as well as the non-uniformity between crystals.

The emission from thoriated tungsten has also been investigated by electron optics and while the pictures are not as striking as those for barium on nickel, they prove rather conclusively that the emission varies from one crystal to the next or from one region of a crystal to the next. Finally the emission from surfaces to which no impurity has purposely been added have been investigated. These too show patchy emissions.

It has generally been felt and frequently stated that the emission from clean surfaces, particularly clean tungsten, varied with applied field according to the Schottky equation. While this is true at moderate and high fields, I have never seen data which showed agreement at low fields, but have seen data which showed disagreement. The deviation is less pronounced than it is for thoriated tungsten but in my opinion it is none the less real. This I have interpreted to mean that even in a polycrystalline surface of clean tungsten, individual crystals have work functions which vary by one-tenth or a few tenths of a volt. This interpretation receives strong support from the work of Farnsworth and Rose.[27] They showed that a (111) surface of a single crystal of copper had a contact potential with respect to a (100) surface equal to 0.463 volt. The direction was such that the work function of the (111) surface was 0.463 volt less than that of the (100) surface. Even after heating the crystals to about 900° C. for as much as 1000 hours, the contact potential difference was still 0.378 volt and had

remained practically constant during the last 700 hours. This decrease is ascribed to the formation of new crystal facets, which produce more nearly equal work functions for the two surfaces. That different crystal planes have different work functions also follows from the photoelectric work of Nitzsche [28] on single crystals of zinc. The work function for a surface normal to the hexagonal axis was found to be 3.28 volts while that for a surface parallel to this axis was 3.09 volts. These observed differences in work function of different surfaces of single crystals are quite large enough to account for the deviations from the Schottky law for clean surfaces.

Still another prediction of the patch theory is verified by experiment. In connection with Fig. 15 it was pointed out that at low applied fields something more than half the area is effective in emitting electrons; as the field is increased the effective area decreases until the log i *vs.* $\sqrt{F}$ curves approach the Schottky line when the effective area attains a constant value whose order of magnitude is 0.1. From this it follows that if Richardson lines are obtained for a series of applied potentials, the intercepts or values of log A should decrease as V increases, but should approach a constant value for sufficiently large values of V. Experimental values of log A *vs.* V are given in Fig. 14 of Brattain and Becker's [29] article on thorium on tungsten. They show the predicted trend. Furthermore, the change in log A with V should be most pronounced for large values of μ. It was shown above that the largest values of μ occur in the neighborhood of $f = 0.6$ while near $f = 1.0$, μ is comparatively small. The experimental curves show the largest dependence of log A on V for $f = 0.6$ and only a small dependence for $f = 1$. In this respect too, experiment confirms the theory.

Non-uniformities on the cathode affect the shape of the retarding potential curves as was explained in a previous section. Here too, there is at least qualitative agreement between theory and experiment.

In connection with the analysis of log i *vs.* $\sqrt{F}$ curves we have, in the course of the last five or six years, developed a number of simple methods for computing approximate values of b and μ. We have also proved a number of useful theorems. If and when the interest in this subject warrants it, we intend to publish these methods and theorems.

Checkerboard with Uniform Charge Distribution

While the agreement between experimental log i *vs.* $\sqrt{F}$ curves with theoretical curves based on a hill and valley charge distribution over a checkerboard array is quite good, it is probable that even better

agreement could be obtained if the charge density over the black squares was assumed to be uniform and equal to $\rho + \bar{\mu}$ while over the white squares it was assumed to be uniform and equal to $\rho - \bar{\mu}$. In terms of the previous notation, $\beta = +1$ for all points of the black squares while $\beta = -1$ for all points of the white squares.

If we use the coordinates indicated in Fig. 11A, such a charge distribution can be represented by the following double Fourier series

$$\sigma = \rho + \frac{16\bar{\mu}}{\pi^2} \sum_N \frac{(-1)^{(N-1)/2}}{N} \cos\frac{\pi N x}{b} \sum_K \frac{(-1)^{(K-1)/2}}{K} \cos\frac{\pi K y}{b}, \quad (74)$$

in which N takes on all values, 1, 3, 5, 7, etc., and for each N, K takes on all the values 1, 3, 5, 7, etc. If such a charge distribution is located at a distance l above the surface while its image is located at a distance l below the surface, then the potential energy of an electron due to this double layer is given by

$$P_\sigma/e = -300 \times 4\pi\rho l - \frac{300 \times 65\bar{\mu}l}{\pi} \sum_N \sum_K \frac{(-1)^{(N+K)/2}}{NK} \times \exp\left(\frac{-\pi(N^2+K^2)^{\frac{1}{2}}}{b} z\right) \cos\frac{N\pi x}{b} \cos\frac{K\pi y}{b}. \quad (75)$$

This formula is accurate provided $l/b \ll 1$, which is always fulfilled in any case in which one is likely to be interested. The electric field normal to the surface due to the double layer is

$$\frac{1}{e}\frac{dP_\sigma}{dz} = \frac{300 \times 64\bar{\mu}l}{b} \sum_N \sum_K \frac{(-1)^{(N+K)/2}(N^2+K^2)^{\frac{1}{2}}}{NK} \times \exp\left(\frac{-\pi(N^2+K^2)^{\frac{1}{2}}}{b} z\right) \cos\frac{\pi N}{b} x \cos\frac{\pi K}{b} y. \quad (76)$$

Equations (74), (75) and (76) reduce to the corresponding equations for the hill and valley distribution if $16\bar{\mu}/\pi^2$ is replaced by μ and if only the first term of the double series is used, i.e., if $N = 1$ and $K = 1$. See equations (70) and (71).

Recently Mr. Albert Rose working with Professor L. P. Smith at Cornell University has made calculations for a checkerboard with uniform charge distribution and has compared his computations with experiment. The agreement is as good as we have found and his computed values of b and μ are about the same as ours.

Linford [6] in an excellent review on the external photoelectric effect has shown that a checkerboard distribution of charge or potential satisfactorily accounts for a number of photoelectric phenomena observed with composite surfaces. His equation (42) is almost

identical with equation (75) above if his $V_0 = 8\pi\bar{\mu}l$, his $2j + 1 = N$ and his $2k + 1 = K$. His equation is not quite as general as ours since he deals only with the case for which $\rho = \bar{\mu}$.

Compton and Langmuir * in 1930 presented an interesting discussion of the poor saturation in composite surfaces. They, too, proposed a checkerboard or patch distribution like the one we are discussing, and on page 151 of their paper they give an equation for the potential above such a surface. Unfortunately there is an error in this equation which was pointed out by Linford.[6] They use only a single summation whereas a checkerboard distribution requires a double summation. However, since they use only the first term of their summation and since this first term is the same as the first term of the correct equation, this error is not serious in their case. Their formula, too, is less general than equation (75) since they deal only with the case $\rho = \bar{\mu}$; this is equivalent to assuming that the white squares are clean tungsten and only the black squares are covered.

They reject their patch theory because (1) "to obtain departures from the Schottky curve comparable to those observed, the patches must be assumed to contain many thousands of atoms," and (2) "the patch theory predicts a departure from the Schottky curve which is small with small fields and increases with large fields, whereas exactly the reverse is the actual case." † From what has been said above it is clear that their second objection is really tied up with their first one, for if larger patch sizes are assumed their statement is incorrect and quite good agreement is found with experiment. They assumed a value of $b = 10^{-6}$ cm. whereas the experimental curves require $b \sim 10^{-4}$ cm. They feel that such "extremely non-uniform distributions" or "such large clusters of adsorbed atoms" are "very improbable." One reason for this belief is that "Becker has shown, for example, that a thorium layer at emission temperatures behaves like a two-dimensional gas on the surface."

In my opinion these objections to the checkerboard or patch theory are not well founded. It seems quite natural to me that various crystals on the surface or various crystal facets may have somewhat different adsorptive properties and that consequently different crystals would be covered to different extents with thorium and would thus have different work functions. It is probable that the size of the squares should be comparable to the crystal size which is of the order of 10^{-4} cm. This is still true if thorium migrates over the surface of the tungsten. The successes of the patch theory presented above far outweigh these objections.

* Reference 1, especially pp. 146–160.
† Reference 1, p. 157.

More than this, some of the very data presented by Compton and Langmuir support the generalized checkerboard theory as we have presented it, i.e., taking into account that both the black and the white squares may be covered with thorium but to different extents. On page 155, Compton and Langmuir discuss two log i *vs.* $\sqrt{F}$ curves obtained by Reynolds[30] for thoriated tungsten. Curve A in their Fig. 4 is a "normal" curve while curve C is taken after the surface has been bombarded by positive ions. They state "this bombardment must have roughened the surface and there is evidence that it also fractured the surface layer of tungsten crystals." In our notation this means that b has been decreased because of the roughening and μ has been increased because the amount of thorium removed in some spots was larger than that removed in others. Now the decrease in b should shift the region at which the curve approaches the Schottky line to higher values of F or $\sqrt{F}$; while the increase in μ should result in a steeper curve and should decrease log i_{ao}. (See our Fig. 16.) But this is precisely what curve C does.

Reynolds[30] in discussing this same data says: "The effect of bombardment was a semi-permanent one. Subsequent activation and deactivation by temperature (below 2700°) shifted the curve along the current axis but did not otherwise alter its unique character. Flashing at 2700° K. or higher, where rapid sintering of tungsten is known to take place, destroyed the effect of bombardment and subsequent activation produced normal log i *vs.* $V^{\frac{1}{2}}$ curves." Every detail of this behavior is just what is to be expected on our view; the "semi-permanent" effect is caused by the decrease in b which does not become normal until the damage to the crystals has been repaired by high temperature treatment; the shifting of the curves along the current axis is caused by changes in ρ and μ brought about by activation and deactivation.

On page 156 Compton and Langmuir[1] discuss Kingdon and Langmuir's data for thoriated tungsten at various degrees of activation (θ or f) and at various temperatures. Some of the results are shown in their Fig. 5 which is a plot of T log i *vs.* $\sqrt{F}$. They point out three and only three distinctive features of these curves. All three support the checkerboard theory. They say: "At the highest field-strengths (about 10,000 volts/cm.) the curves are seen to approach the theoretical slope." Our analysis shows that this means that all surfaces have about the same b irrespective of θ and T. This is predicted by our theory since b is determined by the crystal size which is independent of f and T. In discussing curves for a constant f they say: "In every case the departures from the Schottky line become greater as the

temperature is lowered, —." The theory predicts this as shown by Fig. 18. There may, however, be another reason: As T increases μ may decrease. If there are differences in concentration between neighboring crystals, and if the temperature is high enough for migration to occur, Boltzmann's law would require that the difference in concentration should decrease as T increases.

About the third feature they say: "These results indicate that with nearly complete thoriation of the surface ($f = 0.91$) and with a bare surface ($f = 0.00$) the approach to the Schottky curve is fairly close, but relatively large departures occur with incomplete thoriation." This fact which is abundantly confirmed by my experience not only with thorium on tungsten but also with cesium on tungsten, cesium on oxygen on tungsten, and barium on tungsten means that as f increases, μ increases at first, rises to a maximum and then decreases. Such a variation of μ with f is to be expected from the shape of the log i *vs.* f or φ *vs.* f curve which will be discussed more fully later on. As f increases, φ decreases rapidly at first, then more and more slowly until it passes through a minimum when $f = 1$; beyond this point φ increases again. It is natural to expect that Δf, the difference between f for the black and the white squares, should increase with f; Δf is probably nearly proportional to f. From this and the shape of the $\varphi - f$ curve it follows that $\Delta\varphi$, the difference in φ between black and white squares, is small when f is small; as f increases $\Delta\varphi$ increases at first but later on it decreases; when f approaches 1.0, $\Delta\varphi$ approaches 0 and the surface has a uniform work function. Since μ and $\Delta\varphi$ are proportional, μ should vary in the same way. Hence this feature of Compton and Langmuir's curves as well as the first two is entirely in agreement with the predictions based on the checkerboard theory.

Whether the uniform charge distribution or the hill and valley distribution gives better agreement with experiment has not been decided. This, however, is not very important or very pressing. In an experimental filament the distribution is probably neither one nor the other but something in between. Furthermore, it should be emphasized that in an experimental cathode the patches are not all of the same size nor is the contact potential between two neighboring patches a constant; both of these quantities fluctuate about a mean value. Nevertheless I believe that a sufficiently good case has been made out to show that non-uniformities play an important role in many thermionic experiments, and that the checkerboard theory can be used as a powerful tool in the study of adsorption phenomena, where non-uniformities almost always occur.

This analysis of the effect of non-uniformities has brought out that

the work function is not a characteristic of a given substance but rather of a given surface of a given substance. Strictly speaking, one should not talk about the work function of tungsten but rather of the work function of a particular surface of tungsten. This is true even if the surface is clean tungsten.

The Values of the Work Function for Clean Surfaces

The experimental determination of the thermionic work function or the heat function for clean metal surfaces has been the subject of many investigations. In the case of a number of elements, the determinations by different investigators are not in accord. This is due, in most cases, to adsorbed layers of foreign material caused by either poor vacuum conditions or impurities in the metal which have not been eliminated by a proper heat treatment. Although these measurements have been summarized and discussed in other reviews, it seems advisable that the summary be brought up to date and the most probable values selected from the existing data. Since the photoelectric work function is equal to the thermionic work function,[8] the determination by photoelectric methods should also be included.

A summary of the data is shown in Table IV. The values of the photoelectric work function and the thermionic heat function are expressed in volts. The reference for each value is indicated by the superscript. As discussed in an earlier section, the heat function is the slope of a Richardson line. The photoelectric work functions are mostly calculated from the long wave-length limit except in the case of recent determinations which are made by an analysis of the data by Fowler's [31] method. The photoelectric values listed in the table were selected as representative of values for the best outgassing of each element. For a listing of all determinations see Hughes and DuBridge's book. In most cases, the heat function and the thermionic work function differ only by small amounts so that for practical purposes we can compare the photoelectric work function with the heat function. The most probable values of the heat functions tabulated have been chosen from the several determinations.

Recently several attempts have been made to find an empirical relation between the work function and the atomic properties of the elements. Such a correlation, if applicable to all of the metallic elements, would be of value in predicting values of the work function for the cases in which the existing data are inadequate or no data are available. The work of Rother and Bomke [32] gives the best correlation thus far obtained. In their article they have given a summary of the early attempts at a correlation and therefore we will not consider them here.

TABLE IV

Compilation of Values of Photoelectric and Thermionic Work Functions in Volts and the Value of the Heat Function

Element	Photoelectric work function	Thermionic heat function	Probable value of heat function
Ag	(4.58 to 4.75)[61] (4.71 to 4.75)[22]	(4.08)[25]	4.7
Al	(2.99)[39]		3.0
Au	(4.75 to 4.84)[42] (4.86 to 4.92)[22]	(4.32)[25]	4.8
Bi	(4.05)[61] (3.74)[7] (4.37)[43]		4.1
C	(4.72)[28] (4.81)[48]	(3.93)[26]	4.7
Ca	(2.76)[60] (3.20)[45]	(3.02)[29] (2.24)[19]	3.2
Cb		(3.96)[67]	4.0
Cd	(4.07)[2]		4.1
Ce	(2.84)[46]		2.8
Co	(4.12 and 4.25)[4]		4.2
Cr	(4.60)[35]		4.6
Cs	(1.67)[50]	(1.81)[34]	1.8
Cu	(4.49)[61] (4.08)[39]	(3.85)[58] (4.33)[25]	4.1
Fe	(4.72)[3] (4.77)[24]	(4.04)[33] (4.77)[52] (4.04)[11]	4.7
Ge	(4.85)[50]		4.9
Hf		(3.53)[64]	3.5
Hg	(4.53)[18, 27, 47]		4.5
K	(1.77)[31] (2.0)[32]		1.8
Li	(2.21)[50]		2.2
Mg	(approx. 2.43)[6]		2.4
Mo	(4.15)[16]	(4.08)[37] (4.39)[55] (4.59)[38] (4.44)[20] (4.38)[63] (3.48)[40] (4.14–4.17)[16] (4.32)[1]	4.3
Na	(1.80)[41] (2.25)[44, 53] (1.94)[50]		1.9
Ni	(5.01)[24]	(4.41)[33] (2.77)[51] (4.31)[52] (4.63)[11] (5.03)[23] (4.7)[8]	5.0
Os			
Pb	(3.50)[44] (3.97)[30] (4.14)[28] (3.97)[39]		4.0
Pd	(4.97)[17]	(4.99)[17]	4.98
Pt	(6.30)[14]	(6.27)[15] (5.40)[56] (5.93)[49]*	6.0
Rb	(1.82)[50]		1.8
Re	(approx. 5.0)[21]		5.0
Rh	(4.95 to 4.57)[12]	(4.58)[12]	4.6
Sb	(4.02)[30]		4.0
Se	(4.62)[28]		4.6
Sn	(β4.50; γ4.38; liq. 4.24)[26] (β4.39)[22]		4.4
Sr	(2.06)[13]		2.1
Ta	(4.10–4.14)[22] (4.12–4.19)[5] (4.12)[45]	(4.2)[36] (4.51)[33] (4.18)[54] (4.07)[20] (4.04)[46]	4.10
Th	(3.34)[44] (3.57)[26] (3.38)[45]	(3.35)[64]	3.4
U	(3.63)[45]		3.6
W	(4.69 and 4.54)[59] (4.60)[45]	(4.52)[9] (4.53)[11]*	4.52
Zn	(3.61)[28] (3.68)[48] (3.08)[39] (3.32 and 3.57)[10]		3.3
Zr	(3.73)[45]	(4.13)[64]	4.1

* For complete listing see S. Dushman, *Rev. Mod. Phys.* **2**, 381 (1930).

[1] A. J. Ahearn, Phys. Rev. **44**, 277 (1933).
[2] H. Bomke, Ann. d. Physik **10**, 579 (1931).
[3] A. B. Cardwell, Proc. Nat. Acad. Sci. **14**, 439 (1928).
[4] A. B. Cardwell, Phys. Rev. **38**, 2033 (1931).
[5] A. B. Cardwell, Phys. Rev. **38**, 2041 (1931).
[6] R. J. Cashman and W. S. Huxford, Phys. Rev. **43**, 811 (1933).
[7] Chien Cha, Phil. Mag. **49**, 262 (1925).
[8] Cooke and Richardson, Phil. Mag. **25**, 624 (1913); **26**, 472 (1913).
[9] C. J. Davisson and L. H. Germer, Phys. Rev. **20**, 300 (1922); **24**, 666 (1924).
[10] J. H. Dillon, Phys. Rev. **38**, 408 (1931).
[11] W. Distler and G. Monch, Zeits. f. Physik **84**, 271 (1933).
[12] E. H. Dixon, Phys. Rev. **37**, 60 (1931).
[13] R. Dopel, Zeits. f. Physik **33**, 237 (1925).
[14] L. A. DuBridge, Proc. Nat. Acad. Sci. **12**, 162 (1926).
[15] L. A. DuBridge, Phys. Rev. **32**, 961 (1928).
[16] L. A. DuBridge and W. W. Roehr, Phys. Rev. **39**, 99 (1932).
[17] L. A. DuBridge and W. W. Roehr, Phys. Rev. **42**, 52 (1932).
[18] H. K. Dunn, Phys. Rev. **29**, 693 (1927).
[19] S. Dushman, Phys. Rev. **21**, 623 (1923).
[20] Dushman, Rowe, Ewald and Kidner, Phys. Rev. **25**, 338 (1925).
[21] A. Engelmann, Ann. d. Physik **17**, 185 (1933).
[22] R. H. Fowler, Phys. Rev. **38**, 45 (1931).
[23] G. W. Fox and R. M. Bowie, Phys. Rev. **44**, 345 (1933).
[24] G. N. Glasoe, Phys. Rev. **38**, 1490 (1931).
[25] A. Goetz, Physik. Zeits. **24**, 377 (1923); **26**, 206 (1925); Zeits. f. Physik **42**, 329 (1927); **43**, 531 (1927).
[26] A. Goetz, Phys. Rev. **33**, 373 (1929).
[27] W. B. Hales, Phys. Rev. **32**, 950 (1928).
[28] R. Hamer, J. Opt. Soc. Am. **9**, 251 (1924).
[29] F. Horton, Phil. Trans. **A207**, 149 (1907).
[30] A. L. Hughes, Phil. Trans. **A212**, 205 (1912).
[31] H. E. Ives and A. L. Johnsrud, Astrophys. J. **60**, 231 1924).
[32] H. E. Ives, Jour. Opt. Soc. Am. **8**, 551 (1924).
[33] Jentzsch, Ann. d. Physik **27**, 148 (1908).
[34] K. H. Kingdon, Phys. Rev. **25**, 892 (1925).
[35] Kösters, Zeits. f. Physik **66**, 825 (1930).
[36] I. Langmuir, Phys. Rev. **2**, 450 (1913); Physik. Zeits. **15**, 516 (1914).
[37] I. Langmuir, Trans. Am. Electrochem. Soc. **29**, 125 (1916).
[38] H. Lester, Phil. Mag. **31**, 197 (1916).
[39] Lukirsky and Prilezaev, Zeits. f. Physik **49**, 236 (1928).
[40] M. J. Martin, Phys. Rev. **33**, 991 (1929).
[41] R. A. Millikan, Phys. Rev. **7**, 355 (1916).
[42] L. W. Morris, Phys. Rev. **37**, 1263 (1931).
[43] T. J. Parmley, Phys. Rev. **30**, 656 (1927).
[44] Pohl and Pringsheim, Verh. d. Phys. Grs. 1911–14.
[45] Rentschler, Henry and Smith, Rev. Sci. Inst. **3**, 794 (1932).
[46] N. B. Reynolds, Phys. Rev. **35**, 158 (1930).
[47] Roller, Phys. Rev. **36**, 738 (1930).
[48] S. C. Roy, Proc. Roy. Soc. **A112**, 599 (1926).
[49] R. W. Sears, Unpublished work.
[50] E. F. Seiler, Astrophys. Jour. **52**, 129 (1920).
[51] W. Schlichter, Ann. d. Physik **47**, 573 (1915).
[52] G. Siljeholm, Ann. d. Physik **10**, 178 (1931).
[53] W. Souder, Phys. Rev. **8**, 310 (1916).
[54] H. J. Spanner, Ann. d. Physik **75**, 609 (1924).
[55] E. R. Stoekle, Phys. Rev. **8**, 534 (1916).
[56] H. L. Van Velzer, Phys. Rev. **44**, 831 (1933).
[57] H. B. Wahlin and L. O. Sordahl, Phys. Rev. **45**, 886 (1934).
[58] Wehnelt and Seiliger, Zeits. f. Physik **38**, 443 (1926).
[59] A. H. Warner, Phys. Rev. **38**, 1871 (1931).
[60] G. B. Welch, Phys. Rev. **31**, 709 (1928).
[61] S. Werner, Upsala Univ., 1914.
[62] R. P. Winch, Phys. Rev. **37**, 1269 (1931).
[63] C. Zwikker, Proc. Amst. Acad. Sci. **29**, 792 (1926).
[64] C. Zwikker, Physik. Zeits. **30**, 578 (1929).

In the preceding sections the thermionic work function W was shown to be equal to $P - K$ where P is the difference in potential energy between an electron at rest inside and outside of the metal and K is given by equation (16). If we assume that there is one free electron per atom in the metal for all elements, then

$$K/e = 25.9(D/M)^{\frac{1}{3}}, \tag{77}$$

where D is the density of the metal and M the atomic weight.

From values of K/e given by equation (77) and experimental values of W/e, Rother and Bomke calculated P/e for a number of elements. Their values of P/e were said to be in accord with the empirical equations

$$P/e = 12.6(Dz/M)^{\frac{1}{3}} \text{ for some elements} \tag{78}$$

and

$$P/e = 16.3(Dz/M)^{\frac{1}{3}} \text{ for all other elements,} \tag{79}$$

where z is the maximum chemical valence of the element.

We have computed values of K/e from equation (77) and with the most probable values of W/e from Table IV have determined the probable values of P/e. Since the work function and the heat function differ by only small amounts, it is justifiable to use the heat functions for W/e. To test equations (78) and (79) we have plotted log P/e *vs.*, log (Dz/M) in Fig. 23. According to equations (78) and (79), the points should fall on two straight lines in this plot. The two lines are shown in the figure and have a slope of $\frac{1}{3}$. The values of z used in this plot are those given by Rother and Bomke. The points lie in the general neighborhood of the lines but there is no clear indication of a division into two groups. The deviations in about half of the cases are larger than the possible experimental error.

Bomke [33] has recently found that his values of P/e (from calculated K/e and experimental W/e) plotted against the compressibility gave a smooth curve. The equation of this curve was

$$P/e = 0.30k^{-\frac{1}{3}}, \tag{80}$$

where k is the compressibility. Unfortunately he plotted his data on a linear scale and most of the points on his plot were clustered near one axis where the curve was steep, making it difficult to estimate the deviations. A plot of log P/e *vs.* k which is similar to Fig. 23 showed that the deviations were of the same order of magnitude as the deviations in Fig. 23 previously discussed. Hence values of P/e calculated

from equation (80) will only be approximate. The approximation is about the same as computing P/e from equation (78) or (79). Equation (80) has the advantage that P/e is given by a single function.

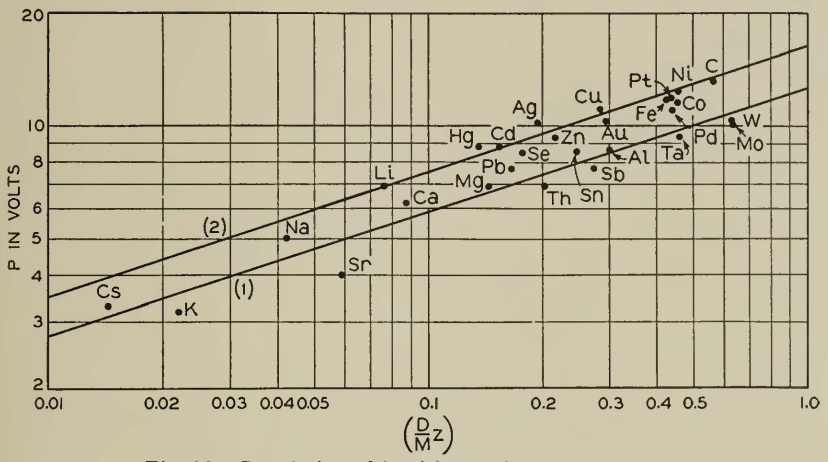


Fig. 23—Correlation of P with atomic properties.

Chittum [34] has related the work function to the bulk modulus of compressibility in a different manner than that used by Bomke. His computed values of the work function deviate from the experimental values by about the same amount as do the values computed from equations (77) and (80).

From the fact that metals with large atomic spacing have a low P while those with small atomic spacing have a high P, it might be expected that P would depend on the spacing of the atoms in the surface layer and that different faces of a single crystal would have different work functions. In fact, Farnsworth and Rose [27] have shown that the contact potential for different faces of a single crystal of Cu varies by about 0.4 volt; and Nitzsche [28] finds the photoelectric work function of two planes of a single crystal of zinc different by about 0.2 volt. Now the values of (D/M) or of the cubic compressibility used in the above calculations do not take into consideration any dependence on the crystal face exposed and, therefore, we would not expect P to be a single-valued function of these properties. This fact may explain the failure to correlate exactly the body properties of the metal with P or the work function. It is quite likely that a better correlation exists between the work function or P and the atomic spacing which prevails on the crystal faces which develop when a metal is heated in a vacuum.

F. Currents Limited by Space Charge

Thus far we have only considered the effect of the surface and applied fields on the number of electrons that escape from a cathode and reach the anode at a particular temperature. However, an electron traveling from the cathode to the anode is also subjected to a field due to all of the electrons in the space between the electrodes. If the electron density in this space is large enough, the current that reaches the anode will be determined by these charges rather than by the work function and temperature of the cathode. The current is then said to be limited by space charge. If, on the other hand, the applied potential is raised to a sufficiently high value, the current is no longer limited by the charges in the space but is then determined by the work function and temperature of the cathode. The current is then said to be saturated or limited by emission. The space charge and saturated emission regions are illustrated by curve 2 in Fig. 24 which is a plot of log i *vs.* log V. In the region to the left of point A, the current is limited by space charge and increases rapidly with the applied potential. To the right of A, the current is limited by emission. Curve 2 has been calculated from equations that will be discussed later. A sharp break point is indicated at A, whereas experimental curves usually show a gradual transition. This gradual transition is due to non-uniformities in work function.

When the current is limited by space charge, the charges in the space increase the height of the potential barrier which electrons must cross in traveling from cathode to anode. The current is determined primarily by the applied potential and electrode geometry and secondarily by the temperature of the cathode and magnitude of the saturated emission. The problem of relating the current to these quantities is very difficult but has been solved on the basis of certain simplifying assumptions for several forms of electrodes by Child,[38] Schottky,[39] Epstein,[40] Fry,[41] Langmuir [42–45] and others. These assumptions together with the solutions will be summarized in this section.

In all of these solutions it is assumed that the maximum of the potential hill which is due to the surface forces and the applied potential, occurs right at the cathode surface. Actually, in the absence of space charge, the maximum in the work distance curve occurs at a small but finite distance from the surface, about 3×10^{-6} cm. for the image equation with moderate applied fields. The space within this distance has a much larger density of electrons than if the potential had its maximum value at the surface. The above assumption, therefore, neglects the influence of these excess charges on the space charge.

This has been justified by Schottky [36] and Laue [37] who concluded that the effect of these charges is negligible. For convenience, the zero of potential is taken not inside the metal but at a point where the electrons

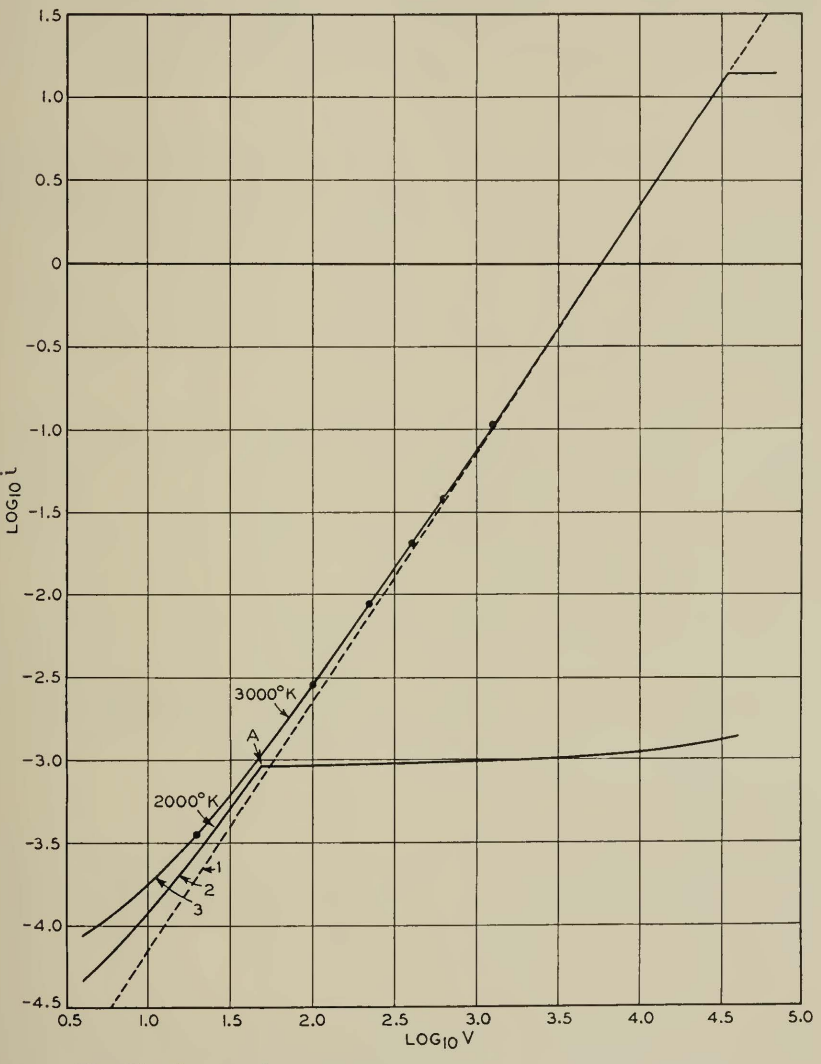


Fig. 24—Currents between parallel plates limited by space charge.

have just overcome the surface forces. Because of the assumption made above, this point is taken at the cathode surface.

In order to obtain a solution, some assumption regarding the velocity

distribution of the emitted electrons must be made. The simplest assumption is that the electrons are emitted from the cathode with zero velocity. From this assumption and the assumption discussed in the preceding paragraph it follows that the potential maximum always occurs at the cathode surface and the emitted electrons are accelerated everywhere in their path between cathode and anode. The current to the anode is determined by the potential distribution in the space.

A better assumption is that the electrons are emitted with a Maxwellian velocity distribution. For this case the potential maximum occurs at some distance from the cathode. The position and value of the maximum depends upon the work function and temperature of the cathode, the geometry of the electrodes and the applied potential. In order that an electron shall reach the anode, its initial velocity normal to the surface must correspond to an energy which is equal to or greater than the potential maximum.

For parallel plates and cylindrical electrodes the following solutions have been obtained for the two assumptions in regard to the velocity distribution of the emitted electrons.

Electrons Emitted with Zero Velocity

For infinite parallel plates:

$$i = (\sqrt{2}/9\pi)(e/m)^{\frac{1}{2}}(V^{\frac{3}{2}}/x^2) = 2.33 \times 10^{-6}(V^{\frac{3}{2}}/x^2), \tag{81}$$

where i is the current to the anode in amp. per cm.^{2}; x the distance in cm. between cathode and anode; and V is the applied potential in volts corrected for the contact potential; e and m are the charge and mass of the electron, respectively.

For long coaxial cylinders:

$$i = (2\sqrt{2}/9)(e/m)^{\frac{1}{2}}(V^{\frac{3}{2}}/R\beta^2) = 1.48 \times 10^{-5}(V^{\frac{3}{2}}/R\beta^2), \tag{82}$$

β is a function of R/r_0 where R and r_0 are the radii of anode and cathode, respectively. Table V shows a few values of β^2 as a function of R/r_0

TABLE V

VALUES OF β^2

R/r_0	1.0	2.0	5.0	7.0	10.0	20.0	40.0
β^2	0.000	0.279	0.767	0.887	0.978	1.072	1.095
R/r_0	70	100	200	400	1000	5000	∞
β^2	1.088	1.078	1.056	1.036	1.017	1.002	1.000

taken from a table given by Langmuir and Blodgett.[44]

The equation for coaxial cylinders only applies to an equipotential cathode. Ordinarily the cathode is a filament and the potential varies along its length because of the heating current. It is of interest to examine the modification of the equation due to this effect. It can be shown that for $V_p > V_f$

$$i = (2\sqrt{2}/9)(e/m)^{\frac{1}{2}}(V_p^{\frac{3}{2}}/R\beta^2) \times [1 - 3V_f/4V_p + 3/24(V_f/V_p)^2 \cdots], \quad (83)$$

where V_p is the applied potential between anode and negative end of the filament and V_f is the total potential drop along the filament.

For the case in which $V_p < V_f$

$$i = (2\sqrt{2}/9)(e/m)^{\frac{1}{2}}(2V_p^{5/2}/5R\beta^2 V_f) = 5.92 \times 10^{-6} V_p^{5/2}/RV_f. \quad (84)$$

For concentric spheres:

$$i = (4\sqrt{2}/9)(e/m)^{\frac{1}{2}}(V^{\frac{3}{2}}/\alpha^2) = 2.96 \times 10^{-5} V^{\frac{3}{2}}/\alpha^2, \quad (85)$$

where α^2 is a function of R/r_0 and has been tabulated by Langmuir and Blodgett.[45] α^2 increases with R/r_0. For $R/r_0 = 5.0$, $\alpha^2 = 1.141$; for $R/r_0 = 10$, $\alpha^2 = 1.777$; for $R/r_0 = 100$, $\alpha^2 = 3.652$.

Electrons Emitted with Maxwellian Velocity Distribution

For infinite parallel plates the space charge limited current is given by

$$i = (\sqrt{2}/9\pi)(e/m)^{\frac{1}{2}}((V - V_m)^{\frac{3}{2}}/(x - x_m)^2) \times [1 + 2.66(kT/(V - V_m)e)^{\frac{1}{2}}]$$
$$= 2.33 \times 10^{-6}((V - V_m)^{\frac{3}{2}}/(x - x_m)^2) \times [1 + 2.48 \times 10^{-2}(T/(V - V_m))^{\frac{1}{2}}]. \quad (86)$$

In this equation

$$V_m = (-2.3Tk/e) \log (i_s/i) = -1.98 \times 10^{-4} T \log (i_s/i) \quad (87)$$

and

$$x_m = 1.092 \times 10^{-6} T^{3/4} \zeta_1 / i^{\frac{1}{2}}, \quad (88)$$

where V is the potential applied between cathode and anode corrected for the contact potential; V_m is the value of the potential maximum measured with respect to the zero of potential previously defined; x the distance between cathode and anode; x_m the distance from the cathode to the potential maximum; T the temperature of the cathode; and i_s the value of the saturation electron emission. ζ_1 is a function of ln (i_s/i). Table VI gives a few values of ζ_1 as a function of ln

TABLE VI

VALUES OF ζ_1.

$\ln (i_s/i)$	0.00	0.30	0.60	1.00	1.60	2.40
ζ_1	0.000	0.979	1.312	1.600	1.881	2.117
$\ln (i_s/i)$	3.4	4.5	7.0	10.0	15.0	25.0
ζ_1	2.293	2.404	2.511	2.544	2.553	2.554

(i_s/i). The values listed in the table were selected from a more extensive table given by Langmuir.[43] From equation (87) it follows that space charge acts as if the work function of the surface were increased by V_m.

An expression for i as a function of V could be obtained by eliminating V_m and x_m between equations (86), (87) and (88). Because of the nature of these equations an analytical expression for i cannot be given. However, for any temperature and electrode spacing it is possible to calculate i as a function of V. The effect of introducing the Maxwellian distribution of velocities can be seen by comparing curves calculated from equations (86), (87) and (88) with equation (81). Such a comparison is made in Fig. 24. Equation (81) gives a straight line with a slope of 3/2 on such a plot and is represented by curve 1. Curves 2 and 3 were calculated from equations (86), (87) and (88) for parallel plates of tungsten spaced 1 cm. apart at 2000 and 3000° K., respectively. The introduction of the Maxwellian velocity distribution causes the currents to be somewhat higher than predicted by the simple 3/2 power law, especially at low applied potentials. Furthermore, at applied potentials considerably less than necessary for saturation, the slope of log i *vs.* log V is less than 3/2. Near the break point, the slope is practically 3/2.

For long coaxial cylinders, Schottky [39] and Langmuir [43, 44] have pointed out that the effect of introducing the Maxwellian velocity distribution is less important than its introduction in the plane parallel case. Langmuir [43] has discussed an approximate formula for this case.

Effect of Fermi-Dirac Velocity Distribution

The introduction of the newer theory that the free electrons in the metal have a Fermi-Dirac velocity distribution requires no modification of the space charge equations deduced on the assumption of a Maxwellian distribution. This is because of the fact that the electrons which escape across the potential barrier at the surface have a Maxwellian distribution, as was shown in connection with Fig. 1. In this connection it is of interest to compare some calculations by Bartlett [46] as-

suming a Fermi-Dirac distribution in the metal at 3000° K. with calculations based on a Maxwellian distribution. This comparison also is shown in Fig. 24, curve 3. The circles were taken from a curve by Bartlett and the line was calculated by equations (86), (87) and (88). The two calculations agree, thus indicating that the assumption of a Fermi distribution in the metal leads to the same result as the assumption of a Maxwellian distribution.

G. Miscellaneous Topics

In order not to lengthen unduly this review we have omitted a discussion of a number of topics. Such topics have either been adequately treated in the reviews and books referred to in the introduction or else no significant advances have been made recently. Some of these topics are: Secondary electron emission, high field emissions, thermionics as related to photoelectricity and contact potential,* and cooling and heating effects accompanying the emission or absorption of electrons. In connection with the last topic we feel that a critical analysis of how the quantities determined by experiment are related to the work function and heat function should be made. Most of these experiments were performed before the day of the Fermi-Dirac-Sommerfeld contributions and should thus be reinterpreted.

Acknowledgments

It gives me pleasure to acknowledge the help I have received in the preparation of this article from my colleagues at the laboratories. I acknowledge in particular the benefit of numerous discussions with Dr. C. J. Davisson and Dr. W. H. Brattain and Mr. R. W. Sears. Professor V. Rojansky, now at Union College, is responsible for some of the more complicated equations used in the checkerboard theory. Mr. R. W. Sears deserves a great deal of credit for his painstaking efforts in preparing the figures and tables and in assembling some of the data.

References

1. K. T. Compton and I. Langmuir, *Rev. Mod. Phys.*, **2**, 123 (1930).
2. S. Dushman, *Rev. Mod. Phys.*, **2**, 381 (1930).
3. A. L. Reimann, *Thermionic Emission*, Aberdeen University Press, Aberdeen, Scotland, or John Wiley & Sons, Inc. (New York, 1934).
4. Müller-Pouillets, *Lehrbuch der Physik*, Vol. IV, F. Vieweg, Braunschweig, 1934.
5. Wien Harms, *Handbuch der Experimentalphysik*, Vol. XIII, Part 2, Akademische Verlagsgesellschaft (Leipzig, 1928).
6. L. B. Linford, *Rev. Mod. Phys.*, **5**, 34 (1933).
7. A. L. Hughes and L. A. DuBridge, McGraw-Hill Book Co. (New York, 1932).
8. J. A. Becker and W. H. Brattain, *Phys. Rev.*, **45** ,694 (1934).
9. P. W. Bridgman, *The Thermodynamics of Electrical Phenomena in Metals*, Macmillan Co. (New York, 1934).
10. S. Dushman, *Phys. Rev.*, **21**, 623 (1923).

* For a critical discussion of this relationship, see Becker and Brattain.[8]

11. A. Sommerfeld, *Zeits. f. Physik*, **47**, 1, 43 (1928); *Naturwiss.*, **15**, 825 (1927).
12. A. Sommerfeld and N. H. Frank, *Rev. Mod. Phys.*, **3**, 1 (1931).
13. R. H. Fowler, *Statistical Mechanics*, University Press (Cambridge, 1929).
14. N. F. Mott, *Proc. Phys. Soc.*, **46**, 680 (1934).
15. W. Hume-Rothery, *The Metallic State*, Oxford University Press (London, 1931).
16. Nordheim, *Physik. Zeits.*, **30**, 177 (1929).
17. L. Tonks, *Phys. Rev.*, **38**, 1030 (1931).
18. Bowden and Rideal, *Proc. Roy. Soc.*, **A120**, 59 (1928); Bowden, *Proc. Roy. Soc.*, **A125**, 446 (1929).
19. W. Schottky, *Ann. d. Physik*, **44**, 1011 (1914).
20. L. H. Germer, *Phys. Rev.*, **25**, 795 (1925).
21. C. J. Davisson, *Phys. Rev.*, **23**, 299 (1924).
22. A. Demski, *Physik. Zeits.*, **30**, 291 (1929).
23. W. B. Nottingham, *Phys. Rev.*, **41**, 793 (1932).
24. J. A. Becker and D. W. Mueller, *Phys. Rev.*, **31**, 431 (1928).
25. J. B. Taylor and I. Langmuir, *Phys. Rev.*, **44**, 423 (1933).
26. W. Knecht, *Ann. d. Physik*, **20**, 161 (1934).
27. H. E. Farnsworth and B. A. Rose, *Proc. Nat. Acad. Sci.*, **19**, 777 (1933); B. A. Rose, *Phys. Rev.*, **44**, 585 (1933).
28. A. Nitzsche, *Ann. d. Physik*, **14**, 463 (1932).
29. W. H. Brattain and J. A. Becker, *Phys. Rev.*, **43**, 428 (1933).
30. N. B. Reynolds, *Phys. Rev.*, **35**, 158 (1930).
31. R. H. Fowler, *Phys. Rev.*, **38**, 45 (1931).
32. F. Rother and H. Bomke, *Zeits. f. Physik*, **86**, 231 (1933).
33. H. Bomke, *Zeits. f. Physik*, **90**, 542 (1934).
34. J. F. Chittum, *Jour. Phys. Chem.*, **38**, 79 (1934).
35. W. Schottky, *Physik. Zeits.*, **15**, 872 (1914).
36. W. Schottky, *Physik. Zeits.*, **15**, 624, 872 (1914).
37. M. v. Laue, *Jahrb. d. Radioakt.*, **15**, 205 (1918); *Berl. Ber.*, **32**, 334 (1923).
38. C. D. Child, *Phys. Rev.*, **32**, 492 (1911).
39. W. Schottky, *Physik. Zeits.*, **15**, 526 (1914).
40. P. S. Epstein, *Verh. d. Phys. Ges.*, **21**, 85 (1919).
41. T. C. Fry, *Phys. Rev.*, **17**, 441 (1921).
42. I. Langmuir, *Phys. Rev.*, **2**, 402 (1913).
43. I. Langmuir, *Phys. Rev.*, **21**, 419 (1923).
44. I. Langmuir and K. B. Blodgett, *Phys. Rev.*, **22**, 347 (1923).
45. I. Langmuir and K. B. Blodgett, *Phys. Rev.*, **24**, 49 (1924).
46. R. S. Bartlett, *Phys. Rev.*, **37**, 959 (1931).

Radio Propagation Over Spherical Earth *

By CHAS. R. BURROWS

The paper shows how Watson's solution for the propagation of electromagnetic waves over perfectly conducting spherical earth merges into the Abraham solution for propagation over a perfectly conducting plane for shorter distances.

The effects of refraction by the lower atmosphere and of the imperfect conductivity of the earth are taken into consideration. The magnitude of the former, which is appreciable, is obtained. The latter is relatively unimportant for ocean water and frequencies of the order of a megacycle and less.

The theoretical solution for radio propagation over perfectly conducting spherical earth with atmospheric refraction is in agreement with available experimental data for propagation over ocean water for frequencies below a few megacycles.

Eckersley's extension of Watson's solution to take into account the effect of the imperfect conductivity of the earth by the phase integral method is found to contain approximations which render its results questionable.

Theory

THE electrical disturbance at the surface of the earth due to a vertical dipole has been calculated by G. N. Watson [1] and others. The results for the case of a perfectly conducting spherical earth with transmitter and receiver both situated on the surface may be reduced to the form:

$$E = \frac{30(2\pi)^{5/3} HI}{a^{5/6}\lambda^{7/6}\sqrt{\sin\theta}} \sum_{n=1}^{\infty} \frac{e^{-\beta_n d\sqrt[3]{2\pi/a^2\lambda}}}{\rho_n}, \qquad (1)$$

where ρ_n and β_n are constants whose values have been calculated as follows:

$\rho_1 = 0.8083$

n	ρ_n/ρ_1	$\beta_n\sqrt[3]{2\pi/a^2}$
1	1.000	0.00376
2	3.188	0.01199
3	4.74	0.0178
4	6.047	0.0227
5	7.236	0.0272
6	8.336	0.0313

* Published in *I. R. E. Proc.*, May, 1935. Presented before U. R. S. I. meeting, Washington, D. C., April 26, 1935.

[1] G. N. Watson, "The Diffraction of Electric Waves by the Earth," *Proc. Roy. Soc.* (*London*) **A95**, 83–99, October 7, 1918.

and

II is the effective height of the transmitting antenna in kilometers,
I is the transmitting antenna current in amperes,
λ is the wave-length in kilometers,
a is the radius of the earth in kilometers ($= 6370$),
d is the distance between transmitter and receiver in kilometers,
θ is the angle at the center of the earth subtended by radii to transmitter and receiver ($= d/a$), and
E is the received field strength in volts per kilometer.

ρ_n and β_n were evaluated for $n = 1, 2$ and 3 by H. M. Macdonald,[2] while the remaining values have been calculated by the present author.

For distances for which this solution would be used (i.e., where the effect of the ionized region of the upper atmosphere may be neglected) $\sin\theta$ very nearly [3] equals θ so that the above formula reduces to the following:

$$E = \frac{30(2\pi)^{5/3} III}{a^{1/3}\lambda^{7/6}d^{1/2}} \sum_{n=1}^{\infty} \frac{e^{-\beta_n d\sqrt[3]{2\pi/a^2\lambda}}}{\rho_n}. \tag{2}$$

This equation may be reduced to a form more readily comparable with the Abraham [4] solution for the field strength over a conducting plane,

$$E = \frac{120\pi III}{\lambda d}. \tag{3}$$

Equation (2) then becomes

$$E = \frac{120\pi III}{\lambda d} f(x), \tag{4}$$

where

$$f(x) = \frac{\sqrt[3]{\pi^2/2a}}{\rho_1} \sum_{n=1}^{\infty} \frac{\sqrt{x}}{\rho_n/\rho_1} e^{-\beta_n\sqrt[3]{2\pi a^2}\, x} \tag{5}$$

and

$$x = d/\sqrt[3]{\lambda}. \tag{6}$$

The constant before the summation sign is equal to 0.1136 when the earth is the sphere under consideration.

[2] H. M. Macdonald, "The Transmission of Electric Waves Around the Earth's Surface," *Proc. Roy. Soc. (London)* **A90**, 50–61, April 1, 1914.

[3] This approximation introduces an error of less than one-tenth of a decibel for distances less than 2250 km.

[4] M. Abraham, "Die Strahlung von Sendedrähten," *Theorie der Elecktrizität*, vol. 2, 2nd edition (1908), 283–294, and "Elektromagnetische Wellen," *Enc. der math. Wissen.*, vol. 5, pt. 2, 482–538, March 18, 1910.

Equation (4) states that the field strength at a point on the surface of a conducting sphere is less than that on the surface of a conducting plane by a factor which is a function of the quotient of the distance along the surface by the cube root of the wave-length.

This factor is plotted in Fig. 1. For small values of $x = d/\sqrt[3]{\lambda}$ it approaches unity so that the Watson solution for radio propagation over the surface of a perfectly conducting sphere merges into the Abraham solution for propagation over a perfectly conducting plane at short distances. In order to depict this graphically the curves that result from neglecting all terms except the first, first two, first three,

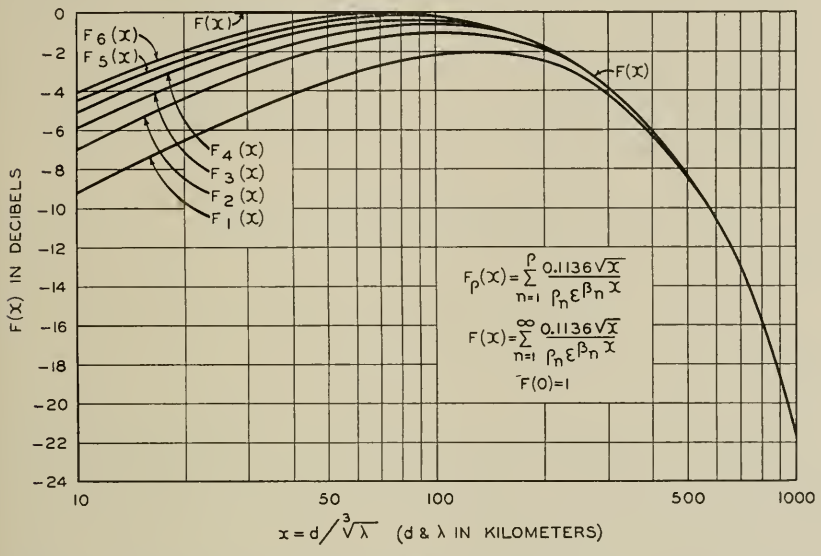


Fig. 1—Successive approximations to Watson's formula for the ratio of the field received over perfectly conducting spherical earth to that over perfectly conducting plane earth.

etc., have been plotted. It will be noted that as more terms of the complete Watson solution are added the resulting curves more nearly approach the Abraham solution for the shorter distances. When x equals 160 the curvature of the earth reduces the field strength one decibel. At this point the error in neglecting all of the terms except the first results in an error of a decibel. For larger values of x the first term approximates the complete series with increasing accuracy, as shown in the curves of Fig. 1. *In other words, no error greater than one decibel is incurred if the Abraham solution is used when* $d/\lambda^{1/3} < 160$ *and only the first term in the Watson solution is employed when* $d/\lambda^{1/3} > 160$.

In obtaining the solution for the propagation of radio waves over the surface of the earth, besides assuming the earth to be a perfect conductor, Watson assumed that the electromagnetic properties of the air were independent of the height above the earth's surface. Data to be presented later indicate that neglecting refraction in the lower atmosphere introduces the greater error for certain frequencies. Fortunately in such cases, it is simpler to extend the solution to take into account atmospheric refraction than the imperfect conductivity of the earth.

It is known [5] that for electromagnetic waves propagated along the surface of the earth, the optical effect of the existing changes in refractivity with height in the lower atmosphere is the same as the effect that would be produced if the earth's radius were increased. If this " effective radius " is substituted for the actual radius in equation (5) the resulting equation for the ratio of the field to that received over a perfectly conducting plane becomes

$$f(y) = 0.1136\sqrt{y}\sum_{n=1}^{\infty}\frac{1}{\rho_n/\rho_1}e^{-\beta_n\sqrt[3]{2\pi/a^2}\,y}, \tag{7}$$

where

$$y = x/\sqrt[3]{K^2} = d/\sqrt[3]{\lambda K^2} \tag{8}$$

and K is the ratio of the effective radius of the earth to the actual radius.

From this it can be seen that the effect of refraction is to multiply the distance at which a given reduction in the field due to the earth's curvature occurs by a factor which is equal to the two-thirds power of the ratio of effective to actual radius of the earth. The analysis of the available meteorological data in the aforementioned article [5] indicates that this radius ratio is about 4/3 on the average. This results in an increase of 1.21 times in the distance at which the reductions in fields occur.[6]

The ratio of the field received over perfectly conducting spherical earth with refraction by the lower atmosphere to that which would be received over a perfectly conducting plane is shown in Fig. 2.

Watson [7] has pointed out the relation of the empirical Austin-Cohen

[5] J. C. Schelleng, C. R. Burrows and E. B. Ferrell, "Ultra-Short-Wave Propagation," *Proc. I.R.E.* **21**, 427–463, March, 1933 and *Bell Sys. Tech. Jour.* **12**, 125–161, April, 1933.

[6] The increase in range is not as great as this due to the inverse distance factor. This advantage would not be realized for waves greater than a certain length. This limit occurs when that part of the atmosphere for which the refractive index no longer decreases at the assumed rate becomes important in the propagation of the waves.

[7] G. N. Watson, "The Transmission of Electric Waves Around the Earth," *Proc. Roy. Soc.* (*London*) **A95**, 546–563, July 15, 1919.

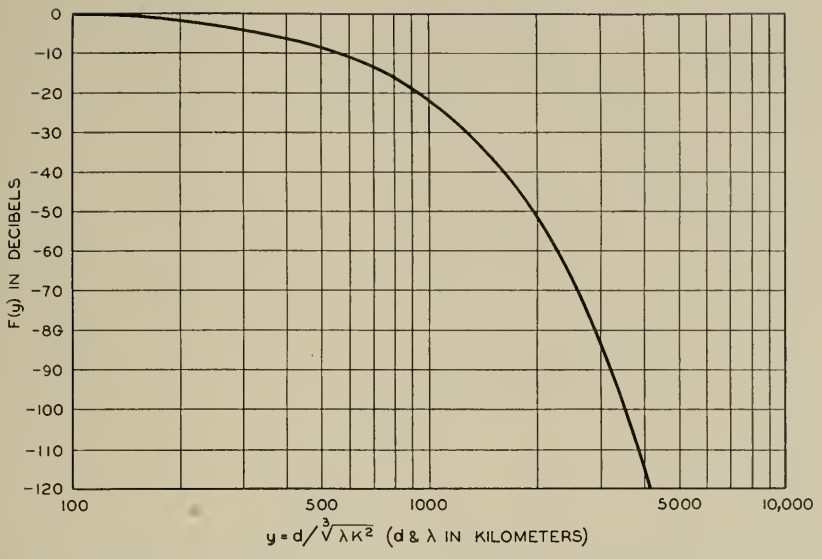


Fig. 2—Ratio of the field received over perfectly conducting spherical earth with refraction by the lower atmosphere to that over perfectly conducting plane earth.

formula for long-distance long-wave communication,

$$E = \frac{120\pi HI}{\lambda d} e^{-0.0015d/\sqrt{\lambda}}, \tag{9}$$

to the above diffraction formulas. He showed that this formula, (9), could be obtained by considering the earth surrounded by a conducting shell some 100 km. above the earth's surface. He also showed that the factor $\lambda^{-1/2}$ instead of $\lambda^{-1/3}$ occurs only when the effect of the upper atmosphere becomes important. Equations (1), (2) and (4) apply only for distances in which the effect of the upper atmosphere may be neglected.

Experiment

In Figs. 3 and 4 the theoretical curve of Fig. 2 has been superimposed upon experimental data[8] obtained for 0.8 and 4 mc. transmission respectively. Theoretical curves are shown for radius ratios of 1, 4/3 and 1.45. The latter gives the best fit with the experimental data. The curve for a ratio of 4/3 estimated from available meteorological data is in fair agreement with the data, but since this is only an estimate of the average value of the ratio it is possible that 1.45 is a better value for the conditions of the experiment. It is doubtful,

[8] All experimental points that represented transmission affected by the ionosphere have been excluded.

however, whether the precision of the experiment would justify distinguishing between these two values. It will be noted that the effect of refraction is appreciable and that the agreement between experiment and theory is greatly improved by taking the effect of refraction into account.

As an indication of the effect of the finite conductivity of ocean water, the theoretical curve for propagation over imperfectly conducting plane earth has been added in each case. Curve 4 for imperfectly conducting plane earth is substantially the same as that for a perfectly conducting plane for 0.8 mc. (Fig. 3), indicating that the effect of the

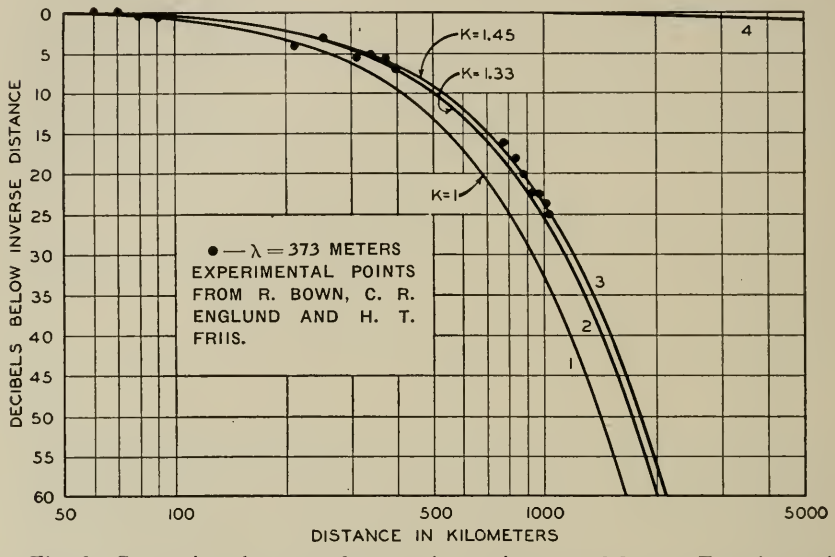


Fig. 3—Comparison between *theory* and *experiment* on 0.8 mc. Experimental points show received field strength from S. S. America, March 1922, taken from Fig. 16 of "Radio Transmission Measurements" by R. Bown, C. R. Englund and H. T. Friis, *Proc. I.R.E.* **11**, 115–152, April 1923.
Curve 1—Theoretical neglecting refraction.
Curve 2—Theoretical assuming average refraction from meteorological data.
Curve 3—Theoretical assuming refraction to give best fit with experimental points.
Curve 4—Theoretical for plane earth taking finite conductivity into account.

imperfect conductivity is negligible on this frequency. For 4 mc., Fig. 4, curve 4, the effect of the imperfect conductivity while not negligible is small compared to the effect of the earth's curvature.

If an attempt be made to take into account the imperfect conductivity of the earth by applying Eckersley's extension of Watson's solution, curve 3 for $K = 1.45$ of Fig. 4 would be moved almost back to curve 1 for $K = 1$. There are, however, several reasons for questioning this extension of Watson's work that will be discussed later.

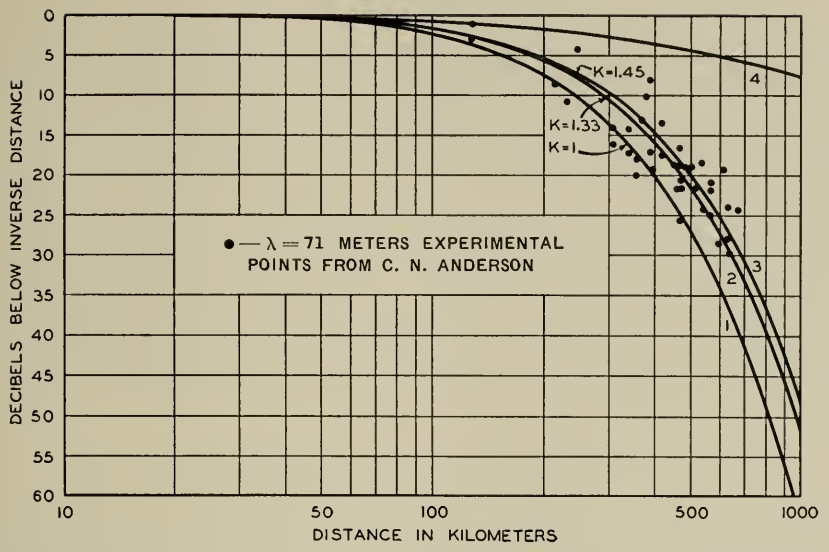


Fig. 4—Comparison between *theory* and *experiment* on 4.2 mc. Experimental points from Fig. 1 of "North Atlantic Ship-Shore Radio Telephone Transmission During 1930 and 1931" by C. N. Anderson, *Proc. I.R.E.* **21**, 81–101, January 1933.
Curve 1—Theoretical neglecting refraction.
Curve 2—Theoretical assuming average refraction from meteorological data.
Curve 3—Theoretical assuming same refraction as curve 3 of Fig. 3.
Curve 4—Theoretical for plane earth taking finite conductivity into account.

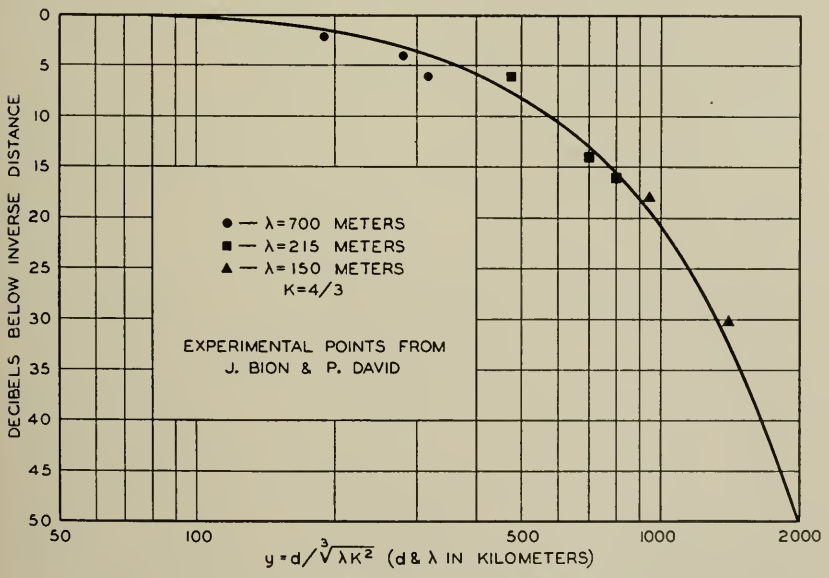


Fig. 5—Comparison between *theory* and *experiment* on 2, 1.4, and 0.43 mc. Experimental points from Bion and David; theoretical curve from equation (7).

Data published by Bion and David [9] to show the inadequacy of Sommerfeld's solution for the propagation over sea water in the wavelength range 150 to 700 meters, have been plotted in Fig. 5. While only eight points are shown they represent data taken at regular intervals on a ship whose distance from the transmitter was continuously increased up to 1050 km. so that their precision is far superior to that possible with single measurements. The points have been plotted against the parameter $y = d/\sqrt[3]{\lambda K^2}$ using 4/3 for the value of K. The points lie close to the theoretical curve, substantiating the theoretical curve and indicating that atmospheric refraction was sufficient to increase the effective radius of the earth by the factor 4/3 for radio propagation (in this frequency range) over the Mediterranean Sea in January and February, 1932.

Effect of Imperfect Conductivity

Due to the complications introduced into the problem of the propagation of electromagnetic energy around the surface of the earth by the effect of imperfect conductivity, no rigorous solution has been made to date. The approximate solution due to T. L. Eckersley,[10, 11] however, has been used [12] to calculate the field strength of the ground wave at distances beyond those for which the solution for transmission over an imperfectly conducting plane applies. The results of this solution will be compared with the rigorous solutions of special cases, leaving a discussion of some of the approximations made and the uncertainties introduced thereby for the appendix.

Theoretical curves obtained by various methods for propagation over the surface of the earth are presented in Fig. 6 for comparison. Curve A is for perfectly conducting spherical earth based on Watson's solution. Curve B is based on Eckersley's solution for a spherical earth whose conductivity is small enough so that its magnitude is unimportant but large enough so that it is essentially a conductor rather than a dielectric.[13] Curves C and D result from using the coefficients given by Eckersley corresponding to the values of $\sigma^{1/2}\lambda^{5/6}$ indicated on the curves. Curves E, F, G and H are for imperfectly

[9] J. Bion and P. David, "Sur L'Affaiblissement des Ondes Moyennes et Intermediaires se Propageant de Jour sur Mer," *Comptes Rendus* **194**, 1723–1724, May 17, 1932.

[10] T. L. Eckersley, "Radio Transmission Problems Treated by Phase Integral Method," *Proc. Roy. Soc.* (*London*) **A136**, 499–527, June 1, 1932.

[11] T. L. Eckersley, "Direct Ray Broadcast Transmission," *Proc. I.R.E.* 20, 1555–1579, October, 1932.

[12] See for example, "Report of Committee on Radio Propagation Data," *Proc. I.R.E.* **21**, 1419–1438, October, 1933.

[13] For detailed explanation of this curve see the appendix.

conducting plane earth based on the solution by Sommerfeld,[14] Weyl,[15] Wise[16] and others and evaluated by Rolf,[17] for the corresponding values of $\sigma^{1/2}\lambda^{5/6}$ indicated on the curves.[18] The part of curve *H* shown also coincides with the solution for perfectly conducting plane earth as determined by Abraham.[4] This indicates that for conductivities greater than those for which $\sigma^{1/2}\lambda^{5/6} = 10^{-5}$ the earth may be regarded as a perfectly conducting sphere.

The fact that the plane earth solution for values of the parameter of the order of 10^{-7} and less (curves *E* and *F*) gives lower fields than Eckersley's solution for spherical earth indicates that the approximations made introduce large errors in these regions of the solution. This inconsistency between the Eckersley solution and the rigorous solution for plane earth in itself would indicate that the solution is

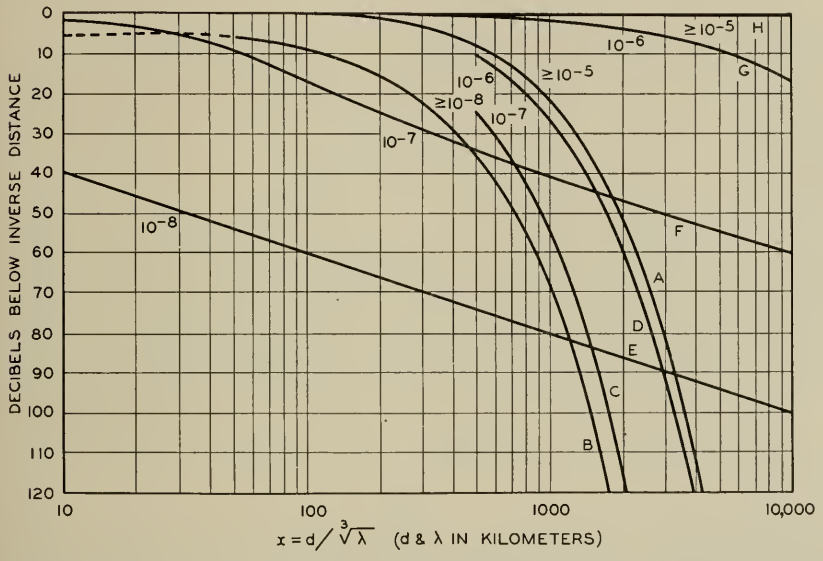


Fig. 6—Comparison of theoretical curves for radio propagation. The numbers on the curves indicate the value of $\sigma^{1/2}\lambda^{5/6}$ for the case represented by the curve in question (conductivity in electromagnetic units and wave-length in kilometers).

[14] A. Sommerfeld, "Über die Aursbreitung der Wellen in der drahtlosen Telegraphie," *Ann. der Phys.* **4**, 28, 665–736, March 16, 1909 and "Ausbreitung der Wellen in der drahtlosen Telegraphie. Einfluss der Bodenbeschaffenheit auf gerichtete und ungerichtete wellenzuge." Jahr. d. drahtlosen T. 4, 157–176, December, 1910.

[15] H. Weyl, "Ausbreitung elektromagnetischen Wellen über einem ebenen Leiter," *Ann. der Phys.* **4**, 60, 481–500, November 20, 1919.

[16] W. Howard Wise, "The Grounded Condenser Antenna Radiation Formula," *Proc. I.R.E.* **19**, 1684–1689, September, 1933.

[17] B. Rolf, "Graphs to Prof. Sommerfeld's Attenuation Formula for Radio Waves," *Proc. I.R.E.* **18**, 391–402, March, 1930.

[18] The usual parameter, $\sigma\lambda^2$, that occurs when the field over imperfectly conducting plane earth is plotted against distance, becomes $\sigma\lambda^{5/3}$ or $(\sigma^{1/2}\ \sigma^{5/6})^2$ when the field is plotted against $d\lambda^{-1/3}$.

not valid for values of $\sigma^{1/2}\lambda^{5/6}$ less than 10^{-7}. While no glaring inconsistencies are evident from Fig. 6 for values of the parameter somewhat greater than this it is the writer's opinion that implicit faith should not be placed in the results without experimental verification due to the nature of the approximations made in obtaining the solution.[19] Comparison of curves *D* and *A* shows that the Eckersley modification of the Watson solution for values of the parameter of the order of 10^{-6} is small which is consistent with the results for plane earth, curve *G*. It is in this region that Eckersley presents experimental data to substantiate his solution.

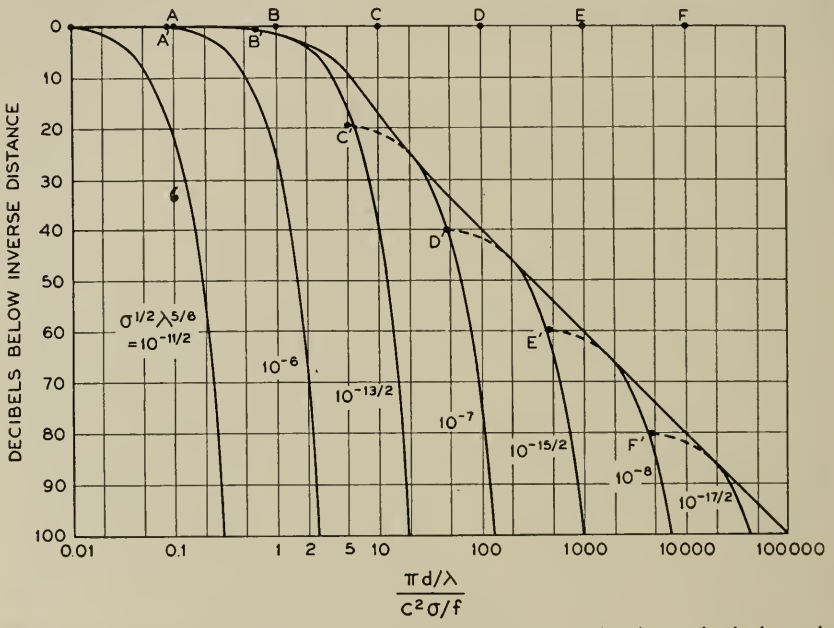


Fig. 7—Revised Eckersley curves for imperfectly conducting spherical earth. (The conductivity in e.m.u. is represented by σ and in e.s.u. by $c^2\sigma$, $c = 3 \times 10^{10}$.)

Recently Eckersley[20] has made the plausible suggestion that his curves should be shifted vertically until they are tangent or nearly tangent to the Sommerfeld curve. This results in the curves of Fig. 7.

[19] Since the writing of this paper an article by Jean Marique entitled "Note sur Quelques Measures due Rayonnement des Stations de Navires," *L'Onde Electrique* **13**, 149–156, March, 1934 has come to the attention of the author. Experimental data are presented from which he concludes that Eckersley's results do not apply for distances of the order 400 to 500 km. at a wave-length of 600 meters.

[20] T. L. Eckersley, "Study of curves of propagation of waves," Document A. G. (1934), No. 11, comm. II; International Scientific Radio Union, Vth Assembly.

Here the abscissa is chosen so that all of the Sommerfeld-Rolf curves coincide. If the effect of imperfect conductivity were unimportant, the curves for spherical earth would begin to depart from unity at the points A, B, C, etc. The effect of imperfect conductivity is to move these points to A', B', C', etc., on the present Eckersley theory. The horizontal motion was calculated by the approximate phase integral method, while the vertical motion is the result of this recent assumption. It can be seen that for the poorer conductivities the recent assumption causes a greater change in the value of the field strength than that calculated by the phase integral method. While this assumption has removed the most obvious inconsistency in the results, the writer believes that they still require experimental verification before reliance should be placed in them.

APPENDIX

The rigorous solution for the perfect conductivity case, equation (1), may be expressed in the form,

$$E = \sum_{n=1}^{\infty} A_n \cos \lambda_n, \tag{10}$$

where A_n and λ_n are functions of ρ_n. By his approximate phase integral method Eckersley was able to evaluate λ_n in the above expression. He found the same relationship between λ_n and ρ_n as Watson. The ρ_n's he obtained, however, differed from those obtained by Watson. The values of ρ_n as determined by Eckersley may be expressed

$$\rho_n = \frac{[3\pi(n - a_n + \eta)]^{2/3}}{2}, \tag{11}$$

where η depends upon the ground constants, being zero for perfect conductivity. a_n is a constant independent of n whose value Eckersley found by comparison with Watson's results to be 3/4. Herein is one of the inaccuracies introduced by the approximate method, for to obtain the correct values of ρ_n, a_n must be allowed to vary with n. While the necessary variation is small [21] for the case of perfect conductivity, without further proof we have no assurance that it is not much larger for the more general case.

Eckersley's method does not tell us anything about the magnitude of A_n in equation (10). He tacitly assumed A_n to be independent of

[21] $a_1 = 0.7819$, $a_2 = 0.7577$, $a_3 = 0.7544$, $a_4 = 0.7530$, $a_5 = 0.7523$, and $a_6 = 0.7519$. For larger values of n, a_n approaches 0.75 more closely.

the conductivity. An equally logical assumption leading to a different result would be that the functional relationship between A_n and ρ_n be independent of the conductivity. Both are undoubtedly incorrect but the error introduced may not be large for the better conductivities.

In obtaining curve B of Fig. 6 the values of a_n given in footnote 21 were used so that Eckersley's solution would be consistent with Watson's solution for the perfect conductivity case. The values of A_n were calculated on the assumption that the functional relationship between A_n and ρ_n be independent of the conductivity. If the magnitude of A_n were assumed independent of the conductivity, curve B would be raised approximately 7 db.

A Single-Sideband Short-Wave System for Transatlantic Telephony *

By F. A. POLKINGHORN and N. F. SCHLAACK

This paper describes the construction of a short-wave single-sideband reduced-carrier system of radio transmission. It also reports the results of comparisons made between this system and an ordinary short-wave double-sideband system between England and the United States. It was found that the single-sideband system gave an equivalent improvement in radiated power over the double-sideband system averaging 8 db. This is in good agreement with the theoretical improvement to be expected.

Introduction

THE single-sideband suppressed-carrier method of transmission has been used to effect economies in the power capacity required, energy consumed, and space in the frequency spectrum on carrier telephone circuits for over fifteen years. On the basis of equal peak amplitudes in a transmitter a single-sideband suppressed-carrier system gives a possible theoretical improvement of 9 db in received signal-to-noise ratio over a double-sideband and carrier system. Six db of this improvement is obtained by omitting the carrier and utilizing the entire available amplitude capacity of the transmitter for the sideband. The other 3 db is obtained by reducing the band width of the receiver to only that required to pass one sideband, thus reducing the noise energy at the receiver output by one-half.

In order that speech may be transmitted without undue distortion over a single sideband system, it is necessary that the carrier frequency at the receiver be within about $\pm$ 20 cycles of the correct value. For the transmission of music a much higher precision is required. The practical construction of a single-sideband radio system at frequencies of the order of 60 kc., such as is used in the long-wave transatlantic telephone circuit, requires only a careful application of known technique to obtain the desired degree of stability of the oscillators. At the short-wave transatlantic radio telephone frequencies of from 5,000 to 20,000 kc., however, the very best crystal oscillators, such as are now used only for the very highest quality laboratory standards, would be required at both transmitter and receiver to obtain the degree of synchronization required.

This high degree of frequency stability can be dispensed with by

* Published in *Proc. I.R.E.*, July, 1935.

transmitting a pilot frequency over the channel. For this purpose the carrier frequency serves as well as, if not better than, any other frequency since it is easily obtainable at the transmitter and is readily utilized at the receiver. If a single-sideband transmitter is fully loaded by two equal side frequencies and a carrier of amplitude 10 db below one of the side frequencies, the power in the carrier is only about 5 per cent of the power in the side frequencies. If the peak voltage in the transmitter is kept the same, each side frequency could be 1.3 db greater when no carrier is transmitted. Practically, it is found that since distortion rather than peak voltage is the limiting factor, the presence of the carrier 10 db down has no appreciable effect on the permissible sideband amplitude. By using a very narrow filter at the receiver to pass the carrier, the same carrier-to-noise ratio can be obtained with the reduced carrier as is ordinarily obtained with a common double-sideband receiver receiving a carrier of full strength. After passing through this narrow filter the carrier may be used to synchronize automatically a local carrier, or by amplifying to a greater extent than the sideband and recombining with the sideband, it may be used for direct demodulation of the sideband. When used in the latter manner it will be called "reconditioned carrier."

In 1928, after extensive tests of short-wave double-sideband transmission had been conducted [1] and while the short-wave transatlantic telephone channels between the United States and England were under construction, some preliminary trials of a short-wave single-sideband system were made under the direction of Mr. R. A. Heising between Deal, New Jersey, and New Southgate, England, using a local carrier supply at the receiver. The local carrier was produced by beating the output of a variable-frequency tuned-circuit oscillator with that of a crystal oscillator. It was necessary to adjust the oscillator continuously in order to keep the oscillator frequency in the proper relation to the incoming sideband.

Notwithstanding the limitations of the equipment, encouraging results were obtained and study of the problem was continued, although along a slightly different line. Receivers were built which were capable of separating the sidebands and carrier of an ordinary double-sideband and carrier transmission in such a manner that single-sideband and other types of reception could be simulated. The carrier could be separately filtered and reconditioned so that even with

[1] Reports of some of these tests were contained in the following articles: "Some Measurements of Short Wave Transmission," R. A. Heising, J. C. Schelleng and G. C. Southworth, *Proc. I. R. E.*, October, 1926. "Transmission Characteristics of a Short-Wave Telephone Circuit," R. K. Potter, *Proc. I. R. E.*, April, 1930. "The Propagation of Short Radio Waves over the North Atlantic," C. R. Burrows, *Proc. I. R. E.*, September, 1931.

considerable selective fading a satisfactory carrier was continuously available. Tests made with these receivers showed that the elimination of one sideband at the receiver did not affect the intelligibility or quality of reception to any extent if allowance were made for the reduction in the received power.

Description of Apparatus

For the purpose of obtaining more complete quantitative information on the improvement to be realized from single-sideband operation and a better understanding of the requirements of commercial single-sideband equipment, apparatus was constructed for a trial of a short-wave single-sideband system across the Atlantic. Transmitter input equipment was constructed which was capable of delivering a single-sideband signal to the input of the water-cooled amplifiers used in the short-wave double-sideband transmitters. This input equipment was sent to Rugby, England, and with the cooperation of the British Post Office installed in conjunction with one of the transatlantic transmitters. For comparison purposes the normal double-sideband output of this same transmitter was used. A single-sideband receiver having a number of novel features was also constructed and installed at the transatlantic receiving station at Netcong, New Jersey. During the latter part of 1933 and the early part of 1934 comparative tests of double and single-sideband transmission were conducted between the British Post Office Headquarters in London and the Bell Telephone Laboratories in New York City.

Transmitting Input Equipment

Figure 1 shows a rear view of the transmitting input equipment. The equipment is mounted on three bays of panels in two welded steel cabinets, each panel being the width of the usual telephone relay rack panel. A schematic of the input equipment is shown in Fig. 2. The incoming speech is applied to the balanced modulator No. 1, to which is also applied voltage having a frequency of 125 kc., obtained through a multivibrator from a 625 kc. crystal oscillator. The low-frequency filter following the first modulator is of the lattice type of construction and uses quartz crystals as elements [2] in order to obtain the necessary attenuation to the carrier frequency and one sideband while passing the other sideband. This filter passes frequencies from 125.1 kc. to 130 kc. The unwanted sideband is suppressed from 40 to 60 db and

[2] For information on the construction of such filters, see article by W. P. Mason, "Electrical Wave Filters Employing Quartz Crystals as Elements," *Bell Sys. Tech. Jour.*, Vol. XIII, No. 3, July, 1934.

the carrier is suppressed approximately 20 db in the modulator and about 15 db more in the filter. In order to obtain a variable amplitude of carrier for experimental purposes, an arrangement was provided for by-passing a variable quantity of the carrier around the first modulator and low-frequency filter. The single-sideband voltage obtained from

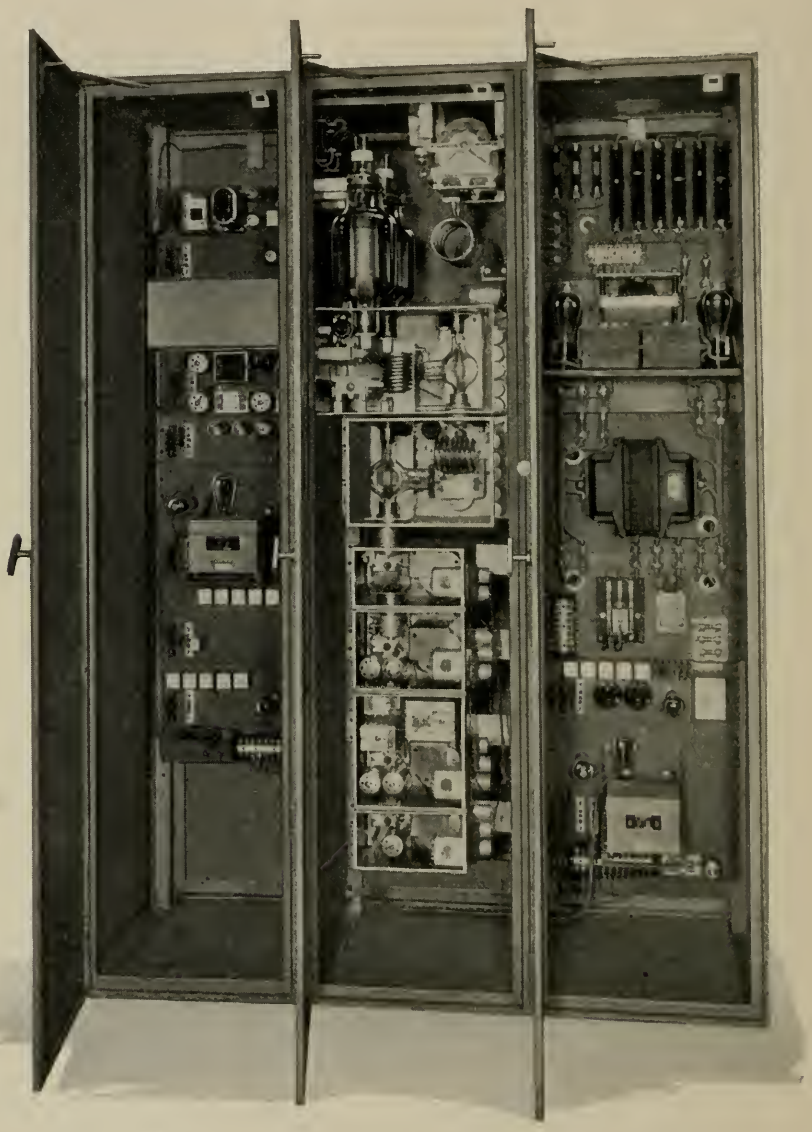


Fig. 1—Rear view of single-sideband transmitting input equipment.

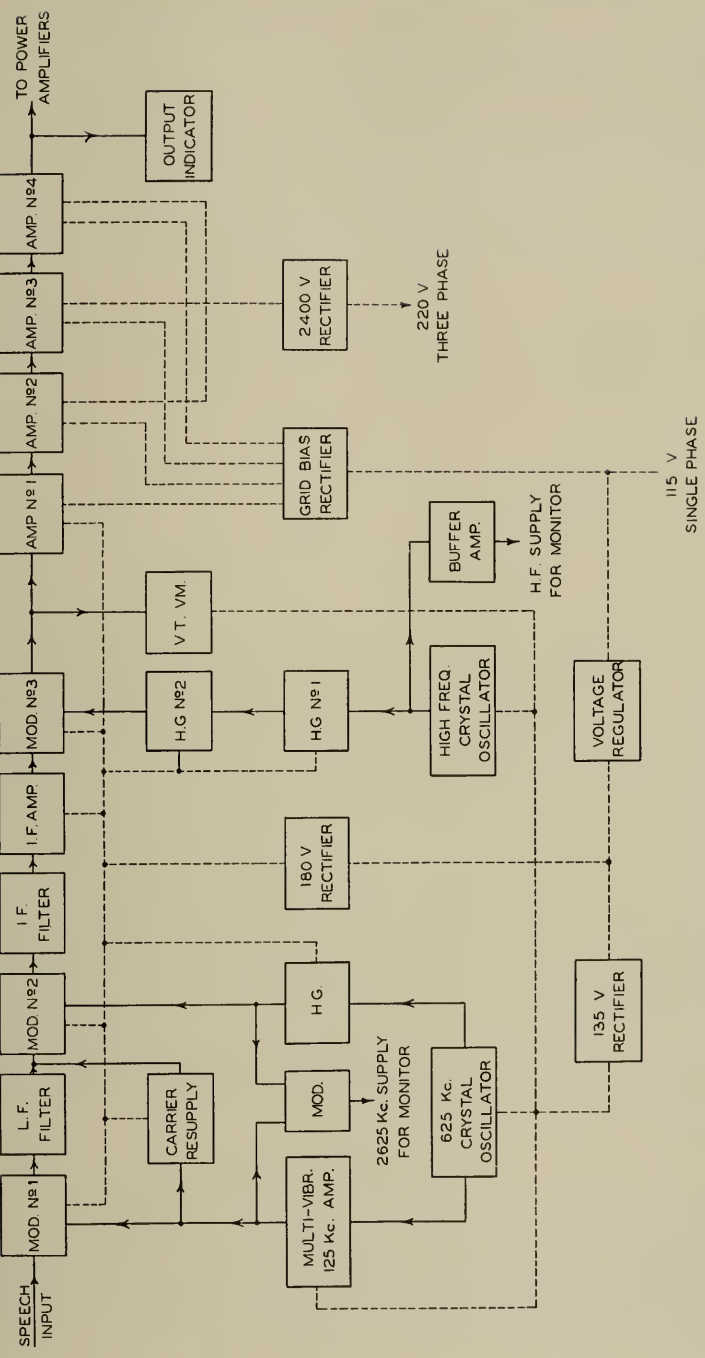


Fig. 2—Schematic of single-sideband transmitting input equipment.

the low-frequency filter, together with the reintroduced 125 kc. carrier, is impressed on the input of balanced modulator No. 2. A 2,500 kc. carrier voltage, which is obtained from the 625 kc. crystal oscillator by means of a harmonic generator, is also supplied to the input of the second modulator. The intermediate frequency filter which follows the second modulator passes the upper sideband generated in the second demodulator (from 2,625.1 to 2,630 kc.) and suppresses the other sideband and the carrier approximately 50 db. The single sideband thus obtained is then amplified before it is impressed on the input of the third modulator. The circuits up to and including the intermediate amplifier are fixed and do not have to be adjusted in order to change the final output frequency of the equipment. The third modulator is of the unbalanced type and both the output of the intermediate frequency amplifier and a third carrier are applied to its input. The third carrier is obtained from a high-frequency crystal oscillator through two harmonic generators in tandem. The frequency of the carrier applied to the third modulator depends on the output frequency desired and since either sideband may be selected the carrier frequency must be 2,625 kc. greater or less than the desired final output carrier frequency. In order to cover the range from 4,700 kc. to 21,000 kc., the carrier must range from 7,325 to 18,375 kc. No filter is required in the output of the third modulator since the output tuned circuits are narrow enough to exclude the third carrier and the unwanted sideband, which are respectively 2,625 and 5,250 kc. away from the desired sideband. The output circuit of the third modulator is the first point in the equipment where the final frequency to be transmitted is obtained. The output voltage of the third modulator is applied to the input of a series of four amplifiers in tandem, which serve to increase the amplitude of the single sideband and the reduced carrier to a value which will excite to full capacity the power amplifiers of a regular double-sideband transmitter. Receiving type screen-grid tubes are used in all but the multi-vibrator, crystal oscillator and the final amplifiers. Amplifiers 2 and 3 consist of one 75-watt screen-grid tube each and amplifier 4 consists of two 1-kw. screen-grid tubes in push-pull.

Transmitting Monitor

It is extremely important in operating the single-sideband equipment to know that the distortion is within reasonable limits. With the ordinary double-sideband type of transmission it is possible to simulate the receiving equipment with a very simple rectifier, thus allowing local distortion tests to be made on the transmitter. With

single-sideband transmissions in which the carrier is either totally or partially suppressed such a simple receiver is not adequate, as the distortion produced in a simple rectifier would be excessive. It is necessary that a carrier of the right frequency and of an amplitude considerably greater than that of the sidebands be present in the rectifier. After a study of the situation it was decided to build up the carrier for monitoring purposes from the same crystal oscillators used in the transmitter. The monitoring device, a schematic of which is shown in Fig. 3, consists of two detectors and a harmonic generator

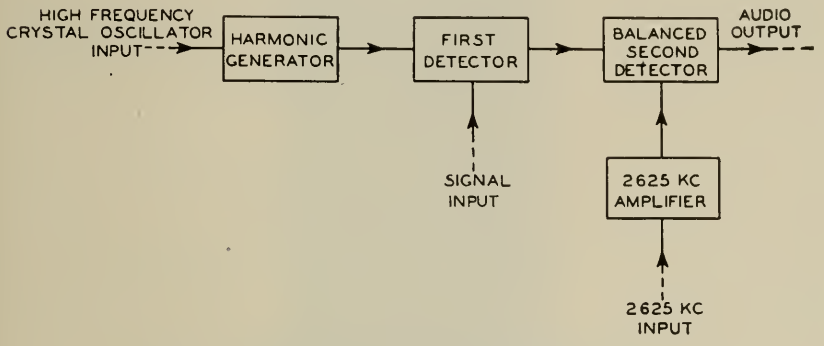


Fig. 3—Schematic of transmitting monitoring device.

which take the high-frequency carrier and combine it with the signal to produce an intermediate frequency, which is in turn beaten with the 2,625 kc. derived from the low-frequency carrier crystal to obtain a demodulated voice frequency.

Receiver

The front view of the receiver is shown in Fig. 4. The receiver is mounted in a steel cabinet seven feet high and a standard telephone bay in width. Figure 5 shows a block schematic of the receiver. The receiver is of the usual double-detection variety, having a high-frequency amplifier stage, a balanced first detector, a three-stage intermediate frequency amplifier and a balanced second detector. A branch circuit, taken from the grid of the third intermediate frequency amplifier tube, contains a narrow crystal filter which passes the carrier, but not the sideband. After passing through the filter the carrier is amplified and rectified by a linear rectifier, the rectified output giving automatic volume control action on the high-frequency tube, first detector, and first and second intermediate frequency amplifiers. Another branch circuit passes the filtered carrier through an overloaded amplifier which reduces the fluctuations of carrier

amplitude which may be present due to fading or modulation. This reconditioned carrier is then used for obtaining automatic frequency control of the beating oscillator and synchronization of the local



Fig. 4—Front view of single-sideband receiver.

carrier oscillator, or it may be applied directly to the second detector for demodulation purposes.

By using an intermediate frequency band of moderate width, an ordinary double-detection receiver for double-sideband operation

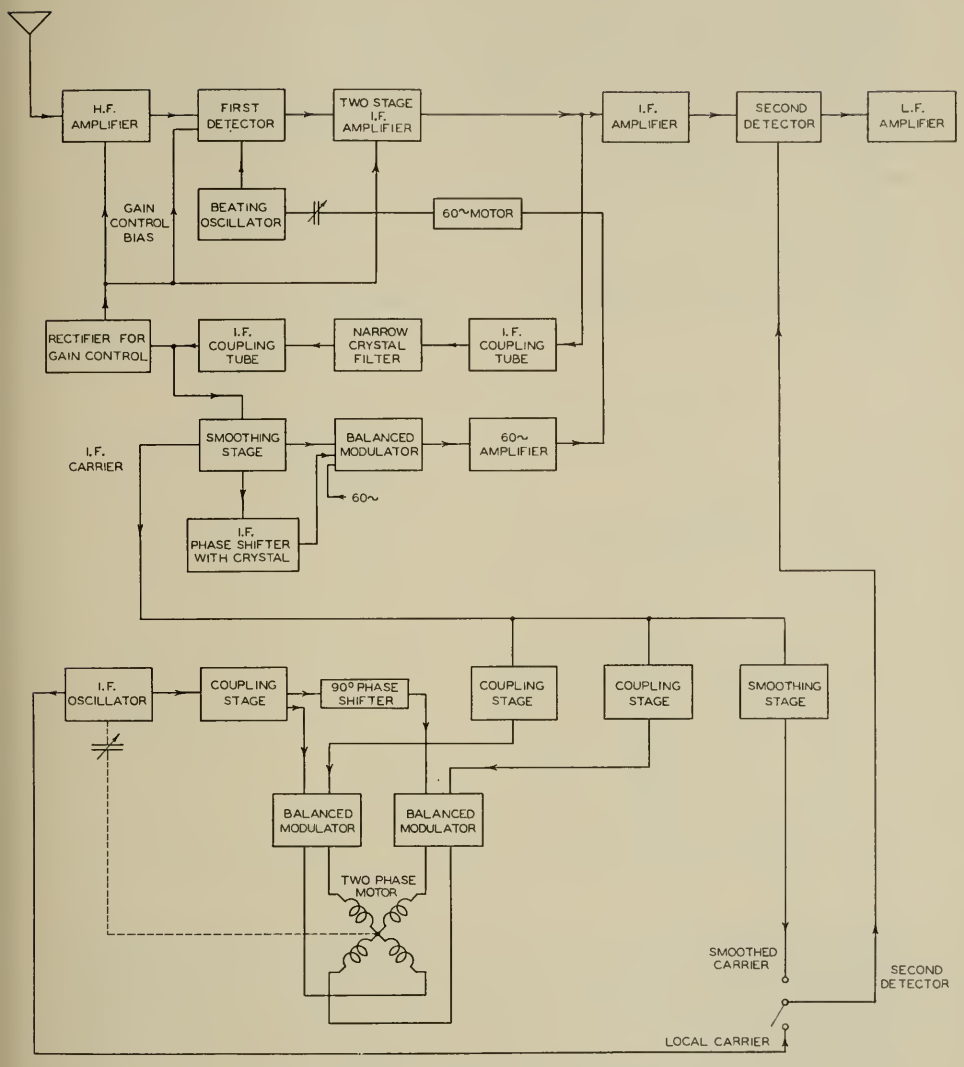


Fig. 5—Schematic of single-sideband receiver.

may be built which will require tuning of the beating oscillator at very infrequent intervals, perhaps only two or three times a day. For receivers in which the carrier is to be separated from the sideband by a narrow filter, a much higher degree of frequency stability is required in both the transmitter and the receiving beating oscillator if frequent or almost continuous tuning is to be avoided. Rather than endeavor to obtain the high-frequency stability required, it was decided to arrange that the incoming carrier automatically tune the beating oscillator of the receiver in such a manner that the carrier at intermediate frequency would always pass through the narrow crystal filter in a satisfactory manner. The manner in which this is accomplished is shown in Fig. 6. The reconditioned carrier is introduced

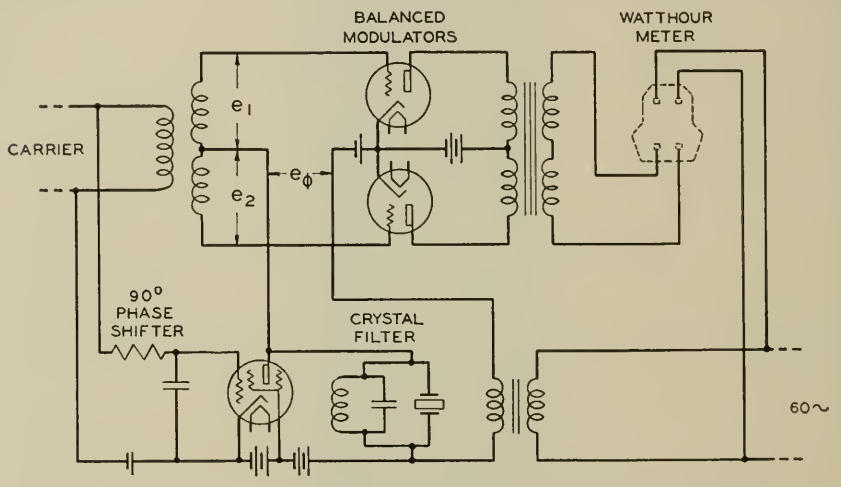


Fig. 6—Schematic of automatic tuning device.

in push-pull fashion on the grids of a balanced modulator system. The same carrier is passed through a circuit having a 90-degree phase shift, through a narrow band suppression filter, and applied to the same two grids in parallel. A small 60-cycle voltage is also applied to the grids in parallel. The 60-cycle output voltage of the balanced modulators is applied through a transformer to the rewound current coils of a watt-hour meter. When the carrier frequency is that of maximum suppression for the narrow filter, equal voltages e_1 and e_2 will be applied to the grids of the two tubes forming the balanced modulator. If the carrier frequency shifts from this position, the voltage applied to each grid will be the vector sum of e_1 or e_2 and a voltage of variable magnitude and phase e_ϕ, which appears in parallel on the two grids. The

magnitude of e_ϕ voltage increases as the frequency of the carrier at intermediate frequency departs from its proper value, causing the voltages on the two grids to change, one becoming higher than the other as shown by $e_1 + e_\phi$ and $e_2 + e_\phi$ of Fig. 7. As the amplitude of

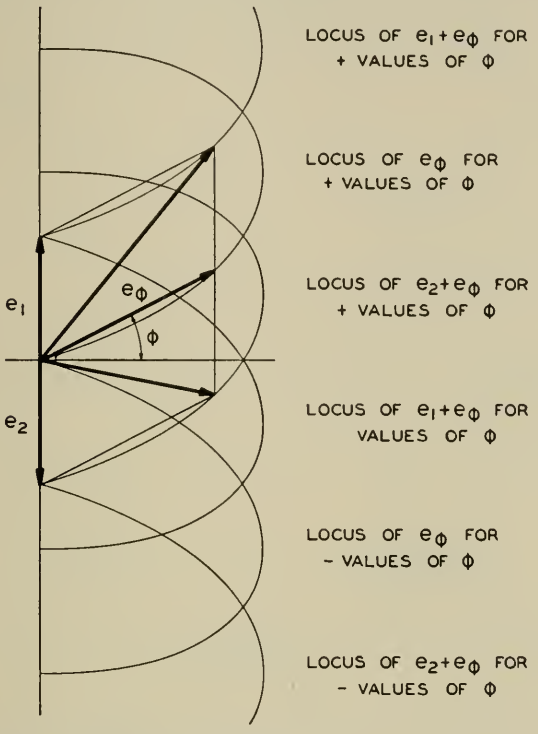


Fig. 7—Vector diagram of input voltages to balanced modulators of the automatic tuning device.

the applied radio frequency voltage increases, the mutual conductance of the modulator tube decreases and consequently a greater amount of 60-cycle current flows in the plate circuit, the phase of which depends upon which tube has the higher mutual conductance. The voltage coil of the watt-hour meter is permanently connected to the supply lead. When the frequency of the carrier at intermediate frequency is too high, the phases will be such that the watt-hour meter runs in one direction, and when the frequency of the carrier is too low the watt-hour meter runs in the other direction. A very small condenser is substituted for the registering mechanism of the watt-hour meter. This condenser is connected to the beating oscillator circuit and the whole circuit arranged in such a manner that the watt-hour meter runs

until the beating oscillator gives the proper frequency, when the action stops.

Since this automatic tuning unit holds the carrier at intermediate frequency in a fixed relation with respect to a crystal filter, which may drift slightly in resonant frequency from time to time, and not in synchronism with a local carrier oscillator, it is necessary that a sepa-

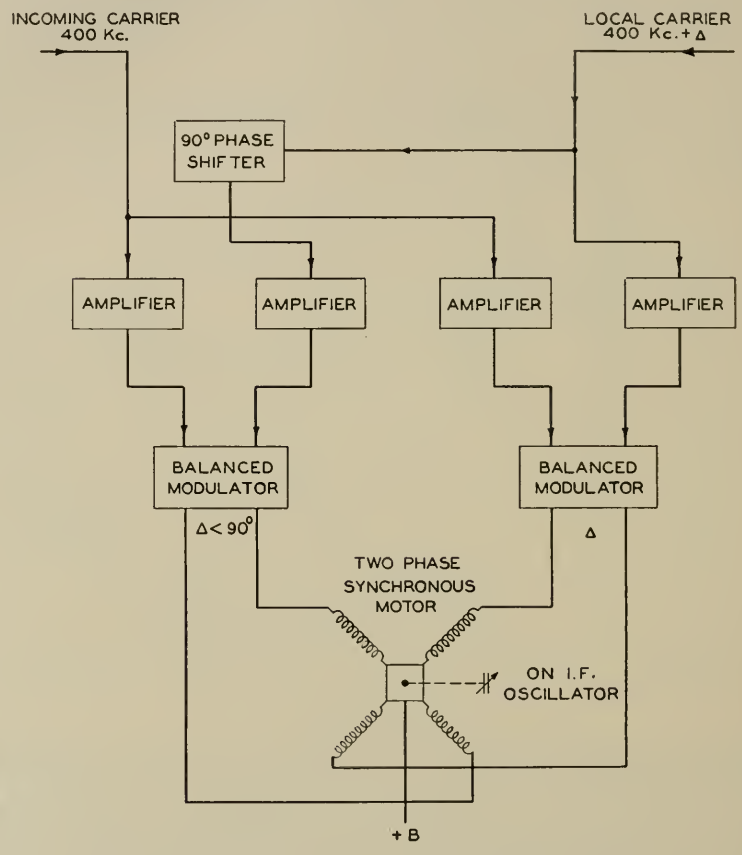


Fig. 8—Schematic of automatic synchronizing equipment.

rate mechanism be provided for synchronizing the local carrier if a local carrier is to be used. A schematic diagram of the circuit for doing this is shown in Fig. 8. The reconditioned carrier at intermediate frequency is introduced through amplifiers to two balanced modulators. The output of the local carrier oscillator is introduced to the same modulators, to one of them directly and to the other through a device which shifts the phase 90 degrees. The phases of

the outputs of these balanced demodulators will be in quadrature and the frequency will be the beat frequency, Δ, between the incoming and local carriers. These voltages operate a variable reluctance type synchronous motor [3] which is mechanically connected to a condenser which forms a part of the local carrier oscillator circuit. The motor operates until the frequency of the local carrier oscillator is exactly the same as the carrier at intermediate frequency, when the frequency applied to the two-phase motor becomes zero.

For distortion testing the receiver can be used as an harmonic analyzer, the frequency of the beating oscillator being shifted so that only the desired distortion product passes through the narrow crystal filter. Measurements made in this way when the transmitter and receiver were close together checked very well with measurements made using the monitoring unit previously described.

A balanced second demodulator system was used, as the distortion is much less than with other types. No attempt is made to separate the incoming carrier from the sideband in the second demodulator, the amplitude of the reconditioned carrier or the local carrier supplied to the second demodulator being several times the amplitude of the carrier transmitted with the sidebands.

Experimental Results and Discussion

To determine experimentally in a quantitative manner the relative merits of two radio systems, such as the single and double-sideband systems, is a matter of considerable difficulty. However, as a practical matter, the percentage of increased commercial time and the increased satisfaction which a customer may obtain are of great interest. Three types of tests have been used in the past for rapidly obtaining information on the performance of radio circuits. They are: (*a*) determining the signal-to-noise ratio, (*b*) articulation tests, and (*c*) observations of circuit merit.

A measurement of the signal-to-noise ratio is made by modulating the transmitter a given amount and measuring the tone at the receiving point. The tone is then removed and the noise measured with the same equipment. When fading conditions are severe a considerable degree of skill is needed to obtain consistent measurements.

Articulation tests may be made in the manner which has been described by Fletcher and Steinberg.[4] They may consist of the reading and recording of meaningless syllables, carefully chosen words inserted

[3] U. S. Patent No. 1,959,449.

[4] "Articulation Testing Methods," H. Fletcher and J. C. Steinberg, *Bell System Technical Journal*, October, 1929.

in a variety of sentences, or a simple list of words chosen at random and inserted in a common phrase.

In the routine operation of the transatlantic channels, the operators record a value of "circuit merit" which is a composite figure representing the operator's judgment of the commercial value of the circuit. All three of these types of test were used in comparing the single and double-sideband systems.

All observations were made on 9,790 kc. with an audio-frequency band of from 250 to 2,800 cycles. The carrier during single-sideband transmissions was 10 db in amplitude below one of two equal side frequencies which loaded the transmitter to its maximum amplitude capacity. Only a single tone was used to modulate the transmitter when measuring signal-to-noise ratios. The degree of modulation of

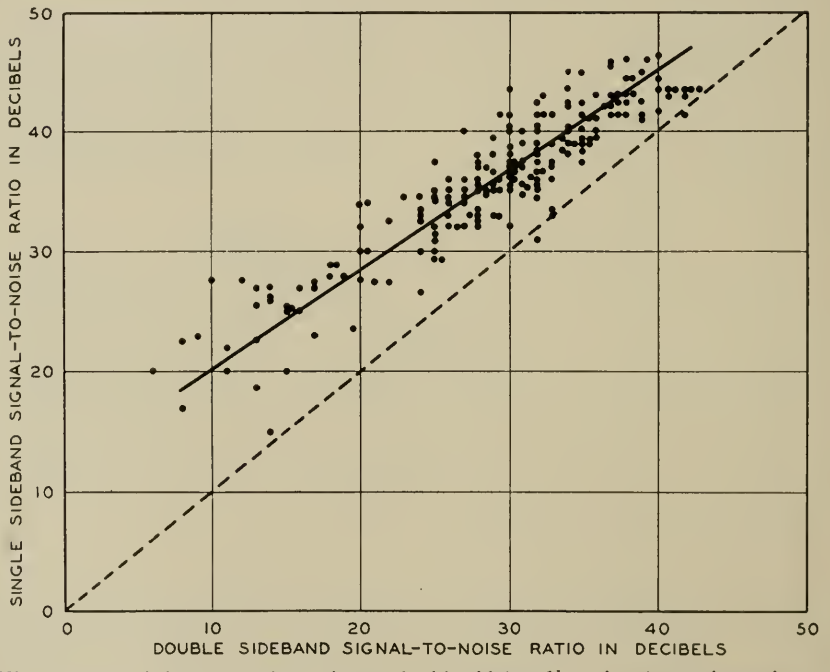


Fig. 9—Plot of signal-to-noise ratios on double sideband vs. signal-to-noise ratios on single sideband.

the double-sideband comparison signal was 45 per cent when tone modulation was used. Speech modulation was made the same as for two tones for both double and single-sideband transmissions. Directional antennas were used for both transmitting and receiving. Successive observations were made on single and double-sideband trans-

missions for 9-minute intervals. A reconditioned carrier was used at the receiver most of the time on account of the time required to synchronize the local oscillator when changing from double to single-sideband reception. The signal-to-noise ratio as well as the articulation was found to be the same for either reconditioned or local carrier except when the fields were very low, at which times the local carrier was found to be more satisfactory. Since it was convenient to use a slightly different degree of modulation on double-sideband than on single-sideband transmissions and the filter on the single-sideband receiver passed only 1.2 db less noise than the double-sideband receiver rather than the theoretically possible 3 db, a theoretical difference of 8.1 db instead of 9 db in signal-to-noise ratio was to be expected.

Each point shown on Fig. 9 represents the signal-to-noise ratio which was observed on the single-sideband system at a particular

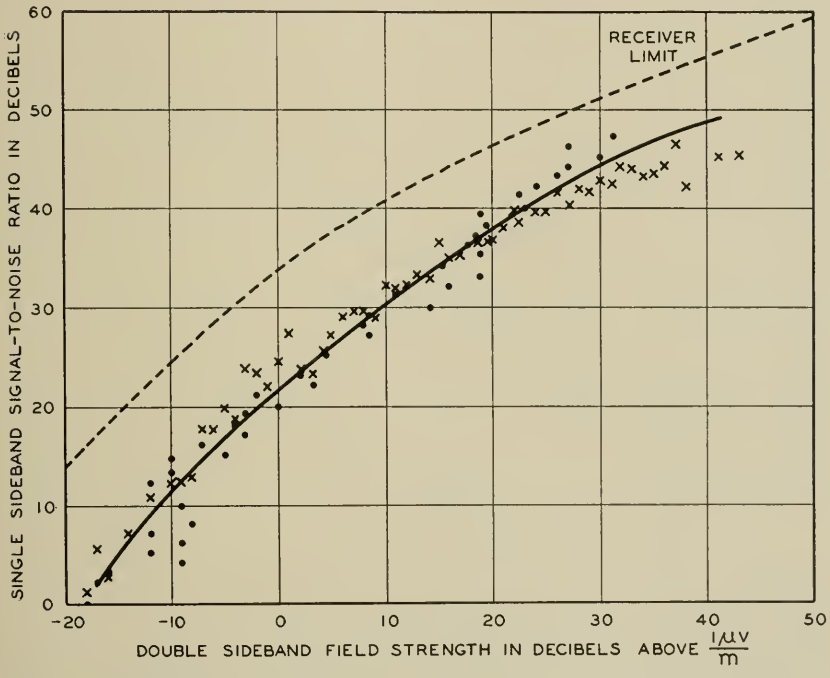


Fig. 10—Plot of signal-to-noise ratios on single sideband vs. field strength.

period plotted against the average of the preceding and succeeding values of signal-to-noise ratio measured on the double-sideband system. It will be seen that when the signal-to-noise ratio on the double-sideband system was 10 db the average signal-to-noise ratio

on the single-sideband system was 10 db higher, and when the signal-to-noise ratio on the double-sideband system was 40 db the average signal-to-noise ratio on the single-sideband system was 5 db higher. The lesser improvement with the single-sideband system for the higher signal-to-noise ratios was probably due to limitations in the maximum signal-to-noise ratio obtainable from the transmitting equipment.

Figures 10 and 11 are plots of the average signal-to-noise ratio versus field for the single and double-sideband systems. Only double-sideband fields were measured and the average single-sideband signal-to-noise ratios are plotted against the average of the preceding

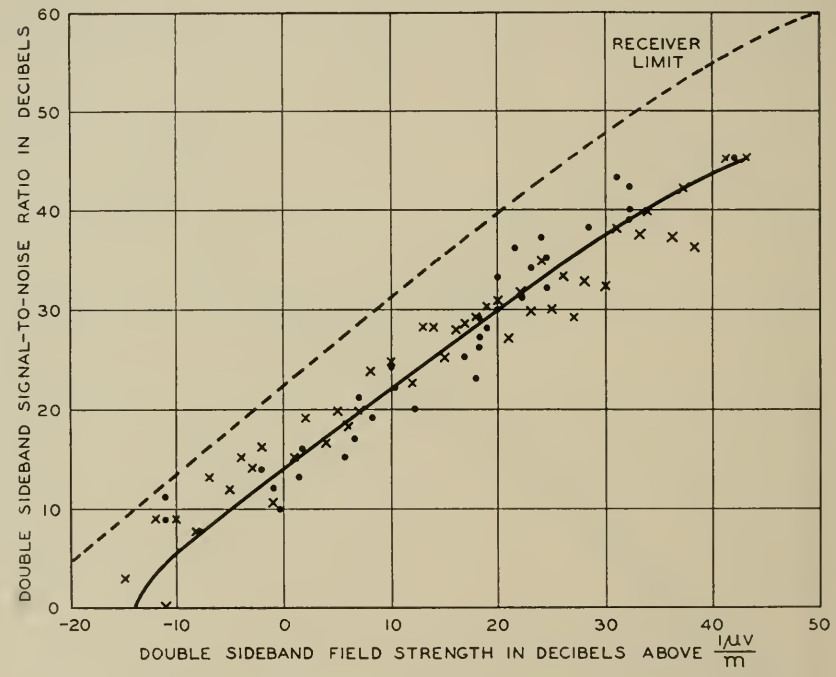


Fig. 11—Plot of signal-to-noise ratios on double sideband vs. field strength.

and succeeding double-sideband measurements. The crosses shown on the curve are the averages of all signal-to-noise ratios in 1 db intervals of field and the dots shown are averages of all fields obtained when the signal-to-noise ratio lay within 1 db intervals. The dotted lines represent the maximum signal-to-noise ratio which the receivers will give for various values of field. It is seen that on the average the set-noise was not the limiting factor determining the signal-to-noise ratios obtained.

Upon occasions, advantages considerably higher than the average were obtained for the single-sideband system. Reeves [5] has shown that at times the two sidebands of a double-sideband radio system are likely to be shifted in phase relative to each other and the carrier in such a manner that the demodulated audio-frequency components add at random rather than directly in phase. Under such circumstances the received signal-to-noise ratio of the transmissions would be reduced by 3 db, and in comparison with the single-sideband system the latter would show a correspondingly greater improvement. Further, under bad fading conditions, some advantage might be expected from using a receiver in which provision is made to insure an adequate carrier in the second detector at all times. The single-sideband receiver used in these tests had such provision while the double-sideband receiver did not.

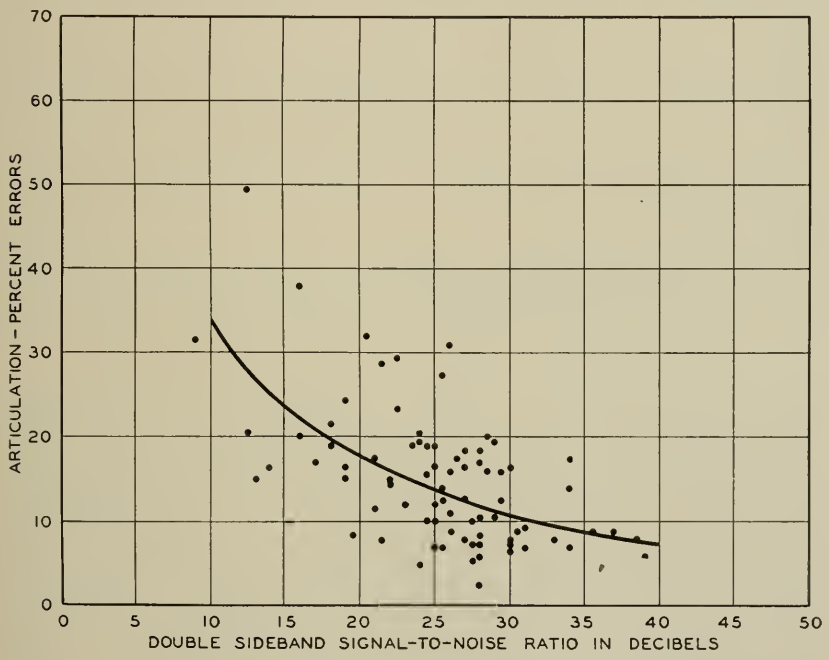


Fig. 12—Plot of per cent articulation errors on single-sideband vs. double-sideband signal-to-noise ratios.

The articulation of the two systems was compared by using words, which averaged approximately three syllables, taken at random from the dictionary. They were inserted in the phrase "Write down

[5] *Journal of the Institution of Electrical Engineers*, September, 1933, page 245.

. . . ". Native English callers were used at the transmitting end of the circuit almost exclusively and experienced articulation observers were used at the receiving end. Figures 12 and 13 show the articu-

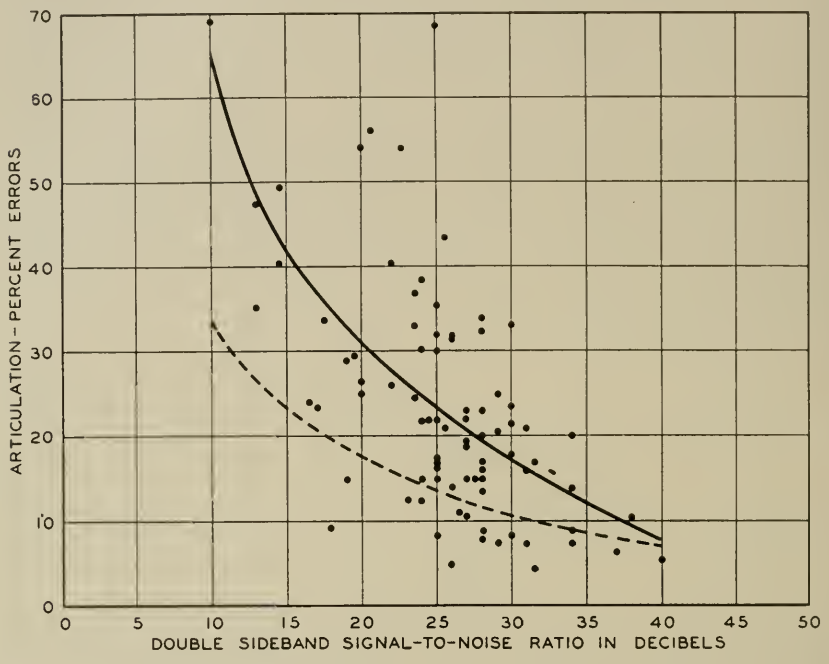


Fig. 13—Plot of per cent articulation errors on double-sideband vs. double-sideband signal-to-noise ratios.

lation errors observed on the single and double-sideband transmissions plotted against the signal-to-noise ratios measured on the double-sideband receiver. When plotting the single-sideband data, the average of two successive signal-to-noise ratio readings on double-sideband was taken as the signal-to-noise ratio for plotting the intervening single-sideband observation. The improvement due to the use of the single-sideband system expressed in decibels is the difference in the abscissæ of the two curves for a given ordinate. The second curve of Fig. 12 has been dotted in on Fig. 13 to facilitate the comparison of the two systems. The improvement is seen to average about 8 db for intermediate values of signal-to-noise ratio.

Figures 14 and 15 show the circuit merits obtained on the single and double-sideband systems respectively plotted against the field strength measured on the double-sideband receiver. A circuit having a merit of 5 is an extremely good circuit, while a circuit having a merit

of 3 is only just commercial and one having a circuit merit of 2 is useful only as an order wire. It will be noted that the difference between the curves for a circuit merit of 3 is about 8 db, for a circuit

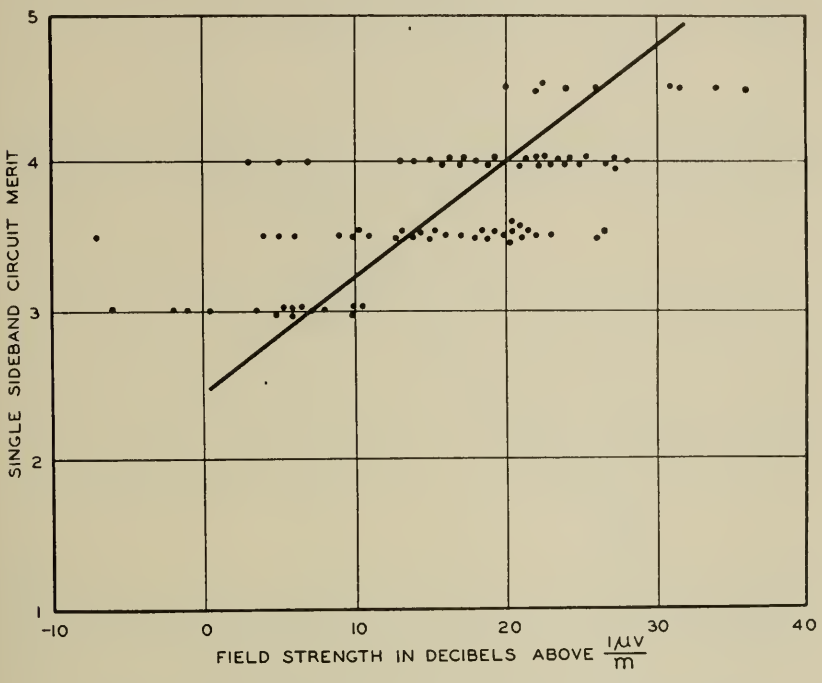


Fig. 14—Plot of circuit merit vs. field strength on single sideband.

merit of 4 about 5.5 db, and for a circuit merit of 5 about 4.5 db. This is in fair agreement with the signal-to-noise and articulation data.

The comparison of two circuits in this manner is undoubtedly of value if the observations extend over a period of time and if the individual observations are separated by a sufficient interval. When the observations are spaced at short intervals, however, the observer is bound to be influenced by the previous observation and it seems likely that the resulting comparison may be considerably in error. For instance, the observer may notice a small difference in circuit merit and consequently consistently rate one circuit a half point higher than the other, when the actual difference might be nearer to ¾ of a point. For this reason it is believed that the comparison of the two systems by means of circuit merit gives a less accurate result than by either signal-to-noise or articulation tests.

Outside of the general observation that, as might be expected, the improvement in signal-to-noise ratio at times when the circuit was poor was greater than at times when the circuit was good, no particular

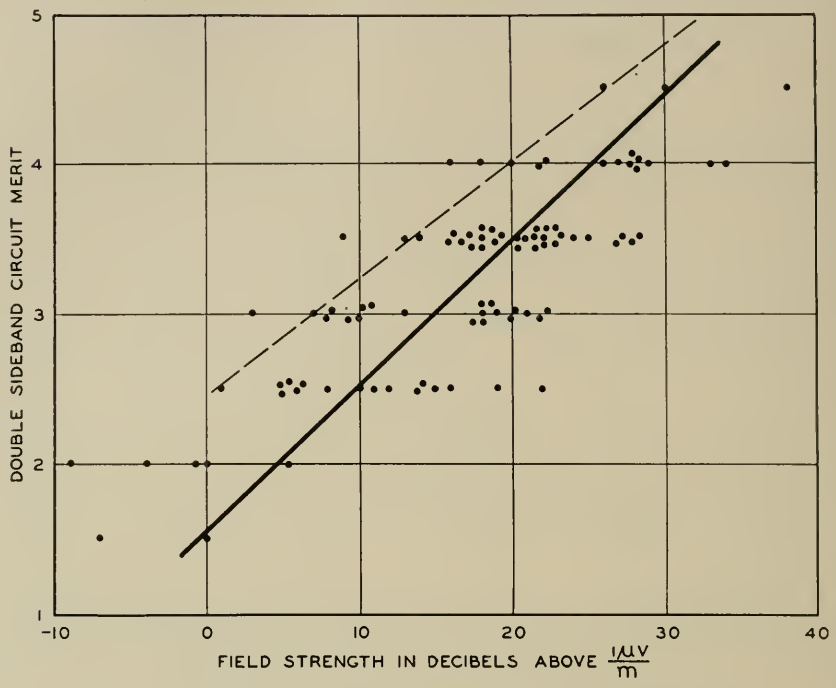


Fig. 15—Plot of circuit merit vs. field strength on double sideband.

connection was found between the improvement obtained by the use of single-sideband and transmission conditions. Only one magnetic storm of any consequence occurred during the test and the transmission was so poor on that day that no results were obtained, and, therefore, no conclusions can be drawn as to the effect of magnetic storms.

Mutual Impedances of Parallel Wires *

By RAY S. HOYT and SALLIE PERO MEAD

This is a theoretical paper relating to circuits of straight parallel wires traversed by alternating currents under such conditions (frequency of the alternating currents, diameter and spacing of the wires) that the resulting non-uniformity of the current distribution is sufficient to play an important part in determining the mutual and self impedances. The paper deals primarily with the mutual impedances; but incidentally the self impedances are dealt with almost as fully, except that no numerical calculations are made for them.

Part I is mainly a discussion of the physical nature of the mutual and self impedances in the generalized manner necessitated by the non-uniformity of the current distribution. It deals with wires which are short enough compared with the wave-length so that the complicating effects of propagation are negligible and so that the current in each wire can be regarded as an aggregate of filamentary currents.

Part II establishes, by recourse to electromagnetic wave theory, calculation formulas for the mutual and self impedances per unit length of a pair of long straight parallel transmission circuits forming a square array. Values of the mutual impedance are calculated over a frequency-range of 1 to 1000 kilocycles per second, for three cases of the circuits, and are compared with measured values.

Introduction

THE concept of the mutual impedance per unit length between two straight parallel filamentary conductors is well understood by engineers, and its calculation formula is simple. This mutual impedance is a pure reactance (directly proportional to the frequency), the induced electromotive force being in phase quadrature with the inducing current.

In the case of open-wire circuits, even when operating with carrier currents of very high frequency, the mutual impedance can be calculated with high accuracy by regarding the wires as filamentary.

For cable circuits, however, the foregoing statement is not true, because of the close juxtaposition of the wires. In such circuits the wires may be termed "thick," meaning that their diameter is appreciable compared with their interaxial separation. Depending in a complicated manner on the conductivity, permeability, diameter, and interaxial separation of such wires, the frequency may easily be so high as to render the filamentary formulas for the mutual impedance of even straight wires quite inaccurate and unreliable. In such cases it is necessary to consider the current distribution over the cross-section

* The two parts of this paper are distinct, though complementary. Part I was written by Ray S. Hoyt, Part II by Sallie Pero Mead.

of the wires.[1] When this is done it is found that the mutual impedance comprises not only a reactance component, which is no longer proportional to the frequency, but also a resistance component, which does not vary in any simple way with the frequency. Both of these component departures of the mutual impedance from its simple filamentary value increase the difficulties of balancing out crosstalk, and the resistance component has also an important effect on the attenuation at carrier frequencies. These matters have recently assumed considerable importance on account of the rapidly increasing interest in the possibilities of communication transmission over non-loaded cable circuits with the aid of carrier currents having frequencies high compared with those of speech. As an approximate guide to the behavior of twisted circuits in cables the theory and formulas for straight wires, as developed in this paper, have proved to be of considerable service.

The present paper deals with the mutual impedances of two or more straight parallel wires from two aspects: In Part I the physical theory is developed and expounded. The current in a wire is there regarded as made up of an indefinitely large number of parallel filamentary current elements. On this basis it is shown (among other things) that the current distribution over the cross-section of each conductor is necessarily non-uniform, and that this non-uniformity gives rise to a mutual resistance term in the mutual impedance, besides a change in the mutual reactance term. In Part II electromagnetic wave theory is applied to develop formulas for the mutual and self impedances of a pair of long straight parallel transmission circuits in close juxtaposition. Calculations of the mutual impedance made with these formulas over a very wide range of frequencies (1 to 1000 kilocycles per second) are found to be in very satisfactory agreement with available experimental results.[2] In both parts of the paper an endeavor has been made to bring engineering concepts and formulas into closer relationship with electromagnetic theory.

[1] The convenient term "proximity effect" when applied to the distribution of the current over the cross-section of a given conductor means the deviation of this distribution from the "intrinsic distribution," the latter meaning the distribution when the given conductor is far enough from all other conductors so that the distribution in it is sensibly unaffected by them.

When the given conductor is a straight uniform wire of circular cross-section, its "intrinsic distribution" is of course axially symmetrical.

Not every axially symmetrical distribution is the same as the corresponding intrinsic distribution, as is evidenced by the case of two coaxial conductors, where the proximity effect in the outer conductor may be large although the current is axially symmetrical in each conductor.

[2] See the paper by R. N. Hunter and R. P. Booth, in the April issue of this *Journal*, entitled "Cable Crosstalk—Effect of Non-Uniform Current Distribution in the Wires," which includes the results of some rather extensive sets of measurements of the mutual impedance of straight wire circuits, and also of twisted circuits in cables, and a brief physical discussion with particular regard to the effect of non-uniform current distribution.

PART I

Physical Theory

The Physical System; Analysis of the Wire Currents into Filaments

Since the mutual impedance between any two parallel circuits can be expressed wholly in terms of the mutual impedances between the various wires composing the circuits,[3] it will suffice in Part I to discuss the mutual impedance between the two wires A and B in Fig. 1. These are each of uniform cross-section, but they need not be alike in cross-sectional shape and area nor in material.

The wires in Fig. 1 will be assumed very long compared to the dis-

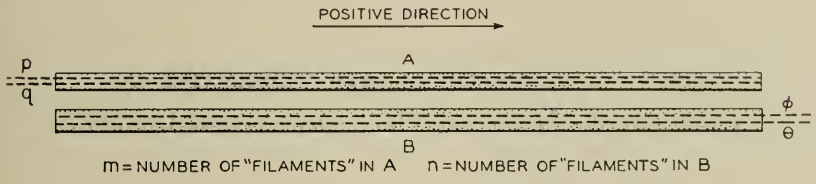


Fig. 1—Two "thick" straight parallel wires. p, q designate any two "filaments" of wire A; ϕ, θ any two of B.

tance between them, so that the end-effects [4] in the current distribution will be negligible, yet short enough compared with the wavelength so that the charging current will be a negligible fraction of the total current and therefore the current in each wire of sensibly the same value throughout its length.[5] These circumstances enable the current in each wire to be treated as an aggregate of filamentary currents which are purely longitudinal, and correspondingly enable the mutual and self impedances of the wires to be described and formulated in terms of the mutual and self impedances of such filaments, thus correlating well with the familiar treatment of a system of fine parallel wires. This treatment by analysis into filaments has been chosen

[3] For example, the mutual impedance Z_{ab} between two circuits a and b, of which a comprises wires 1 and 2 and b comprises 3 and 4, is given by $Z_{ab} = Z_{13} - Z_{14} - Z_{23} + Z_{24}$. However, since the wires are in general "thick," the value of each mutual impedance (also each self impedance) must depend on the presence of all four of the wires.

[4] These consist in the currents not being purely longitudinal near the ends of the wires.

[5] Negligibility of the charging current does not by any means imply that the distributed charges on the surfaces of the wires are negligible as regards the voltages which they produce, for extremely small charging currents suffice to establish charges which can produce relatively large voltages.

For a discussion of this very important fact and other underlying concepts of circuit theory, the reader is referred to a paper by John R. Carson, "Electromagnetic Theory and the Foundations of Electric Circuit Theory," published in this *Journal* for January, 1927.

because it lends itself well to a physical exposition and to the derivation of the simple formulas needed in that exposition.

Since in general the various filamentary currents in a wire are not in phase the total, or resultant, current in the wire, which is the complex algebraic sum of the filamentary currents, must be less than the arithmetic sum of the filamentary currents. An extreme instance of this fact is presented by a wire, short compared with the wave-length, which is on open circuit and is situated in the field due to other currents; for although the total, or resultant, current traversing any cross-section of this open wire must be zero, the individual filamentary currents are not zero.

The Two Parts of a Voltage, and Their Resultant [6]

For clearness in describing and formulating the mutual and self impedances of the wires, even when these are filamentary, it is necessary to recognize that the voltage along any specified path (which may, in particular, be a filament in a conductor) is in general the sum, or resultant, of two voltages which are simultaneously present along the path, namely the voltage due to all charges, and the voltage due to all currents; for brevity, these two parts of the total voltage will be called merely the "charge voltage" and the "current voltage" respectively—or, somewhat more fully, the "charge-produced voltage" and the "current-produced voltage." They will be denoted by V and U respectively, and their resultant by W, so that $W = V + U$.

The two parts of a voltage have the sharply contrasting properties constituting principles "1" and "2" in the following set of four principles, all of which are of much importance for the understanding of electric circuit theory and transmission theory.

1. A "charge voltage" (V) has exactly the same value along every path between any two fixed points, and hence is zero around every closed path.

2. A "current voltage" (U) has in general unequal values along any two different paths between any two fixed points, the difference in these values being accounted for by the time rate of change of the magnetic flux in the space between the two paths; thus a "current voltage" is in general not zero around a closed path.

3. The total, or resultant, voltage (W) must evidently have the same properties as the "current voltage" (U) in " 2."

4. For any current filament f in a conductor the product of the resistance R_f of the filament and its current I_f is, by Ohm's law, equal

[6] This section is based on certain fundamentals of electromagnetic theory summarized in an appendix placed at the end of the whole paper.

to the total, or resultant, voltage along the filament; that is $R_f I_f = W_f = V_f + U_f$. Hence $V_f = R_f I_f - U_f$, which is the most convenient form in many applications, particularly those involving inductances.

Before taking up (in the next section) the more complicated subject of the mutual and self impedances of "thick" wires, some of the foregoing principles will be illustrated by applying them to the simple system represented by Fig. 2, which comprises two filamentary wires

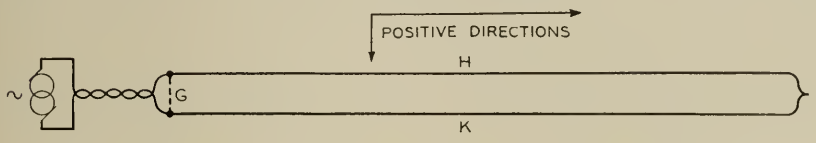


Fig. 2—An illustrative circuit of two "filamentary" wires, H and K.

H and K forming a loop. The dotted line G is merely a geometrical path traced directly between the initial terminals of the two wires.

The first form of principle "1," when applied to the two separate paths G and HK between the initial terminals, gives:

$$V_G = V_H + (-V_K). \tag{1}$$

The following two equations result from the last form of principle "4," when supplemented by the definitions of the self and mutual inductances of filamentary wires, which enable the U's to be expressed in terms of the I's:

$$V_H = R_H I_H - U_H = R_H I_H + i\omega L_H I_H + i\omega L_{HK} I_K, \tag{1a}$$
$$V_K = R_K I_K - U_K = R_K I_K + i\omega L_K I_K + i\omega L_{KH} I_H, \tag{1b}$$

where L_H denotes the self inductance of wire H, L_{HK} the mutual inductance[7] between H and K, $\omega = 2\pi$ times the frequency, and $i = \sqrt{-1}$. Further, on account of the choice of positive directions shown in Fig. 2, $I_K = -I_H$. Accordingly, replacing I_K by $-I_H$ and substituting the resulting values of V_H and V_K into equation (1) gives:

$$V_G = (Z_{HH} + Z_{KK} - 2Z_{HK})I_H, \tag{1c}$$

where $Z_{HK} = i\omega L_{HK} = i\omega L_{KH} = Z_{KH}$, $Z_{HH} = R_H + i\omega L_H$, etc. It will be observed that while the "current voltages" have been eliminated (through the self and mutual inductances and the currents), the "charge voltages" remain and play the role of "applied voltages."

For wire H (Fig. 2), the equation (1a), when written in the form

$$R_H I_H = W_H = V_H + U_H = V_H - i\omega(L_H - L_{HK})I_H, \tag{2}$$

[7] The first subscript designates the "disturbed" wire, the second the "disturbing" wire ("inducing" wire).

and its "vector diagram" (Fig. 3) both show that, in the limiting case of a perfectly conducting wire ($R_H = 0$), V_H and U_H would exactly balance each other, their values being exactly equal and opposite;

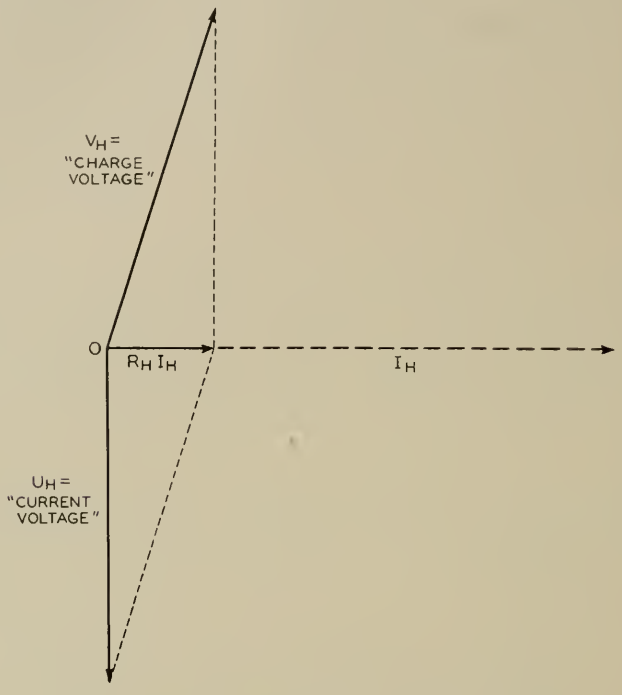


Fig. 3—Vector diagram relating to wire H of Fig. 2.

and that, in the case of an actual wire of low resistance, V_H and U_H nearly balance each other, their values being nearly equal and nearly opposite, so that their resultant $W_H = R_H I_H$ is a small residual voltage, although V_H and U_H individually may be very large compared with W_H.

The Mutual and Self Impedances

The mutual impedance (and similarly the inductive part of each self impedance) of the wires A and B (Fig. 1) cannot be defined as the negative of the voltage induced in either by unit current in the other, because the induced voltage necessarily has unequal values along the various filaments of which the disturbed wire may be regarded as composed. Thus in general this definition, which might be called the elementary definition, is applicable only to the individual filaments

composing the wires, or to wires which themselves are fine enough to be regarded as filamentary.[8]

The self and mutual impedances of the wires could be formulated in terms of the self and mutual impedances of their filaments by eliminating the filamentary currents. However, without such elimination it is possible to obtain a type of formulation which is simpler, more compact, and in many ways more enlightening, as will now be shown by aid of the foregoing section:—Considering, in Fig. 1, any filament p of wire A, let R_p denote the resistance of that filament and I_p the current through it; then, by Ohm's law, the product R_pI_p is equal to the total voltage W_p along p. But W_p is the resultant of the "current voltage" U_p due to all of the wire currents, and the "charge voltage" V_p due to all of the charges; however, V_p is equal to V_A, the "charge voltage" along wire A as a whole, since the "charge voltages" along all of the various filaments in a wire must be equal. These various facts are expressed by the equation [9]

$$R_pI_p = W_p = U_p + V_p = U_p + V_A. \tag{3}$$

But, from the definitions of the filamentary self and mutual impedances,

$$U_p = -Z_pI_p - \sum_{q \neq p} Z_{pq}I_q - \sum_{\phi} Z_{p\phi}I_\phi, \tag{4}$$

where $Z_p = i\omega L_p$ denotes the inductive part of the self impedance $Z_{pp} = R_p + i\omega L_p$ of filament p, $Z_{pq} = i\omega L_{pq}$ the mutual impedance[7] between p and any other filament q of A, and $Z_{p\phi} = i\omega L_{p\phi}$ that between

[8] The case where one wire is "thick" and the other filamentary is on the border line, the elementary definition of the mutual impedance being applicable when the "thick" wire is the disturbing wire but not when it is the disturbed wire.

The generalized definition, to be formulated later herein, must of course be such that the mutual impedance between any two wires will have exactly equal values in the two directions.

[9] The distribution of V over the cross-section of the wire being uniform, equation (3) shows that if U is non-uniform I also must be, and conversely. This is exemplified in skin effect and proximity effect.

By averaging the whole set of equations, of which (3) is typical, relating to all of the filaments in wire A, and denoting the total current in this wire by I and its *direct current* resistance by R,[0] we find that

$$R^0I = \overline{W} = \overline{U} + \overline{V} = \overline{U} + V,$$

a bar indicating an average value over the cross-section. The relation $R^0I = \overline{W}$ appears sufficiently useful and interesting to justify its enunciation in the form of a theorem, as follows: *When the varying current in a single piece of uniform wire, which may have any cross-sectional shape, has sensibly the same total value I throughout the length of the wire, whose direct current resistance is R^0, the product R^0I is equal to the cross-sectional average $\overline{W}$ of the total, or resultant, voltage W along the wire between its two ends.* For a wire which is fine enough to be regarded as filamentary, the above equation reduces to $RI = W = U + V$. For a wire carrying direct current, it reduces to $R^0I = \overline{V} = V$.

filament p of A and filament ϕ of B. On substituting (4) into (3) we get for filament p the "voltage equation:"

$$V_p = \sum_q Z_{pq} I_q + \sum_\phi Z_{p\phi} I_\phi = V_A. \tag{5}$$

Next we multiply this equation through by I_p, add together the m such resulting equations corresponding respectively to the m filaments of wire A, and introduce the condition that the sum of the filamentary currents in A is equal to I_A. Finally, we divide the resulting equation through by I_A and denote the current-ratios I_p/I_A, I_q/I_A, I_ϕ/I_B by J_p, J_q, J_ϕ respectively, each J thus denoting the ratio of a filament current to the total current in the wire to which that filament belongs, so that J may be called a "relative filamentary current." We thus get the equation

$$V_A = I_A \sum_p \sum_q Z_{pq} J_p J_q + I_B \sum_p \sum_\phi Z_{p\phi} J_p J_\phi. \tag{6}$$

Comparison of this with the equation

$$V_A = Z_{AA} I_A + Z_{AB} I_B, \tag{6a}$$

which is the "voltage equation" for wire A as a whole, yields the following formulas for the self impedance Z_{AA} of wire A and the mutual impedance Z_{AB} to A from B (Fig. 1):

$$Z_{AA} = \sum_p \sum_q Z_{pq} J_p J_q, \tag{7}$$

$$Z_{AB} = \sum_p \sum_\phi Z_{p\phi} J_p J_\phi. \tag{8}$$

Similarly, for wire B,

$$Z_{BB} = \sum_\phi \sum_\theta Z_{\phi\theta} J_\phi J_\theta, \tag{9}$$

$$Z_{BA} = \sum_\phi \sum_p Z_{\phi p} J_\phi J_p, \tag{10}$$

Z_{BA} denoting the mutual impedance to B from A. It will be recalled that p, q designate any two typical filaments of wire A, and ϕ, θ any two of B.

The presence of the relative filamentary currents (the J's) in these equations accounts for the fact that the self impedance of a wire depends on the current-distribution over its cross-section, and the mutual impedance between two wires on the current-distributions over their cross-sections. The self and mutual impedances of two wires, such as A and B, must thus depend on the currents in any other wires that may be present, because the voltages induced in A and B by these other currents will partly determine the current-distributions in A and B. Although the *values* of the summation expressions in equations (7) to (10) depend on the currents in any other wires that

may be present, nevertheless the *forms* of these expressions do not. Thus, so far as the *forms* of the expressions are concerned, the two wires A and B need not be alone but may be any two of a system of parallel wires $A, \cdots, B, \cdots, D$ carrying arbitrary currents $I_A, \cdots, I_B, \cdots, I_D$ respectively; still further, A and B may even be any two of the parallel longitudinal parts of which any wire may arbitrarily be regarded as composed.

Equations (8) and (7) respectively show that the mutual and self impedances of "thick" wires have the following significance:

The mutual impedance between two wires is equal to the sum of the weighted mutual impedances from every filament in one wire to every filament in the other, the weighting factor of any filamentary mutual impedance being the product of the corresponding two relative filamentary currents (the J's).[10]

The self impedance of a wire is equal to the sum of the weighted mutual impedances from every filament to every other filament, including the weighted mutual impedance from every filament to itself, the weighting factor of any filamentary mutual impedance being the product of the corresponding two relative filamentary currents (the J's).[11]

Or, more briefly, *the self impedance of a wire is equal to the sum of the weighted mutual impedances from every filament to every other filament and to itself.*

Several matters of interest regarding the "thick" wires A and B (Fig.1) will next be discussed, mainly from the physical viewpoint corresponding to equations (7) to (10).

Reciprocity of the Two Mutual Impedances

Since $Z_{p\phi}$ and $Z_{\phi p}$ are unquestionably equal, because they relate to filaments, comparison of formulas (8) and (10) shows that the mutual impedances Z_{AB} and Z_{BA} between the wires A and B are equal. The same conclusion follows also from the first italicised paragraph above, which is based on formulas (8) and (10).

Complex Nature of the Mutual and Self Impedances

Although every mutual impedance between different filaments is a pure reactance which is directly proportional to the frequency, nevertheless the mutual impedance Z_{AB} between the wires A and B has in

[10] In other words, the mutual impedance of two wires is equal to the sum of the weighted mutual impedances between all of the various filaments taken in pairs each pair consisting of one filament from each wire.

[11] In other words, the self impedance of a wire is equal to twice the sum of the weighted mutual impedances between all of the various filaments taken in pairs, plus the sum of the weighted self impedances of the filaments. (The weighting factor of the self impedance of any filament is evidently the square of its relative filamentary current.)

general not only a reactance component which is not quite proportional to the frequency, but also a resistance component which does not vary in any simple way with the frequency. On the basis of formula (8) these facts are to be accounted for by the consideration that in general the various filamentary currents in a wire are not only not in phase but have no simple phase relations. Thus if (8) is written in the form

$$Z_{AB} = i\omega \sum_{p} \sum_{\phi} L_{p\phi} J_p J_\phi = R_{AB} + i\omega L_{AB}, \tag{11}$$

then R_{AB} is not zero, and R_{AB} and L_{AB} vary with ω although $L_{p\phi}$ does not.

That the self impedance Z_{AA} of the wire A is not a pure reactance can be accounted for similarly, with the additional reason that the self impedance of each filament is complex, because of its resistance. Thus if (7) is written in the form

$$Z_{AA} = \sum_{p} (R_p + i\omega L_p) J_p^2 + i\omega \sum_{p} \sum_{q \neq p} L_{pq} J_p J_q = R_A + i\omega L_A, \tag{12}$$

then R_A is not zero, and R_A and L_A vary with ω although R_p, L_p, L_{pq} do not.

It may be noted that in the idealized case of perfect conductivity the mutual and self impedances of the wires would be pure reactances and directly proportional to the frequency; for in this case the filamentary currents in any wire would all be in phase and their distribution would be independent of the frequency. (The current distribution would be the same as the charge distribution and hence purely superficial.)

Case of Negligible Proximity Effect

The case here considered is that in which the wires A and B are of circular or of annular cross-section (but external to each other) and are far enough apart so that the proximity effect [1] is negligible and so that therefore the current distribution over the cross-section of each wire is sensibly axially symmetrical.

For this particular case the mutual impedance $Z_{AB} = Z_{BA}$ is *not* complex but is pure reactance, being equal to the mutual impedance $Z_{A'B'} = Z_{B'A'}$ between two filamentary wires A' and B' having the same interaxial spacing as the given "thick" wires A and B. Although this statement is clearly true when only one of the wires is "thick," it really needs a proof in the general case where *both* are "thick." The following simple proof depends on the fact that everywhere (except near its ends) outside of a long straight wire, of circular or of annular section, carrying an axially symmetrical current the magnetic field produced by that current is the same as though the current were con-

centrated in the axis, and the proof also utilizes the reciprocity relation for the mutual impedances in the two directions between the two wires involved; thus,[7]

$$Z_{AB} = Z_{AB'} = Z_{B'A} = Z_{B'A'} = Z_{A'B'},$$
$$Z_{BA} = Z_{BA'} = Z_{A'B} = Z_{A'B'} = Z_{B'A'}.$$

The statement at the beginning of this paragraph is thus proved.

PART II

Mathematical Theory and Calculations

The theoretical investigation of the self and mutual impedances per unit length of two long parallel pairs of wires in space is an application

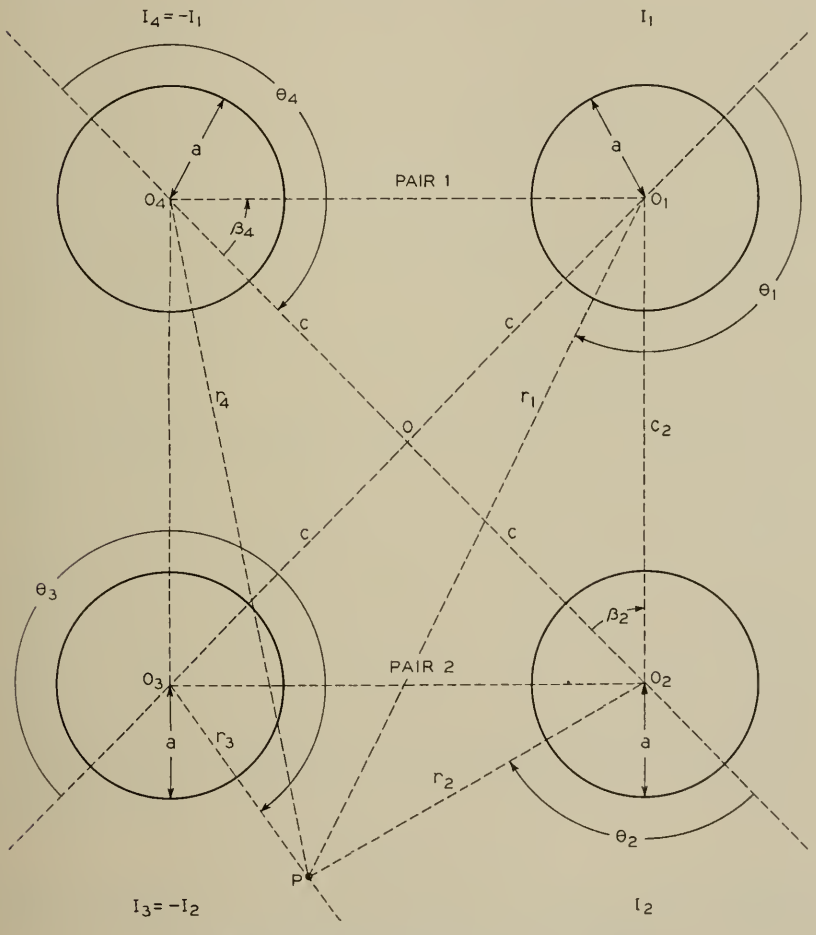


Fig. 4—Cross-sectional diagram of 4-wire system.

of two dimensional wave propagation theory. The specific case of four wires in a square array was selected as the basis of a comparison of measured and theoretical values of mutual inductance. The configuration with four equal wires is shown in cross section in Fig. 4 where wires No. 1 and No. 4, centered at O_1 and O_4 and carrying currents I_1 and $-I_1$, respectively, form the first pair or primary and wires No. 2 and No. 3 at O_2 and O_3 and carrying currents I_2 and $-I_2$, respectively,

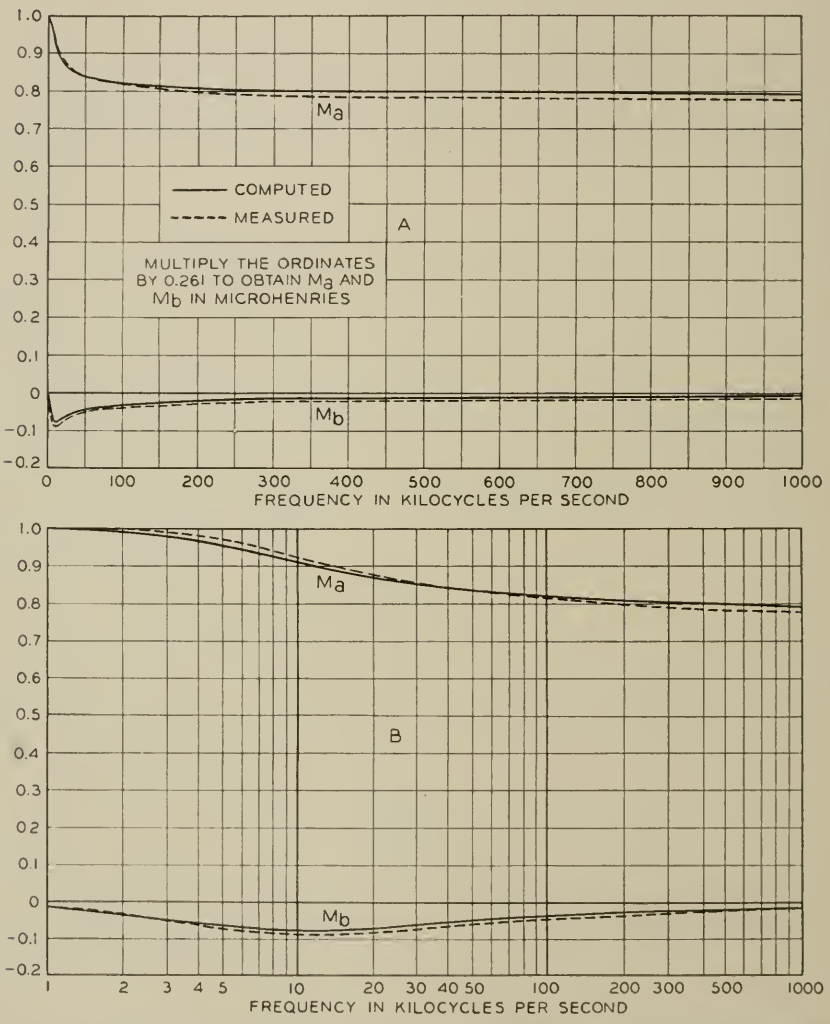


Fig. 5—Real and imaginary components, M_a and M_b, respectively, of the mutual inductance between a pair of No. 10 A.W.G. wires and a pair of filaments. Length of wires = 74 inches; interaxial separation = 0.14 inch.

form the second pair or secondary. We have

$$O_1O_3 = O_2O_4 = 2c$$

and

$$O_1O_2 = O_1O_4 = \sqrt{2}c.$$

The notation for the dimensions and coordinate systems is shown in Fig. 4. The theoretical values of mutual inductance are calculated from the geometry and electrical constants of this system by means of the formulas which will be derived herein, while the measured values are those obtained for this system by R. N. Hunter and R. P. Booth.[2]

Numerical Results

A close agreement between the values of mutual inductance computed on the basis of the approximate formulas derived below and the experimental results is shown by the curves in Figs. 5 and 6. In fact, for No. 18 gauge wires, in which case the proximity effect is comparatively small, the computed and measured values are indistinguishable in Figs. 6A and 6B. Evidently the error introduced by the fact that actually the line is comparatively short while theoretically we assume it of doubly infinite length, is inappreciable. The drawings give relative values of the real and imaginary components of the complex mutual inductance $M = M_a + iM_b$, for 74 inch lengths of wires with vertical and horizontal interaxial spacing of 0.14 inch over a frequency range of 1 to 1000 kilocycles per second. (The value 0.565×10^{-3} emu. is assumed for the conductivity of the wires and unit permeability for both wires and dielectric.) The solid curves represent computed values and the dotted curves measured values. The values shown are the ratios of M_a and M_b to the value of M_a at 1 kilocycle. In Figs. 5A and 6A the frequency scale is linear, while in Figs. 5B and 6B it is logarithmic. The computed curves of Fig. 5 (obtained from formula (13) below) assume a pair of No. 10 A.W.G. wires (0.102 inch in diameter) as the primary and a filamentary secondary. Actually the secondary was a pair of No. 28 A.W.G. wires. In the two cases shown in Fig. 6, computed from formula (14) below, both pairs of conductors are of the same size; namely, No. 10 and No. 18 A.W.G. wires, respectively (the latter being actually 0.0410 inch in diameter).

It will be observed that we have the relation

$$M = Z_m/i\omega,$$

Z_m denoting the mutual impedance, $\omega/2\pi$ the frequency and i the

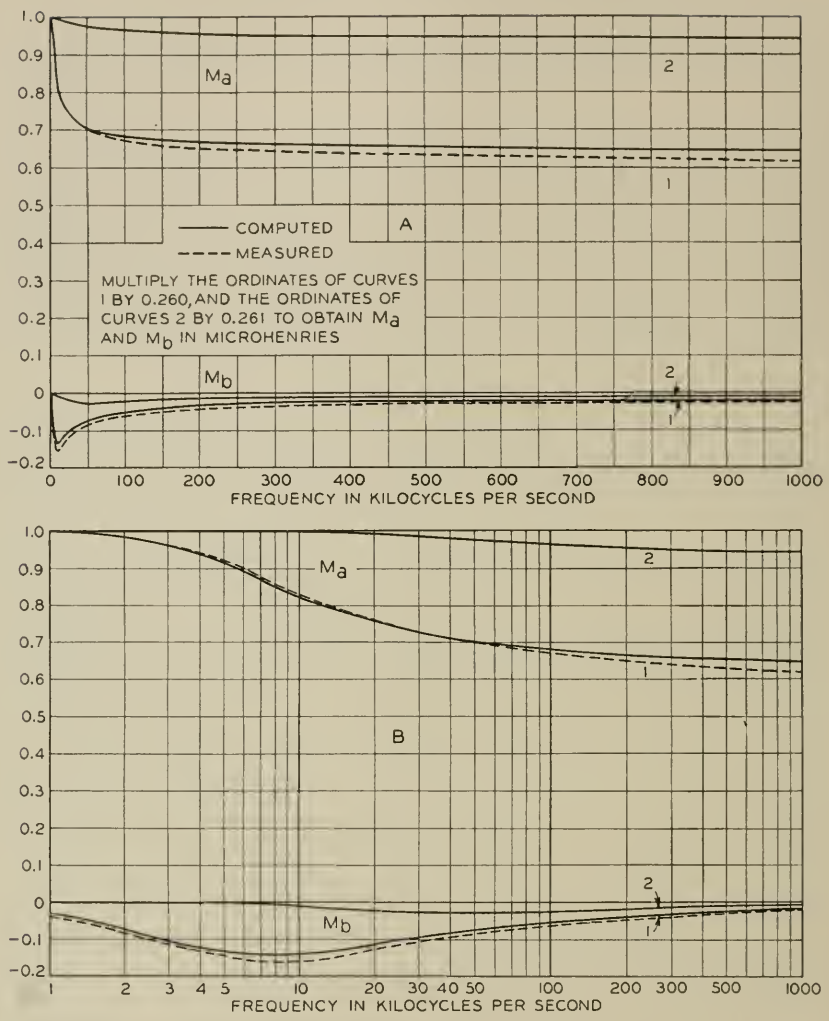


Fig. 6—Real and imaginary components, M_a and M_b, respectively, of the mutual inductance between (1) two pairs of No. 10 A.W.G. wires and (2) two pairs of No. 18 A.W.G. wires. Length of wires = 74 inches; interaxial separation = 0.14 inch.

imaginary. In the last step of the analysis below the mutual impedance of two circuits of wires of large [12] cross-section is derived by a method of successive approximations. A first approximation is obtained by assuming the field due to each wire of the second pair cir-

[12] The wires of a pair are to be considered "large" for a given frequency (f), provided the values of the radius (a), interaxial separation (d), conductivity (σ) and permeability (μ_c) are such that the magnitude

$$|(2a/d)(J_1(z)/J_0(z))^2|$$

is not small compared to unity. Here J_0 and J_1 are Bessel functions of the first kind of zeroth and first orders, respectively, and of complex argument, $z = ia\sqrt{4\pi\sigma\mu_c i\omega}$.

cularly symmetrical. Physically this is equivalent to assuming the concentration of the current on the axes of the secondary as if it were filamentary so that the proximity effect in this pair is eliminated. Thus equation (13) formulates the solution of the case which is represented in Fig. 5. Regarded as an approximation to the solution when both pairs of wires are of large cross-section, it will be seen that these values account for about 50 per cent of the departure of the final results from the d.c. value. (This is 0.261 microhenry for the square arrangement.) The second approximation (formula (14)) takes into account the circularly unsymmetrical components of the field due to the unsymmetrical distribution of current density in the wires of the secondary as well as of the primary and so adds the proximity effect due to the thickness of the secondary. A summary of the formulas for mutual inductance follows:

Formulas

With the notation

$\lambda = a/2c$
a = radius of wires in centimeters
$2c$ = diagonal interaxial separation of wires in centimeters
σ = conductivity of wires in emu.
$f = \omega/2\pi$ = frequency

and denoting by $M^{(0)}$ the complex mutual inductance per unit length of two circuits, one of wires of large cross-section and one filamentary or, from the other point of view, a first approximation to the mutual inductance of two circuits of wires of large cross-section and, by M, a second approximation to the latter, we have

$$M^{(0)} = 4(\log_e \sqrt{2} - k_1), \tag{13}$$

$$M = M^{(0)} - 4(k_1 - 7k_1^2 + 4k_2), \tag{14}$$

where

$$k_1 = \frac{\lambda^2\tau_1}{1 - 2\lambda^2\tau_1},$$

$$k_2 = \frac{\lambda^4\tau_2}{1 - 12\lambda^4\tau_2},$$

$$\tau_1 = 1 + i\frac{2}{x}\frac{\operatorname{ber}' x + i \operatorname{bei}' x}{\operatorname{ber} x + i \operatorname{bei} x},$$

$$\rightarrow 1 - \nu + i\nu \text{ when } x \rightarrow \infty,$$

$$\tau_2 = 1 - i\frac{8}{x^2} - \frac{4}{x}\frac{\operatorname{ber} x + i \operatorname{bei} x}{\operatorname{ber}' x + i \operatorname{bei}' x},$$

$$\rightarrow 1 - 2\nu + i2\nu \text{ when } x \rightarrow \infty,$$

$$x = a\sqrt{4\pi\sigma\omega} = \sqrt{2}/\nu,$$

$$\nu = 1/(2\pi a\sqrt{f\sigma}).$$

For values of ber x, bei x, ber′ x and bei′ x see Jahnke u. Emde, "Funktionentafeln." $M^{(0)}$ and M are given above in emu. per centimeter. We assume $M^{(0)}$ and M proportional to the length and multiply by 0.1880 to obtain michrohenries per 74 inches.

No assumptions with respect to frequency are made in formulas (13) and (14) but terms of the order of magnitude with respect to unity of $9k_1^3 \approx 9\lambda^6$ or smaller are neglected. That is, the accuracy of (13) and (14) is limited by the dimensions rather than by the frequency. But for frequencies of about 100 kilocycles or higher and not too small wires (that is, when $x \geqq$ about 10) formulas (13) and (14) may be expressed in interpretable form; namely,

$$M^{(0)} = M_a{}^{(0)} + iM_b{}^{(0)} \tag{15}$$

and

$$M = M_a + iM_b, \tag{16}$$

where

$$M_a{}^{(0)} = 4\left(\log_e \sqrt{2} - \frac{\lambda^2}{1-2\lambda^2} + \frac{\lambda^2}{1-4\lambda^2}\nu\right),$$

$$M_b{}^{(0)} = -4\frac{\lambda^2}{1-4\lambda^2}\nu,$$

$$M_a = M_a{}^{(0)} - 4\left(\frac{\lambda^2}{1-2\lambda^2} - 3\lambda^4\right) + 4\nu\left(\frac{\lambda^2}{1-4\lambda^2} - 6\lambda^4\right),$$

$$M_b = M_b{}^{(0)} - 4\nu\left(\frac{\lambda^2}{1-4\lambda^2} - 6\lambda^4\right).$$

The asymptotic values (when f or σ or both approach infinity) are, therefore,

$$M^{(0)} = 4\left(\log_e \sqrt{2} - \frac{\lambda^2}{1-2\lambda^2}\right), \tag{17}$$

$$M = M^{(0)} - 4\left(\frac{\lambda^2}{1-2\lambda^2} - 3\lambda^4\right) \tag{18}$$

and the d.c. value (when f approaches zero) is, of course,

$$M = 4\log_e \sqrt{2}.$$

Thus M is real (i.e., $M_b = 0$) when the frequency is either zero or infinite.

Formal Solution

The following derivation of these results is an application of the general method of calculating the self and mutual impedances in a system of parallel wires which is outlined in Section V of John R. Carson's paper "Rigorous and Approximate Theories of Wave Transmission along Wires," *B. S. T. J.*, Jan., 1928. This method of solution

has proved valuable in investigating a variety of problems in wave propagation along parallel conductors. Reference may be made to the paper itself for a fuller exposition of the underlying theory, much of which is omitted from the present analysis.

In terms of a vector potential A and a scalar potential V, the electric and magnetic forces E and H per unit length in the dielectric are given by the relations,

$$\begin{aligned} E &= -\operatorname{grad} V - i\omega A \\ \text{and} \qquad \mu H &= \operatorname{curl} A, \end{aligned} \tag{19}$$

where μ is the permeability of the dielectric. Assuming that the wave varies as $\exp(i\omega t - \gamma z)$ and putting $F = i\omega A_z$, the axial electric force E_z (omitting the subscript z) may be written

$$E = \gamma V - F, \tag{20}$$

where γ is the propagation constant per unit length of the system.

Now, as we shall show below, the electric force E inside of the conductors and the wave function F in the dielectric may be expressed as linear functions of the conductor currents. That is, at the surface of the jth conductor, for example, we may write

$$\begin{aligned} E_j &= e_{1j}I_1 + e_{2j}I_2 + e_{3j}I_3 + e_{4j}I_4 \\ \text{and} \qquad F_j &= f_{1j}I_1 + f_{2j}I_2 + f_{3j}I_3 + f_{4j}I_4, \end{aligned} \tag{21}$$

where e_{jk} and f_{jk} are determined by the geometry and electrical constants of the system. (V_j, the potential at the surface of the jth conductor, is determinable from the geometry of the conductors as a linear function of the conductor *charges*, Q_1, Q_2, $\cdots$; that is, V_j may be written

$$V_j = p_{1j}Q_1 + p_{2j}Q_2 + p_{3j}Q_3 + p_{4j}Q_4,$$

the p coefficients being the Maxwell potential coefficients of the system. These, however, are not required in the present problem.)

But relation (20) must hold at the surface of the conductors. Thus, since the electric force is continuous at the surface of the conductors, relations (20) and (21) give

$$\begin{aligned} &Z_{11}I_1 + Z_{21}I_2 + Z_{31}I_3 + Z_{41}I_4 = \gamma V_1 = E_1 + F, \quad r_1 = a, \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &Z_{14}I_1 + Z_{24}I_2 + Z_{34}I_3 + Z_{44}I_4 = \gamma V_4 = E_4 + F, \quad r_4 = a, \end{aligned} \tag{22}$$

where

$$Z_{jk} = e_{jk} + f_{jk} = Z_{kj},$$

and

$$Z_{jj} = Z_{kk},$$

the Z coefficients being the self and mutual impedances of the individual conductors. The required self and mutual impedances, Z_s and Z_m, respectively, however, are the impedances of the circuits 1–4 and 2–3. Owing to the relations, $I_1 = -I_4$ and $I_2 = -I_3$, of the currents, Z_s and Z_m are given by

$$\begin{aligned} Z_s &= 2(Z_{11} - Z_{41}) \\ \text{and} \qquad Z_m &= 2(Z_{21} - Z_{31}). \end{aligned} \tag{23}$$

Thus, from equations (22) and (23), we have

$$Z_s I_1 + Z_m I_2 = \gamma(V_1 - V_4) = 2(E_1 + F), \quad r_1 = a. \tag{24}$$

The problem is then reduced to the determination of E and F in terms of I_1 and I_2.

The function F must satisfy Laplace's equation in two dimensions and may be resolved into four waves centered respectively on the axes of the four wires, each satisfying Laplace's equation. Thus, at any point (r_j, θ_j) in the dielectric, F may be written

$$F = F_1 + F_2 + F_3 + F_4, \tag{25}$$

where

$$F_j = A_{0j} \log r_j + \sum_{n=1}^{\infty} \left(A_{nj} \frac{\cos n\theta_j}{r_j{}^n} + B_{nj} \frac{\sin n\theta_j}{r_j{}^n} \right), \qquad j = 1, 2, 3, 4.$$

The arbitrary constants A_{0j} are determined by the relations

$$4\pi\mu i\omega I_j = -\int_0^{2\pi} \left(\frac{\partial F}{\partial r_j} \right)_{r_j = a} a\, d\theta_j. \tag{26}$$

But owing to the specific configuration and to the conditions

$$I_4 = -I_1 \qquad \text{and} \qquad I_3 = -I_2, \tag{27}$$

the $8n$ arbitrary constants A_{nj}, B_{nj} may be reduced to $4n$. Thus, we have

$$\begin{aligned} A_{n1} &= -A_{n4} = A_n, \qquad B_{n1} = B_{n4} = B_n, \\ A_{n2} &= -A_{n3} = C_n, \qquad B_{n2} = B_{n3} = D_n, \end{aligned} \tag{28}$$

and also

$$A_{01} = -A_{04} = -2\mu i\omega I_1, \qquad A_{02} = -A_{03} = -2\mu i\omega I_2.$$

Inside of the conductors the axial electric force E_j must satisfy the wave equation in two dimensions. It may, therefore, be expressed as the Fourier-Bessel series,

$$E_j = g_{0j}J_o(r_jz/a) + \sum_{n=1}^{\infty} J_n(r_jz/a)(g_{nj}\cos n\theta_j + h_{nj}\sin n\theta_j). \quad (29)$$

The constants g_{0j} are given by the relations

$$4\pi\mu_c i\omega I_j = \int_0^{2\pi} \left(\frac{\partial E_j}{\partial r_j}\right)_{r_j=a} a d\theta_j,$$

or

$$g_{0j} = Z_j I_j \frac{J_o(r_jz/a)}{J_o(z)}, \quad (30)$$

Z_j being the internal impedance per unit length of the jth conductor with concentric return. Here μ_c is the permeability of the conductor, $J_n(z)$ is the Bessel function of the first kind of nth order and argument $z = ai\sqrt{4\pi\sigma\mu_c i\omega}$ and the arbitrary constants g_{nj} and h_{nj} are to be determined by boundary conditions; it is evident, however, that we must have

$$\begin{aligned} g_{n1} &= -g_{n4}, & h_{n1} &= h_{n4}, \\ g_{n2} &= -g_{n3}, & h_{n2} &= h_{n3}. \end{aligned} \quad (31)$$

At the surfaces $r_j = a$, the boundary relations are

$$\frac{\partial F}{\partial r_j} = -\frac{\mu}{\mu_c}\frac{\partial E}{\partial r_j}$$

and (32)

$$\frac{\partial F}{\partial \theta_j} = -\frac{\partial E}{\partial \theta_j}.$$

Hence, introducing (25) and (29) in (32), applying (32) at the two surfaces $r_1 = a$ and $r_2 = a$ and equating harmonic coefficients, gives $8n$ equations in the $8n$ arbitrary constants A_n, B_n, C_n, D_n, g_{n1}, g_{n2}, h_{n1}, h_{n2}. This procedure requires that F be expressed in terms of r_1, θ_1 and of r_2, θ_2 by suitable transformations of coordinates.[13] Thus, for all points in the neighborhood of $r_1 = a$, for example, F may be written

$$\begin{aligned} F = {} & 2\mu i\omega I_1 \log\frac{c_2}{r_1} + 2\mu i\omega I_2 \log\frac{2c}{c_2} - \sum_{n=1}^{\infty}\left(-\frac{1}{2c}\right)^n C_n \\ & - \sum_{n=1}^{\infty}\left(-\frac{1}{c_2}\right)^n \left[\left(\cos\frac{n\pi}{4}\right)(A_n - C_n) + \left(\sin\frac{n\pi}{4}\right)(B_n - D_n)\right] \\ & + \sum_{n=1}^{\infty}(\cos n\theta_1)(A_n' r_1^n + B_n' r_1^{-n}) + \sum_{n=1}^{\infty}(\sin n\theta_1)(C_n' r_1^n + D_n' r_1^{-n}), \end{aligned} \quad (33)$$

[13] The necessary formulas for these transformations are derived in Note II of the paper "Transmission Characteristics of the Submarine Cable" by John R. Carson and J. J. Gilbert, *Journal Franklin Institute*, December, 1921.

where A_n', B_n', C_n' and D_n' are expressible in terms of A_n, B_n, C_n and D_n and of the currents, electrical constants and dimensions of the system. In the neighborhood of $r_2 = a$, F is given by a similar expression in the coordinates r_2, θ_2. The application of the boundary relations at $r_1 = a$ and $r_2 = a$ then, as explained above, leads to a set of equations which determine the arbitrary constants in terms of the currents, electrical constants and dimensions of the system. When these equations are solved and the arbitrary constants are known, equation (24) becomes

$$Z_s I_1 + Z_m I_2 = 2\left(Z_1 + 2\mu i\omega \log \frac{c_2}{a} + \Delta_s\right) I_1 + 2\left(2\mu i\omega \log \frac{2c}{c_2} + \Delta_m\right) I_2, \quad (34)$$

where

$$\Delta_s I_1 + \Delta_m I_2 = -\sum_{n=1}^{\infty}\left(-\frac{1}{2c}\right)^n C_n - \sum_{n=1}^{\infty}\left(-\frac{1}{c_2}\right)^n \left[\left(\cos\frac{n\pi}{4}\right)(A_n - C_n) + \left(\sin\frac{n\pi}{4}\right)(B_n - D_n)\right].$$

The formal solution is then complete, $\Delta_s I_1 + \Delta_m I_2$ representing the correction in the series voltage drop of the primary circuit due to the proximity effect.

Solution by Successive Approximations

As the set of simultaneous equations, upon which depends the determination of the arbitrary constants A_n, B_n, C_n and D_n, involves an infinite number of unknowns, a direct solution is, in general, impossible. Consequently, some method of successive approximation is required. The convergence of the harmonic sequences indicates the practicability of the following procedure in the present problem.

(1) Determine first approximations $A_n^{(0)}$ and $B_n^{(0)}$ by boundary conditions at $r_1 = a$, neglecting the summations in C_n and D_n. For the first approximation only $A_1^{(0)}$ and $B_1^{(0)}$ will be required and the series may be represented by their leading terms.

(2) Determine $C_n^{(1)}$ and $D_n^{(1)}$ in terms of $A_n^{(0)}$ and $B_n^{(0)}$ by conditions at $r_2 = a$.

(3) Determine $A_n^{(1)}$ and $B_n^{(1)}$ in terms of $C_n^{(1)}$ and $D_n^{(1)}$ by conditions at $r_1 = a$.

. . . . . .

Then, we have, for example,

$$\sum_{n=1}^{\infty}\left(-\frac{1}{2c}\right)^n C_n = \sum_{n=1}^{\infty}\left(-\frac{1}{2c}\right)^n (C_n^{(0)} + C_n^{(1)} + C_n^{(2)} + \cdots) \quad (35)$$

and similar expressions for the other summations. Now, putting $C_n^{(0)} = D_n^{(0)} = 0$, the first approximation to the proximity effect is given by

$$\Delta_s^{(0)} I_1 + \Delta_m^{(0)} I_2 = \frac{A_1^{(0)}}{2c} + \frac{B_1^{(0)}}{2c}. \quad (36)$$

Next, since, for example, $A_2^{(0)}/(2c)^2$ is of the same order of magnitude as $A_1^{(1)}/(2c)$, the increment due to $C_n^{(1)}$ and $D_n^{(1)}$ will be

$$\Delta_s^{(1)} I_1 + \Delta_m^{(1)} I_2 = \frac{A_1^{(1)}}{2c} + \frac{B_1^{(1)}}{2c} - \frac{D_1^{(1)}}{2c} - 2\frac{B_2^{(0)}}{(2c)^2} - \frac{C_2^{(1)}}{(2c)^2} + 2\frac{D_2^{(1)}}{(2c)^2}. \quad (37)$$

Then a second approximation to $\Delta_s I_1 + \Delta_m I_2$ will be

$$(\Delta_s^{(0)} + \Delta_s^{(1)}) I_1 + (\Delta_m^{(0)} + \Delta_m^{(1)}) I_2$$

and, in general,

$$\Delta_s I_1 + \Delta_m I_2 = \sum_{n=0}^{\infty} (\Delta_s^{(n)} I_1 + \Delta_m^{(n)} I_2). \quad (38)$$

Applying this method we assume unit permeability for wires and dielectric. Then putting the first approximation in equation (34) gives equation (13) above for $M^{(0)}$. Neglecting terms containing λ^6, we find $C_1^{(2)}/2c$ and $D_1^{(2)}/2c$ ignorable. In $B_2^{(0)}/(2c)^2$, $C_2^{(1)}/(2c)^2$ and $D_2^{(1)}/(2c)^2$ we require the first terms. We then have

$$\Delta_s^{(1)} I_1 + \Delta_m^{(1)} I_2 = -2i\omega I_1(k_1 - 6k_1^2 + \cdots + \tfrac{9}{2}k_2 + \cdots) - 2i\omega I_2(k_1 - 7k_1^2 + \cdots + 4k_2 + \cdots). \quad (39)$$

Hence

$$Z_m = 4i\omega(\log\sqrt{2} - 2k_1 + 7k_1^2 - \cdots - 4k_2 + \cdots) \quad (40)$$

and

$$Z_s = 2Z_1 + 4i\omega\left(\log\frac{\sqrt{2}c}{a} - 3k_1 + 6k_1^2 - \cdots - \frac{9}{2}k_2 + \cdots\right), \quad (41)$$

where

$$k_1 = \frac{\lambda^2\tau_1}{1 - 2\lambda^2\tau_1},$$

$$k_2 = \frac{\lambda^4\tau_2}{1 - 12\lambda^4\tau_2}$$

and

$$\tau_n = 1 - \frac{2n}{z}\frac{J_n(z)}{J_{n-1}(z)}.$$

The relations

$$J_o(z) = \text{ber } x + i \text{ bei } x$$

and

$$J_1(z) = \frac{1}{\sqrt{-i}}(\text{ber}' x + i \text{ bei}' x),$$

where

$$z = x\sqrt{-i}$$

give the expressions in equations (13) and (14) for τ_1 and τ_2.

For the asymptotic values we have

$$\frac{J_n(z)}{J_{n-1}(z)} \to -i, \text{ when } x \to \infty,$$

so that

$$\tau_n \to 1 + i\frac{2n}{z}$$

or

$$\tau_1 \to 1 - \nu + i\nu$$

and

$$\tau_2 \to 1 - 2\nu + i2\nu,$$

where

$$\nu = \sqrt{2}/x = 1/(2\pi a\sqrt{f\sigma}).$$

Also,

$$k_1 \to \frac{\lambda^2}{1 - 2\lambda^2}\left(1 - \frac{\nu}{1 - 2\lambda^2}\right) + i\frac{\lambda^2\nu}{(1 - 2\lambda^2)^2},$$

$$k_1^2 \to \frac{\lambda^4}{(1 - 2\lambda^2)^2}\left(1 - \frac{2\nu}{1 - 2\lambda^2}\right) + i\frac{2\lambda^4\nu}{(1 - 2\lambda^2)^3},$$

and

$$k_2 \to \frac{\lambda^4}{1 - 12\lambda^4}\left(1 - \frac{2\nu}{1 - 12\lambda^4}\right) + i\frac{2\lambda^4\nu}{(1 - 12\lambda^4)^2}.$$

Thus equations (15) and (16) readily follow.

In addition, the high frequency value of the self *impedance*, Z_s, is given by

$$Z_s^{(0)} = 2Z_1\left(1 + \frac{4\lambda^2}{1 - 4\lambda^2}\right) + 4i\omega\left(\log\frac{\sqrt{2}c}{a} - \frac{2\lambda^2}{1 - 2\lambda^2} + \frac{2\lambda^2\nu}{1 - 4\lambda^2}\right)$$

and

$$Z_s = Z_s^{(0)} + 2Z_1\left(\frac{2\lambda^2}{1-4\lambda^2} - 6\lambda^4\right) - 4i\omega\left(\frac{\lambda^2}{1-2\lambda^2} - \frac{3}{2}\lambda^4 - \nu\left(\frac{\lambda^2}{1-4\lambda^2} - 3\lambda^4\right)\right),$$

where

$$Z_1 \to \frac{1}{a}\sqrt{\frac{f}{\sigma}}(1+i) = \omega\nu(1+i).$$

APPENDIX *

Production and Properties of Electric Field Intensities and Voltages

This appendix gives a summary of certain points in fundamental electromagnetic theory which are necessary for a thorough understanding of some portions of Part I of the paper.

Precisely defined, "voltage" (W) means the line-integral of the electric field intensity (E) along a specified path (s) between two specified points.[14] Thus

$$W = \int_{(s)} E_s ds = \int_{(\cdot)} E \cdot ds. \tag{1}$$

At any point, in a dielectric or in a conductor, the total electric field intensity E is the resultant of a part E_q due to all charges and a part E_u due to all currents; thus $E = E_q + E_u$. (E_q and E_u might be called the "charge electric intensity" and the "current electric intensity" respectively.)

Precisely stated, the phrases "due to all charges" and "due to all currents" have the same meanings respectively as in the formulations of the "retarded scalar potential" Ψ and the "retarded vector potential" A of electromagnetic theory, as summarized in the following paragraphs. "All charges" and "all currents," respectively, include polarization charges and polarization currents in a dielectric, thus allowing (indirectly) for a specific inductive capacity of any specified value. Furthermore, "all currents" include also such additional currents (current whirls) as would account for a magnetic permeability of any specified value. On the other hand, displacement currents are *not* included and should not be, for they do not play the role of true

* This appendix relates to Part I.

[14] The "electric field intensity" (or, briefly, "electric intensity") is often called the "electric force."

physical "causes" when "retardation" is allowed for in the formulation of the effects.[15]

It is of course possible to give, in a single step, formulas for E_q and E_u in terms explicitly of the charges and currents to which they are respectively due. However, it is much preferable, both mathematically and physically, to proceed in two steps, of which the first consists in giving the formulas for the two potential functions, Ψ and A, and the second in giving the formulas expressing E_q and E_u in terms of Ψ and A respectively. For convenience these four formulas will now be given together. For completeness the formula for the magnetic field intensity H will be added, although it is of only secondary interest here and in Part I of this paper; further, the formula for the relation between Ψ and A will be included, since it underlies the formulas for Ψ and A. These six formulas, which are classical, follow. The functional notation $q(t - r/c)$, in formula (2), indicates that the charge-density q is to be evaluated at the time $t - r/c$, as discussed in the next paragraph; similarly for the current-density u in (3).

$$\Psi = \int \frac{q(t - r/c)}{r}\, dv, \qquad (2) \qquad\qquad A = \frac{1}{c}\int \frac{u(t - r/c)}{r}\, dv, \qquad (3)$$

$$E_q = -\operatorname{grad} \Psi, \qquad (4) \qquad\qquad E_u = -\frac{1}{c}\frac{\partial A}{\partial t}, \qquad (5)$$

$$H = \operatorname{curl} A, \qquad (6) \qquad\qquad \operatorname{div} A + \frac{1}{c}\frac{\partial \Psi}{\partial t} = 0. \qquad (7)$$

Although usually the application of the first two of these formulas to specific cases is difficult and laborious, their physical meaning is rather simple, as will shortly appear in the following description and discussion of them.

The six formulas in the above set constitute a complete explicit solution of Maxwell's differential equations of the electromagnetic field, and form the connecting link between those differential equations and electric circuit theory. They express the potentials (Ψ, A), and thence the field intensities (E_q, E_u, H), at a specified point P and time t, due to all of the distributed charges and currents contemplated. The point P may be anywhere, in a dielectric or in a conductor; and the time t is that observed at P. dv is a *fixed* element of volume or of surface (as the case may be[16]) at any typical point in the contemplated

[15] For a mathematical treatment relating to the various matters touched on in this paragraph reference may be made to the appendix of the paper by John R. Carson cited at the end of footnote 5.

[16] For brevity the term "volume-element" will throughout be used generically to include "surface-element" as a limiting case, with "charge-density" being interpreted as "volume charge-density" and "surface charge-density" respectively.

system of charges and currents; r is the distance between dv and P; and c is the velocity of light in free space. $q(t - r/c)$ and $u(t - r/c)$ are the charge-density and the vector current-density, respectively, in dv, not at the time t but at the slightly earlier time $t - r/c$, allowance thus being made for the time of propagation of the effect from dv to P. Thus in (2) the integration, made at the time t, which is that observed at P, must include every volume-element dv which contained any charges at the time $t - r/c$, whatever the motions of those charges; and in (3) the integration must include every volume-element dv which contained any current (moving charges) at the time $t - r/c$; moreover, associated with each volume-element dv is a corresponding value of r.

r denoting distance, Ψ and A are called "potentials" because of their inverse dependence on r and their direct dependence on the charge-density q and the current-density u respectively. Ψ is called the "scalar potential" because it does not have direction in space; A the "vector potential" because it has direction. These potentials are qualified as being "retarded" potentials[17] because the values to be taken for the charge-elements and current-elements are not their actual values at the contemplated instant t but their "retarded" values, that is, their values at the earlier instants $t - r/c$. (It is to be remembered that the time t is that observed at the point P where Ψ and A are to be calculated.)

In the way of a summary statement regarding the set of formulas (2) to (7), we may say that electric charges, whether stationary or moving, produce a scalar potential Ψ calculable from (2), and thence an electric field intensity E_q calculable from (4); and that if the charges are in motion, thus constituting currents, they produce also a vector potential A calculable from (3), and thence an additional electric field intensity E_u calculable from (5) and a magnetic field intensity H calculable from (6). Thus the total, or resultant, electric field intensity $E = E_q + E_u$ is calculable from

$$E = -\operatorname{grad} \Psi - \frac{1}{c}\frac{\partial A}{\partial t}. \tag{8}$$

If the contemplated point P for which E is calculated is in a conductor, of resistivity ρ, where the current-density is u', there exists the additional relation $E = \rho u'$, in accordance with Ohm's law.

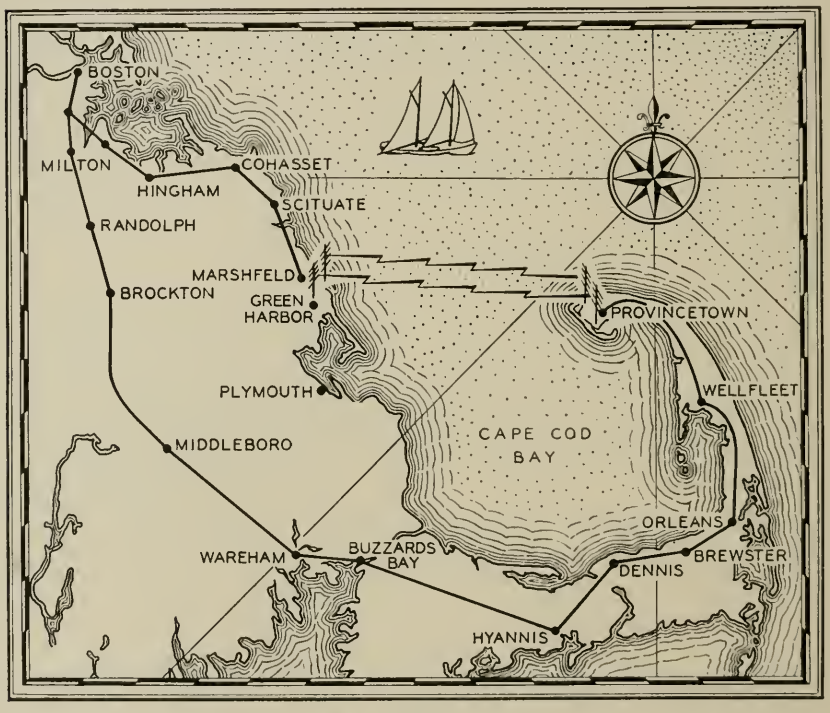
Of the important contrasting principles enunciated in the section entitled "The Two Parts of a Voltage, and Their Resultant," principle "1" is an immediate consequence of equation (4) of this appendix, and "2" is a consequence of (5) and (6) together.

[17] Sometimes called "propagated" potentials.

An Unattended Ultra-Short-Wave Radio Telephone System *

By N. F. SCHLAACK and F. A. POLKINGHORN

FOR several years attention has been directed by Bell Telephone Laboratories toward determining the characteristics of ultra-high frequencies and their possible application to the telephone plant. This led to the belief that ultra-high-frequency radio might find a useful field as an adjunct to the wire telephone plant in crossing natural barriers where other means might prove difficult or expensive.



Map

In order to make the new facility practicable for use in as many as possible of the situations for which it is technically adapted, it is necessary to keep the total operating costs low. By designing the

* Digest of paper presented at Annual Convention of the Institute of Radio Engineers in Detroit, July, 1935. The paper will be published in full in the *Proc. I. R. E.*, October, 1935.

equipment to include certain features over and above the basic requirements, it is possible to reduce to a minimum the attendance necessary to assure continuous operation. Further economies are effected by using equipment capable of continuous operation out-of-doors.

Since ultra-high-frequency radio circuits are normally quite stable and comparatively free from noise, it is possible to omit volume regulation and voice operated devices such as are used on transatlantic circuits at a considerable saving in cost. Under this condition it is necessary, however, to provide a radio transmitter of somewhat higher power capacity than would be required if volume regulation were used, but the cost of this additional power is small compared to the cost of the features required to provide for regulated volume operation.

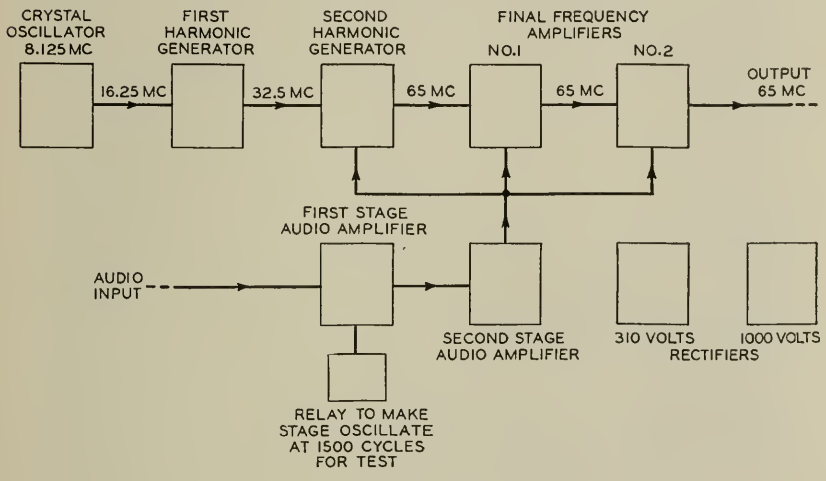


Fig. 1—Block schematic of ultra-short-wave transmitter.

With unattended operation it is desirable that starting and stopping both the transmitter and the receiver be separately controlled from the telephone office. Local testing arrangements should also be provided if possible, to allow the test board operator to determine whether the transmitter and receiver are operating properly.

In order to obtain representative information on the feasibility of operating an ultra-high-frequency telephone circuit on an unattended basis in the telephone plant and to secure a better idea of the mechanical and electrical requirements, equipment was constructed by Bell Telephone Laboratories for such an installation. With the cooperation of the New England Telephone and Telegraph Company this equipment has been used to establish an experimental ultra-short-wave

circuit between Green Harbor and Provincetown, Massachusetts, as indicated on the map. Sand dunes near Provincetown, rising about 80 feet, make it possible to secure an optical path across the bay. The

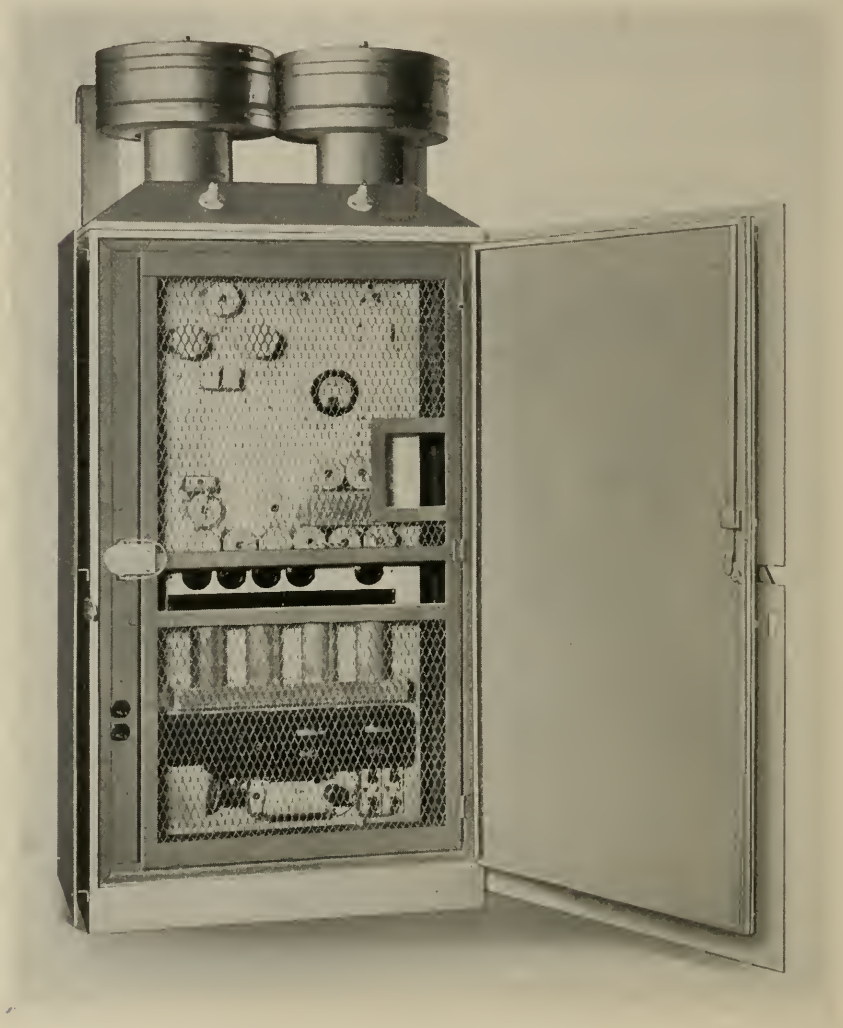


Fig. 2—Ultra-short-wave transmitter mounted in metal container suitable for pole mounting.

radio circuit is extended by wire from Green Harbor to Boston to form a direct Boston-Provincetown toll circuit. It is used as one of a group of terminal circuits and is operated at the normal overall net loss for

this type of circuit, namely, 9 decibels. Transmission from Green Harbor to Provincetown is accomplished on a frequency of 65 mc. and in the reverse direction on 63 mc. This does not represent the minimum possible frequency spacing for this equipment, but was a convenient one for the experiment.

At Boston and at Provincetown the circuit appears at a jack in the switchboard beside the jacks of wire toll circuits. As far as the operator is concerned, switching and ringing operations are performed in the same manner as for other similar grade toll circuits and there is nothing to designate that this toll circuit has a radio link. The insertion of a cord into the jack starts the radio transmitter at that end of the circuit.

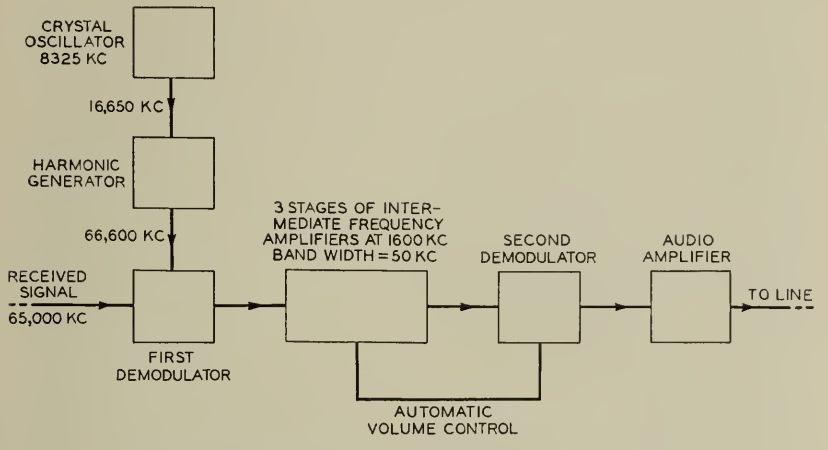


Fig. 3—Block schematic of ultra-short-wave receiver.

The receivers at both ends are kept in constant operation while the circuit is available for traffic but are started and stopped by the operation of a key at the local test board. Ringing is accomplished by sending a 1000-cycle tone interrupted at 20 cycles over the circuit. Privacy equipment similar to that used on the transatlantic short-wave radio channels is installed at the terminal offices.

The transmitters are crystal controlled and are capable of delivering 15 watts of carrier power which can be completely modulated. It was estimated that this would give a reasonably satisfactory circuit. A block schematic of the Green Harbor transmitter is shown in Fig. 1. The Provincetown transmitter is of similar construction.

The receivers are of the double-detection type (see Fig. 3), and to make unattended operation possible and at the same time permit high selectivity at these frequencies, a crystal oscillator is used as the source

of beating frequency. A small amount of automatic volume control is provided to compensate for slight variations in received voltage caused by variation in humidity and other factors. The receivers are capable of delivering 0.3 watt of undistorted power to a 600-ohm



Fig. 4—Ultra-short-wave receiver mounted in metal container suitable for pole mounting.

impedance. This is well in excess of the power required during normal operation.

The transmitting and receiving antennas are identical and each is



Fig. 5—General view of antennas and pole mounted radio equipment at Green Harbor terminal.

mounted on a single wooden pole about ninety feet high. Horizontal exciter and reflector elements are supported on standard cross-arms. Four pairs of half-wave exciter elements, each comprising two half-wave conductors, are spaced one-half wave-length apart in a vertical

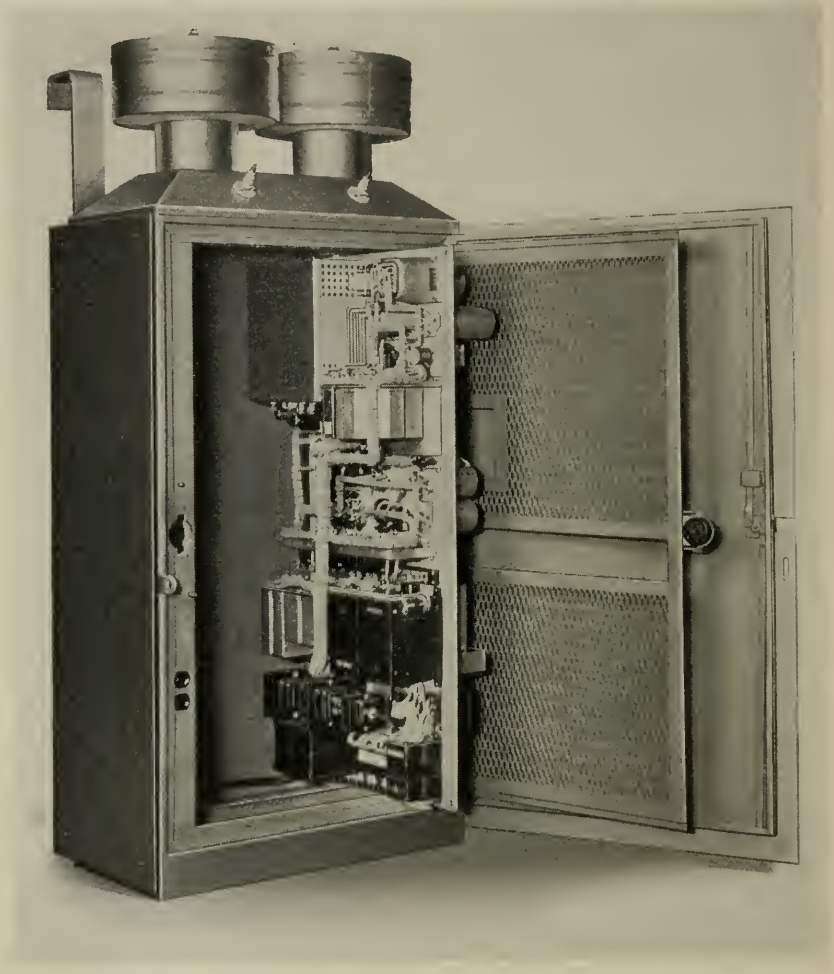


Fig. 6—Open view of ultra-short-wave transmitter.

plane on one side of the pole. Four pairs of half-wave reflector elements are similarly arranged on the opposite side of the pole. The spacing between exciters and reflectors is one-quarter wave-length. These antennas when mounted as shown in Fig. 5, give a gain, measured

at the other end of the circuit, of 12 db over a simple half-wave element with the same power input, at the same mean height. This type of antenna was used as it gave the highest gain and directivity which could be conveniently mounted on a single pole. High directivity was desirable not only as a means of increasing the received signal but also to exclude automobile ignition and other noises originating near the receiving stations. The transmitters and receivers are mounted on the poles with their respective antennas.

Daily observations of the circuit loss have not shown variations greater than $\pm$ 4 db from the normal value. Noise from local thunder storms has never prevented the circuit from being utilized in the normal manner. The several months of traffic operation to which the circuit has been subjected have disclosed no important technical difficulties with this type of system. It has been found that the radio apparatus can remain in operation over periods of several weeks without attention or adjustment.

Around the World by Telephone

THE first two-way telephone conversations completely encircling the earth took place at New York City, during April, 1935. The two telephone instruments used were located in separate rooms on the 26th floor of the Long Distance Building at 32 Sixth Avenue, New York. In connection with these tests, Mr. Gifford, President of the American Telephone and Telegraph Company, spoke with Vice-President Miller, and a number of other persons conversed over this around-the-world circuit for some thirty minutes on April 25th.

These world-encircling conversations were made possible by the very close cooperation of the several communication interests involved, including the British Post Office, the Netherlands Telephone Administration, the Netherlands Indies Telephone Administration, and the Bell System.

The circuit used was completely four-wire and was made up of a cable and open wire carrier telephone circuit from New York to San Francisco, a radio link from San Francisco to Bandoeng, Java, a radio link from Bandoeng to Amsterdam, the Netherlands, a land and submarine cable link from Amsterdam to London, England, and another radio link back to New York.

The approximate lengths of each type of facility and the radio frequencies used in each radio link are indicated in the table below:

Circuit Sections	Type of Facilities	Length	Radio Frequencies Used
New York–San Francisco	Cable and Open Wire Carrier	3500 Mi.	—
San Francisco–Bandoeng	Radio	8700 Mi.	E.–W. 10840 Kc. W.–E. 9415 Kc.
Bandoeng–Amsterdam...	Radio	7300 Mi.	E.–W. 19355 Kc. W.–E. 18535 Kc.
Amsterdam–London.....	Cable	300 Mi.	—
London–New York.....	Radio	3500 Mi.	E.–W. 12150 Kc. W.–E. 18340 Kc.
Total.....		23300 Mi.	

In completing this around-the-world connection the operating personnel at the various points involved were working at vastly differ-

ent times of the day. At New York the time was 9:30 in the morning and at San Francisco 6:30 in the morning. Between San Francisco and Java the voice went from April 25th into April 26th and back reaching Java at 10:00 o'clock at night on April 25. At Amsterdam the time was 2:50 in the afternoon and at London 2:30 in the afternoon.

A total of about 980 vacuum tubes were used and of this number 515 were in the United States. In all, the above vacuum tubes produced a gain of about 2000 decibels in each direction of transmission. The total delay in the transmission of speech over this circuit was about one quarter of a second. While this was not particularly noticeable to the talkers at the two ends of the circuit, at one time during the conversations telephone receivers located near one of the talkers were connected to the opposite end of the circuit. With this arrangement the delay between the speech from the talker as heard by the short cut and that as heard over the circuit was very marked. It is of interest to note that while the land line links accounted for only about 15 per cent of the total distance traversed by the voice, they were responsible for about 55 per cent of the delay.

Abstracts of Technical Articles from Bell System Sources

Receiver Band-Width and Background Noise.[1] C. B. Aiken and G. C. Porter. In doubling the band-width of a radio receiver, it might be supposed that the apparent noise level would increase about 3 db since the noise energy brought in should be doubled. However, the high-frequency components of noise may be very much more troublesome from the standpoint of the listener than the low-frequency components. An experimental study shows that this is actually the case. Thus, if the noise level is low, as it should be whenever an effort is made to employ high fidelity reception, the dependence of the signal-to-noise ratio upon band-width is very apparent. While there are many variables involved, it seems safe to conclude that in doubling received band-width, the required increase in field strength may even be as much as 8 to 10 db.

Cable Sheath Corrosion—Causes and Mitigation.[2] J. B. Blomberg and Norvel Douglas. The causes and mitigation of telephone cable sheath corrosion are dealt with in this paper, which describes particularly a method of applying a counter potential to the cable sheath for the mitigation of corrosion from localized currents. This method, although not new, has had but limited application. It may find extensive future use for controlling corrosion on intercity toll cables, and in localities where street railways have been abandoned. In addition to a unique application of this method, there is also described the method of correcting by current drainage a particularly bad case of corrosion from stray current.

The Detection of Frequency Modulated Waves.[3] J. G. Chaffee. The comparative ease with which pure frequency modulation can be produced in electron oscillators at ultra-high frequencies has led to an examination of the problem of detecting a frequency modulated wave. In this region of frequencies the high ratio of frequency shift to modulating frequency gives rise to a very large number of side bands in the spectrum representing the modulated wave. Detection is usually accomplished by distorting the spectrum by means of a

[1] *Radio Engg.*, May, 1935.
[2] *Elec. Engg.*, April, 1935.
[3] *Proc. I. R. E.*, May, 1935.

selective network and then impressing the output voltages upon the grid of a detector. This process is treated analytically and formulas are given which permit the calculation of low-frequency detection products in terms of the transmission characteristic of the distorting network and the maximum frequency shift during modulation.

Measured detection products produced in such a system are compared with values calculated by means of the formulas which are given, and the results are shown to be in substantial agreement over the region in which certain simplifying assumptions are justified.

Acceptance-Rejection Requirements in Specifications.[4] H. F. DODGE. Specifications for quality of materials and finished products impose requirements for individual quality characteristics to distinguish between what may be considered satisfactory for a given purpose and what may not. For many characteristics, 100 per cent inspection or testing is not feasible; hence reliance must be placed on sampling a part of the whole. Under these conditions, 100 per cent conformance with requirements cannot be achieved with certainty and errors arising from sampling fluctuations cannot be avoided.

The sampling clauses included in specifications often provide criteria for the acceptance or rejection of lots of a product. These clauses constitute interpretations of the intent of the basic quality requirements and serve as a basis for action. With sampling, certain risks are assumed by both the consumer and the producer. One kind of risk is discussed, and the relationship between (1) the distribution of the risk between producer and consumer, and (2) the choice of acceptance criteria and sample size, is indicated for certain conditions.

Selection and Development of Teachers for Communication Engineering Instruction.[5] O. W. ESHBACH. One of a symposium of papers presented at the Conference on Electronics and Electrical Communication at the Ithaca meeting of the S. P. E. E. commenting on normal procedure in the selection and training of teachers, the trend of development in instruction in electronics and communication, attitudes characteristic of good teachers, responsibilities toward young instructors, and the means through which broadening of knowledge may be accomplished. Selection of the right individual, development of effective teaching technique, and the enhancement of knowledge and experience are emphasized as fundamental to successful teaching.

[4] *Proc. Amer. Soc. for Testing Materials*, Vol. 34, Part II, 1934.
[5] *Jour. Engg. Education*, April, 1935.

The Correlation of Distillation Range with the Viscosity of Creosote [6] (Part V of series, "Chemical Studies of Wood Preservation"). C. J. Frosch. The results of viscosity measurements of a series of creosotes distilled from a single tar are given. It was found that these creosotes are truly viscous solutions, which permits the designation of the data as absolute viscosity. The viscosity-temperature data conform to two equations, one an empirical relationship previously found in an analogous series of crude oils, the other developed from theoretical considerations. It is remarkable that in spite of the complex nature of creosote, the viscosity data permit one to regard the material boiling below 355° C. as solvent and the residue above that temperature as solute. This is not true for any other temperature limit customarily used in creosote analysis.

An Electron Diffraction Camera.[7] L. H. Germer. An experimental apparatus is here described for obtaining and photographing electron diffraction patterns from solid substances. It is designed for the study of the crystal structures of thin films and of superficial layers on massive blocks. Electrons from a hot tungsten filament are accelerated within an evacuated metal container by a potential difference of 50 or 60 kv. They are stopped down by appropriate slits to form a narrow beam which strikes the material under investigation. Electrons scattered by this material form a diffraction pattern characteristic of the crystal structure. This pattern is registered directly upon a photographic plate in the path of the scattered electrons.

The Motion of a Bar Vibrating in Flexure, Including the Effects of Rotary and Lateral Inertia.[8] W. P. Mason. In this paper a complete theoretical solution is given for a bar vibrating in flexure taking account of rotary and lateral inertia. The solution shows that the frequency of a bar free to vibrate on both ends, is asymptotic to the frequency given by the usual solution, neglecting rotary inertia, when the ratio of width to length is small, and approaches the frequency of a bar in longitudinal vibration when the width becomes comparable to the length. The theoretical frequencies have been compared with the published results of Harrison on the frequency of a quartz crystal vibrating in flexure, and have been found to agree within one per cent for a crystal whose width is less than half its length.

[6] *Physics*, May, 1935.
[7] *Rev. Sci. Instruments*, May, 1935.
[8] *Jour. Acous. Soc. Amer.*, April, 1935.

Probability in Engineering.[9] E. C. MOLINA. The purpose of this paper is to emphasize the practical value of probability theory in engineering. For this purpose a short introduction on probability theory as such is followed by a discussion of three problems from the domain of engineering with which the author is most familiar, namely, telephony.

The first problem deals with the switching, or trunking, of telephone calls. It illustrates the part played by probability theory in determining the amount of equipment an engineer must install in anticipation of *deviations* from normal or average service demands.

The bearing of probability theory on problems wherein one is confronted with the *cumulative* effect of a multitude of small independent discrepancies is indicated by the second problem presented in the paper. A long distance telephone circuit equipped with repeaters at several points is analyzed with reference to the cumulative effect of slight voltage variations in the battery supply at each repeater station.

The third and last problem is one on sampling. It introduces the engineer to the practical significance of *inverse* or *a posteriori* probability.

Direct-Current Amplifier Circuits for Use with the Electrometer Tube.[10] D. B. PENICK. A number of balanced, single-tube, direct-current amplifier circuits are compared, which are applicable to the four-element, low grid-current vacuum tube. The balance equations are stated for the most generally useful circuit, and magnitudes of the tube characteristics involved are given for the Western Electric No. D-96475 Tube. Experimentally determined values of circuit constants observed under balance conditions are also given. The stability of the circuit is discussed, and a convenient procedure for obtaining a balance by experimental methods is suggested.

Internal Dissipation in Solids for Small Cyclic Strains.[11] R. L. WEGEL and H. WALTHER. This paper presents the results of investigations of dissipation of energy in vibrating solids, mostly metals, by means of longitudinal and torsional vibrations of cylindrical rods. The amplitudes of strain used have been kept between 10^{-5} cm./cm. and 10^{-8} cm./cm., in which range the dissipation of energy is proportional to the square of the strain. The specific dissipative property of a material is expressed in three different ways: (1) Equivalent viscosity or the ratio

[9] *Elec. Engg.*, April, 1935.
[10] *Rev. Sci. Instruments*, April, 1935.
[11] *Physics*, April, 1935.

of stress to dissipative component of strain rate; (2) hysteretic constant defined as the area in ergs of the cyclic stress-strain diagram; and (3) elastic phase constant defined as the ratio of specific elastic reactance to equivalent viscosity. Within a range of frequencies 100 to 100,000 cycles per second the results show that the hysteretic constant is proportional to some power Δ of the frequency, the numerical value of the exponent Δ varying between the limits $-\frac{1}{3}$ and $+\frac{1}{2}$, depending on the kind of material and its internal structural condition. Measurements made with longitudinal and torsional vibration indicate that dissipation is associated with dilatation as well as with pure shear. Preliminary studies are described showing the correlation between internal dissipation in metals and temperature hardness effects of annealing and aging.

Broadcasting Studio Acoustics.[12] S. K. Wolf and C. C. Potwin. It is now of fundamental importance that studios be designed to provide an acoustic transmission characteristic that will insure the fullest benefits from the many recent improvements in transmitting and receiving systems. For this reason, the traditional "dead" studio, which was so common in the early days of radio, is no longer suited to the present technique of broadcasting.

This paper deals with improved methods of analysis and treatment, particular consideration being given to the problems of the small studio. A description of the high-speed level recorder, its operating characteristics and application to studio analysis, are included. The increased accuracy of instrumental measurement over computational methods in the solution of the problems of reverberation, multiple reflection and room resonance at various frequencies is explained. The factors governing the proper selection and distribution of acoustic materials are discussed and supported by actual measured data taken with the level recorder in studios designed in accordance with the methods advocated.

Two typical studio designs are illustrated, one suggesting sound reflective angular wall and ceiling surfaces adjacent to the performers, a moderate sound absorbent on the intermediate surfaces and a highly efficient absorbent on the surfaces adjacent to and surrounding the microphone. Distant pickup employed in this type of studio is briefly described.

Quantitative Studies on the Singing Voice.[13] S. K. Wolf, D. Stanley and W. J. Sette. The field of singing has been handicapped by the

[12] *Communication and Broadcast Engineering*, April, 1935.
[13] *Jour. Acous. Soc. Amer.*, April, 1935.

lack of suitable quantitative means for simply evaluating the various voice factors. With the aid of recently developed acoustic instruments, the authors have investigated physical characteristics of vocal tones, including attack, quality, vibrato, and power as a function of pitch. Measurements have been made and repeated on more than fifty singers in various stages of development. On the basis of the results, it is possible partially to evaluate and criticize a singer's technical equipment, and determine by periodic tests whether the voice is improving or deteriorating.

Better singers were found to attack a tone more vigorously and sustain it more uniformly, to possess a vibrato with a rate of about six per second, and to excel in those phases of artistry dependent upon proper control. They are also capable of producing relatively high amounts of acoustic power over wider singing ranges, the power increasing gradually with increasing pitch. Harmonic analyses, with the intensity levels of individual partial tones each averaged over an interval of about .5 second, have as yet failed to reveal consistent differences between good and bad voices.

Contributors to this Issue

J. A. Becker, A.B., Cornell University, 1918. Bureau of Standards, 1918; Westinghouse Electric and Manufacturing Company, 1919; Western Electric Company, 1919. Ph.D., Cornell University, 1922; National Research Fellow, California Institute of Technology, 1922–24; Assistant Professor, Stanford University, 1924. Western Electric Company, 1924–25; Bell Telephone Laboratories, 1925–. Since 1924 Mr. Becker has been engaged in research work on thermionics, oxide coated filaments, semi-conductors, varistors, and copper oxide rectifiers.

Charles R. Burrows, B.S. in Electrical Engineering, University of Michigan, 1924; A.M., Columbia University, 1927; E.E., University of Michigan, 1935. Research Assistant, University of Michigan, 1922–23. Western Electric Company, Engineering Department, 1924–25; Bell Telephone Laboratories, Research Department, 1925–. Mr. Burrows has been associated continuously with radio research and is now in charge of a group investigating the propagation of ultra-short waves.

Arthur B. Crawford, B.S. in Electrical Engineering, Ohio State University, 1928. Member of Technical Staff, Bell Telephone Laboratories, 1928–. Mr. Crawford has been engaged chiefly in work relative to radio communication by ultra-short waves.

Carl R. Englund, B.S. in Chemical Engineering, University of South Dakota, 1909; University of Chicago, 1910–12; Professor of Physics and Geology, Western Maryland College, 1912–13; Laboratory Assistant, University of Michigan, 1913–14. Western Electric Company, 1914–25; Bell Telephone Laboratories, 1925–. As Radio Research Engineer Mr. Englund is engaged largely in experimental work in radio communication.

Ray S. Hoyt, B.S. in Electrical Engineering, University of Wisconsin, 1905; Massachusetts Institute of Technology, 1906; M.S., Princeton, 1910. American Telephone and Telegraph Company, Engineering Department, 1906–07. Western Electric Company, Engineering Department, 1907–11. American Telephone and Telegraph Company, Engineering Department, 1911–19; Department of Develop-

ment and Research, 1919–34. Bell Telephone Laboratories, 1934–. Mr. Hoyt has made contributions to the theory of transmission lines and associated apparatus, theory of crosstalk and other interference, and probability theory with particular regard to its applications in telephone transmission engineering.

Sallie Pero Mead, A.B., Barnard College, 1913; M.A., Columbia University, 1914. American Telephone and Telegraph Company, Engineering Department, 1915–19; Department of Development and Research, 1919–34. Bell Telephone Laboratories, 1934–. Mrs. Mead's work has been of a mathematical character relating to telephone transmission.

William W. Mumford, B.A., Willamette University, 1930. Bell Telephone Laboratories, 1930–. Mr. Mumford has been engaged in radio receiving work, chiefly on the problem of propagation and measurement in the ultra-short-wave region.

F. A. Polkinghorn, B.S., University of California, 1922; U.S. Naval Radio Laboratory at Mare Island Navy Yard, California, 1922–24; A-P Radio Laboratories, San Francisco, 1924–25. Pacific Telephone and Telegraph Company, San Francisco, 1925–27; Bell Telephone Laboratories, 1927–. Mr. Polkinghorn's work has been primarily in connection with the design of radio receiving and test equipment for use at high and ultra-high frequencies.

N. F. Schlaack, B.S., University of Michigan, 1925. Bell Telephone Laboratories, 1925–. Mr. Schlaack has been engaged primarily in the development of short and ultra-short-wave transmitting equipment.

E. C. Wente, A.B., University of Michigan, 1911; S.B. in Electrical Engineering, Massachusetts Institute of Technology, 1914; Ph.D., Yale University, 1918. Engineering Department, Western Electric Company, 1914–16 and 1918–24; Bell Telephone Laboratories, 1924–. As Research Physicist, Dr. Wente has worked principally on general acoustic problems and on the development of special types of acoustic devices.

VOLUME XIV OCTOBER, 1935 NUMBER 4

THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

Dr. George A. Campbell—*F. B. Jewett* 553
Dr. Campbell's Memoranda of 1907 and 1912 558
Some Aspects of Low-Frequency Induction Between Power and Telephone Circuits—*H. R. Huntley and E. J. O'Connell* 573
Circulating Currents and Singing on Two-Wire Cable Circuits—*L. G. Abraham* 600
Operation of Ultra-High-Frequency Vacuum Tubes—*F. B. Llewellyn* 632
Further Extensions of Theory of Multi-Electrode Vacuum Tube Circuits—*S. A. Levin and L. C. Peterson* 666
Transatlantic Long-Wave Transmission—*Austin Bailey and H. M. Thomson* 680
Technical Digests—
1. Superiorities of Lead-Calcium Alloys for Storage Battery Construction—*H. E. Haring, U. B. Thomas, E. E. Schumacher and G. S. Phipps* 699
2. Marine Radio Telephone Service for Boston Harbor—*F. A. Gifford and R. B. Meader* 702
3. Harbor Craft Ship-to-Shore Radio Telephone Service in Puget Sound Area—*E. B. Hansen* 708
4. Ship Sets for Harbor Ship-to-Shore Service—*H. N. Willets* . . 713
5. An Electromechanical Representation of a Piezoelectric Crystal—*W. P. Mason* 718
Abstracts of Technical Papers 724
Contributors to this Issue 728

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

50c per Copy *$1.50 per Year*

THE BELL SYSTEM TECHNICAL JOURNAL

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

EDITORIAL BOARD

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

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, 1935

PRINTED IN U. S. A.



Photo by Garber

George A. Campbell

The Bell System Technical Journal

Vol. XIV *October, 1935* *No. 4*

Dr. George A. Campbell

By F. B. JEWETT

ON the first of December next, after thirty-eight years of active and unusually productive service as a mathematical physicist and inventor, Dr. George A. Campbell retires from active membership on the staff of the Bell Telephone Laboratories.

As the history of an art can often be written most effectively in terms of the personalities who have been responsible for its upbuilding, I feel that I am not departing from the objectives of the *Bell System Technical Journal* in bringing to the attention of its readers a brief note concerning one of the chief artificers of telephone transmission. Dr. Campbell's achievements in this field entitle him beyond question to rank first among his generation of theoretical workers in electrical communication. Yet, in common with many truly great minds, it has been his nature to avoid publicity, so that outside the circle of his immediate associates and a few of the more mathematically gifted students of his chosen branch of electrical science, his fame is far from being commensurate with his achievements.

In 1897, thirty-eight years ago, the art of telephone transmission was in its infancy. Circuits of even a few hundred miles' length were rare, and the longest distance over which communication had been held was that separating New York and Chicago. It was at this time that Campbell, as a young man, after graduating from Massachusetts Institute of Technology and spending four years in graduate study at Harvard, Göttingen, Vienna and Paris, joined the staff of the American Telephone and Telegraph Company to engage in research. Familiar with the work of Rayleigh and Heaviside, Campbell's early studies sought some method of mitigating the attenuation, which levied heavy toll upon the voice currents and formed a theretofore unyielding barrier against telephone communication over very long distances.

Heaviside had suggested that inductance, if properly applied in a

long telephone circuit, should diminish rather than increase the attenuation. Campbell followed this suggestion and developed a theory of loading, but in his case there occurred one of those coincidences—fortunately rare in the history of science—of two investigators arriving at substantially the same result at the same time. Independently of Professor M. I. Pupin, he worked out a complete theory of the telephone loading coil. They both applied for patents, with the result that an interference was declared, and Pupin was able to establish a slightly earlier date of conception. The loading coil interference was decided in Pupin's favor and the famous patents issued to him. The fact should be recorded, however, that Campbell's analysis of the problem—actually more detailed than Pupin's—led him to formulate rules for the design of loading coils and their spacing which were, from the very beginning, the only ones employed in this country.

As the effectiveness of telephone instruments increased and the lengths of circuits grew, noise and crosstalk became an outstanding obstacle to telephone advance. It had been shown that this crosstalk was a complex effect resulting partly from electromagnetic and partly from electrostatic induction. In unpublished memoranda written between 1903 and 1907, Campbell pointed out the importance of Maxwell's capacity coefficients in the calculation of crosstalk and coined the much-used term "direct capacity," now modernized to "direct capacitance." It was also at this time that he designed his well-known "shielded" balance, which in one form is a bridge for measuring direct capacities. He showed in these early memoranda that crosstalk between two circuits depends, to a considerable extent, and particularly in the case of loaded circuits, on a function of the various direct capacities between the wires of two circuits. He termed this function the "direct capacity unbalance." This work led to the invention of the well-known capacity unbalance test set, hundreds of which have been used in countless measurements in the manufacture and installation of toll cables.

This study of Campbell's marked an important advance since, for the precise but unwieldy theory of crosstalk, it substituted a simple approximation—an approximation which was to remain adequate until the advent of carrier systems with their higher frequencies and shorter wave-lengths.

As control was gradually extended over the characteristics of telephone circuits, both from the standpoint of their transmission effectiveness and their freedom from crosstalk and noise, the art reached the point at which development emphasis shifted to the

telephone repeater. Here, an entirely new line problem arose,—namely, that of avoiding singing when repeaters are adapted to two-way amplification. Up to 1912, the only type of repeater circuit used was the so-called 21-type, in which a single repeater element amplifies messages which reach it from both directions and which requires that the two associated sections of line have very similar characteristics. A well-known limitation of the 21-type repeater is its tendency to "sing" when line unbalance or amplification exceed certain rather low limits.

On the other hand, the 22-type repeater has two amplifying elements and two artificial lines, one to balance each associated section of actual line. While the basic idea of the 22-type was old, it remained for Campbell, in a memorandum dated March 7, 1912, to reveal its properties of inherent stability. He points out that "singing will not be introduced by any possible unbalance however large, in either of the lines, provided the unbalance of the other line does not exceed a certain critical magnitude." Also his words, "the use of a compensating device such as an artificial line, to reduce the amplification at the resonant frequencies to the level of the amplification at other telephonic frequencies" suggest broadly the idea of equalizing for amplitude-frequency distortion which is brought in by the selective characteristics of the line circuits or other apparatus in a long system. Furthermore, "if it became necessary merely to eliminate certain frequencies lying outside of the range required for telephony, the use of an artificial selecting circuit" is definite anticipation of the use, subsequently common in all repeaters, of low-pass filters to cut off frequencies outside of the band transmitted and thus minimize line balance difficulties.

Campbell also indicated that the stability of a circuit as regards singing could be improved if the amplification were distributed among a number of properly spaced points along the line rather than concentrated at a single point.

Moreover, the great amplification made possible in telephone circuits by the perfection of the vacuum tube and its associated circuits permitted the use of cables for long distances with manifest advantages for congested routes. The great amplifications required for the longer cable circuits, however, could most effectively be handled by the use of "four-wire" circuits both for voice frequencies and later for carrier systems. Campbell was the originator of this type of circuit. In the same memorandum of 1912, which discussed the 22-type repeater, he suggested it as the logical extension of the one-way paths in the 22-type repeater, each path containing as many one-way amplifiers and line sections as desired.

The present situation in transmission, particularly toll transmission, is characterized by the growing use of high frequencies, the transmission of broader frequency bands and the use of the so-called "carrier" method. This we find in the most advanced form in the proposed very broad band cable circuits of the coaxial and other types. Transoceanic telephony and broadcasting are other outstanding uses. The electrical wave-filter in its many forms is one of the most important elements in all such systems. The filter appears as a means of sharply separating the currents of different continuous bands of frequencies in carrier telephone and telegraph systems, for sharp selectivity in radio systems and for various other uses in these and other forms of transmission such as telephone repeaters, telephotography, composite sets and testing apparatus. Indeed, the filter has, within the last few years, become almost as ubiquitous as the vacuum tube. The fundamental conception of the electric wave-filter arose out of Campbell's analysis of loaded lines. The patent was issued to him in 1917. It is evident to one reading his famous paper on "Loaded Lines in Telephonic Transmission," published in the *Philosophical Magazine* of 1903, that even at that time he had begun to envisage the high-pass and low-pass wave-filters.

Effective station sets are fundamental to all good transmission. In a memorandum dated October 8, 1906, Campbell disclosed the single-transformer anti-sidetone station circuit which is achieving almost world-wide acceptance. Later, he carried out a comprehensive and conclusive piece of work in revealing all of the possible circuit arrangements for doubly conjugate branches and in setting down the impedance relations of the line, network, transmitter and receiver of these various branches. This systematic analysis of the problem greatly facilitated a comprehensive survey, giving assurance that all types of circuits would be considered and that those which best fitted the available transmitters and receivers would be selected. The work was summarized in an extensive paper entitled "Maximum Output Networks for Telephone Substation and Repeater Circuits" by Campbell and Foster in the *A. I. E. E. Transactions* of 1920.

It would appear that Campbell also originated the articulation test which now finds a use wherever telephone development work is in progress. In a paper entitled "Telephonic Intelligibility," which appeared in the *Philosophical Magazine* of January, 1910, he describes how, in connection with tests he had been conducting, he made up and employed successfully articulation lists consisting of meaningless monosyllables.

This brief note is intended only to enumerate without elaboration

the more outstanding of Dr. Campbell's contributions to the art of electrical communication as they fit into the history of that art. Their diversity is such as to establish the unusual versatility of Campbell's genius. His is a career unusually productive of discoveries, inventions and patents. Many of his important memoranda, however, were never worked up in the detailed form which would render them suitable for publication, and still reside only in the Company's engineering files. It would be regrettable to pass this occasion by without some notice being taken of these unpublished documents, and perhaps as fitting a commemoration as any is to print a few of the briefer ones just as they were written. Choosing somewhat at random, we are selecting the above cited memorandum of March 7, 1912, in which the 22-type repeater and the four-wire circuit are suggested, and two memoranda of earlier dates discussing capacity unbalances and crosstalk. The memorandum on repeaters is perhaps particularly interesting because of its historical flavor. Written twenty-three years ago, it refers to the measurement of attenuation in miles of cable, not decibels, and to such considerations as the natural period of the mechanical repeater diaphragm.

Dr. Campbell's Memoranda of 1907 and 1912

Introductory Note

As mentioned in the preceding article by Dr. Jewett, the first and second of the three following memoranda were the basis of methods of designing transpositions for voice-frequency circuits. They applied particularly to non-loaded circuits but the theory was readily extended to cover loaded circuits. In an earlier and more general study written in 1904, Dr. Campbell considered the involved equations necessary to an exact solution of the crosstalk problem and deduced simplifying approximations and convenient artifices for avoiding lengthy derivations.

He first assumes a line having the circuits substantially perfectly balanced to each other by means of very frequent transpositions. He then considers the effect of an unbalanced condition in a short length of line such as might arise from an irregularity in wire or transposition spacing or an unbalanced series impedance which might be due to a poor joint. Dr. Campbell refers to such effects as "slight alterations in the impedances, mutual impedances and admittances of the system." He shows how the crosstalk can be readily computed if these alterations in impedance and admittance are known.

He then considers the case of a short untransposed length in which the coupling between circuits is systematic rather than accidental. He shows that the crosstalk in such a short length can be computed in terms of mutual impedances and admittances in just the same manner used for accidental coupling in a short length nominally perfectly balanced. He shows that the mutual impedance per unit length (which is substantially proportional to the mutual inductance) is a measure of the crosstalk effect of the magnetic field of the disturbing circuit and can be computed from a knowledge of the spacing and diameters of the wires of the disturbing and disturbed circuits. The mutual admittance per unit length is a measure of the crosstalk effect of the electric field of the disturbing circuit and is shown to be proportional to the "direct capacity unbalance" which may readily be measured or computed from measurements of the individual direct capacities. This notion of direct capacity unbalance which was deduced in the earlier memorandum of 1904 has been of the greatest usefulness in crosstalk problems both in open wire and in cable. The mutual admittance defined in this way takes account not only of the

charges on the wires of the disturbing circuit but also of the charges induced by the disturbing circuit on all other wires. Thus, the electric shielding effect of other wires is taken into account. Magnetic shielding is ignored since this is unimportant with a line transposed at intervals very small compared with the wave-length. Dr. Campbell gave data for comparing the relative importance of the electric and magnetic components of the disturbing field and showed that for severe exposures both effects are of importance.

The equations given in the latter part of the memorandum of September 14, 1907, formed a basis of transposition design. They show how the crosstalk in a long transposed line may be computed with sufficient accuracy by simply summing up the effects computed individually for each short element of line. This important approximation is discussed in some detail in the earlier memorandum of 1904. Dr. Campbell prophetically says, "It must, however, always be borne in mind that we are working only with a first approximation and that in certain cases it may be necessary to continue the investigation to a higher order of approximation." In making the approximation, Dr. Campbell was, of course, thinking of voice frequency telephone circuits and at such frequencies, if the interval between transpositions is sufficiently short to guard against noise due to irregular power exposures, that interval will be but a very small fraction of the wave-length and it is unnecessary to consider the second approximation or as Dr. Campbell says to calculate "crosstalk-of-crosstalk."

When transpositions were designed for carrier frequency operation up to 30 kc. it was obviously impracticable to make a transposition interval a very small fraction of the wave-length and "crosstalk-of-crosstalk" could not be ignored. In other words, it was necessary to consider the crosstalk in each short element of line from the disturbing circuit into all the other wires on the line, the propagation of these crosstalk currents (and charges) along the line, and their effect in inducing currents in the disturbed circuit in other short elements of line. This effect has been termed interaction crosstalk since it takes account of the interaction between elements of line instead of simply summing up individual effects in each element. Thus it indeed proved true that "in certain cases" it was necessary to continue the investigation to a higher order of approximation.

I. Crosstalk Formulæ for Non-Loaded Circuits *

Take first the simple case of two perfectly symmetrical uniform circuits having the same transmission constants, which terminate at

* Memorandum dated September 14, 1907.

the same places in sets having the same impedance as the lines. Transmission upon either circuit can then give rise to no crosstalk upon the other circuit. Circuits such as two well transposed pairs on a pole lead are here to be understood; that is, there may be any number of circuits in the system but the mutual impedances and admittances between conductors connect points which are equi-spaced with respect to the impedances in each conductor.

Now suppose that at a point distanced x from the transmitting end of the circuits, slight alterations are made in the impedances, mutual impedances and admittances of the system. The effect of each change will be small and the total effect will be approximately equal to the sum of the individual results. That is, we may neglect the second-order terms, or crosstalk of crosstalk. Furthermore, unbalancing one of the given circuits alone cannot produce crosstalk. It is necessary that both circuits be unbalanced simultaneously by a single change in the system. Now, adding impedances to either side of one of the given circuits or to any third circuit will not unbalance both of the original circuits. Mutual impedance or admittance between the two sides of any circuit does not unbalance the circuit. Mutual impedance or admittance, added between either of the given circuits and any third circuit of the system, will not unbalance both of the given circuits. This leaves admittance shunted directly from one given circuit to the other given circuit and mutual impedance between the two circuits as the only source of crosstalk.

Let the admittances added between the two circuits be a, b, c, d connected between conductors 1 and 3, 3 and 2, 2 and 4, 4 and 1, respectively, where conductors 1/2 form one circuit and conductors 3/4 form the other circuit. These admittances may be resolved into the sum and difference of four admittances, as shown by the following table:

1 to 3	$a=$	$\frac{(a+b+c+d)}{4}$	$+\frac{(a+b-c-d)}{4}$	$+\frac{(a-b-c+d)}{4}$	$+\frac{(a-b+c-d)}{4}$
3 to 2	$b=$	$+$	$-$	$-$	
2 to 4	$c=$	$-$	$-$	$+$	
4 to 1	$d=$	$-$	$+$	$-$	

By the principle of superposition the effect of the given admittances a, b, c, d will be practically the same as the sum of the effects of the four component admittances taken individually. The first component admittance $(a + b + c + d)/4$ is added symmetrically between the two wires of one circuit and the two wires of the other circuit. This will not disturb the symmetry of either circuit and will, consequently, not

give rise to crosstalk. The second admittance is added between the conductor 3 and conductors 1/2 and subtracted between conductor 4 and conductors 1/2. This does not destroy the symmetry of circuit 1/2, and it can in consequence not give rise to crosstalk. The third admittance $(a - b - c + d)/4$ is added in the same way as the second with an interchange of circuits. It will also not give rise to crosstalk.

Crosstalk due to the added admittances, a, b, c, d, must therefore be due to the last component $(a - b + c - d)/4 = Y/4$ where Y is what we may call the direct admittance unbalance.

In order to determine the crosstalk occasioned by this admittance unbalance Y, when the electromotive force E is impressed upon one of the circuits, we may proceed as follows:

The circuits are connected as shown by Fig. 1. This is equivalent

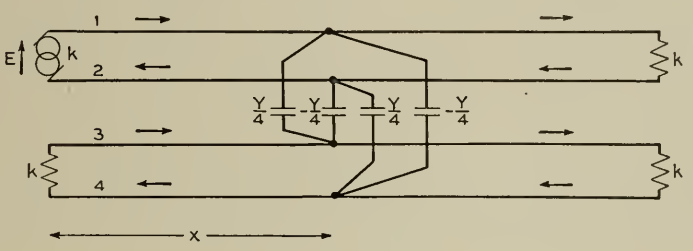


Fig. 1.

to the bridge of Fig. 2. For if the unbalancing admittances $Y/4$ were removed circuit 1/2 would be clear. Then as the e.m.f. E acts through an impedance k upon a line whose impedance is k, the potential difference at the sending end of the line would be $E/2$ and in traversing a distance x this would be attenuated by the factor $e^{-\gamma x}$. The im-

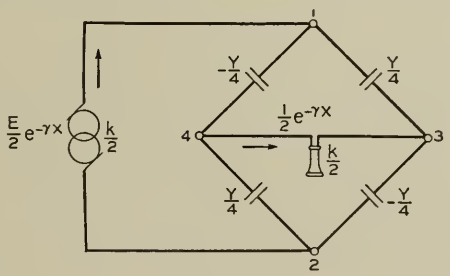


Fig. 2.

pedance of each end of circuit 1/2 at the point x will be k and therefore the entire circuit with its two ends in parallel will have the impedance $k/2$. We may therefore replace circuit 1/2 of Fig. 1 by a branch in

Fig. 2 having the impedance $k/2$ and containing an e.m.f. $Ee^{-\gamma x}/2$. Similarly circuit 3/4 will be replaced in Fig. 2 by a branch having the impedance $k/2$. As the current reaching circuit 3/4 will divide equally between the two ends of the line and the part reaching the beginning of the line will be further attenuated by the factor $e^{-\gamma x}$ the termination k in Fig. 1 is to be replaced by the receiver in Fig. 2 which indicates only $e^{-\gamma x}/2$ of the current flowing through it. Substituting these values in the expression for the galvanometer current in a bridge [1] we find for the crosstalk current

$$\Delta I_y = -\frac{EY}{16} e^{-2\gamma x} \left[\frac{1}{1 - \frac{k^2 Y^2}{64}} \right]$$

or approximately

$$\Delta I_y = -\frac{EY}{16} e^{-2\gamma x} \text{ as } Y \text{ is very small.}$$

This is the current at the end corresponding to the transmitting station. At the other end the attenuation factor will be that corresponding to transmission over the entire length of line l—making the crosstalk

$$\Delta I_y' = \frac{EY}{16} e^{-\gamma l}.$$

It will be noticed that the crosstalk at the farther end of the line is independent of the position of the admittance unbalance, while the crosstalk at the transmitting end of the system will diminish as the point of unbalance is moved farther from this end. In one case the wave must traverse the entire distance between terminal stations; in the other it must travel down the line to the point where it is carried across from one circuit to the other and then back from this point to the beginning of the line where the crosstalk is received.

The mutual impedances between the four conductors composing the two circuits may be treated in a manner similar to that which has been employed for the admittances between these conductors. Assume that any four mutual impedances are added and then divide them into four components, of which three may be shown to give rise to no crosstalk. The remaining component is the mutual impedance unbalance $Z/4$ and the crosstalk due to it may be found as follows. The circuit is shown by Fig. 3, which may be replaced by the trans-

[1] Maxwell I, § 347.

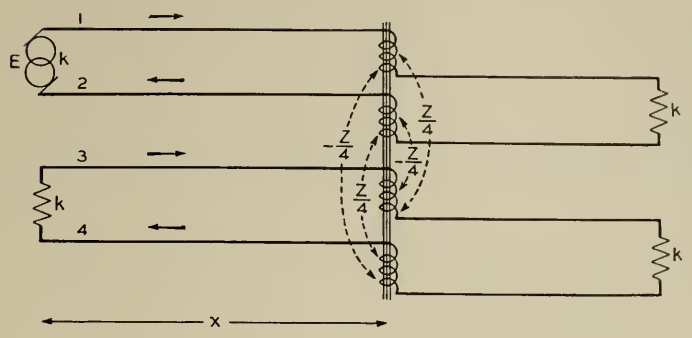


Fig. 3.

former circuit of Fig. 4. If there is no unbalance Z so that we have the original uniform system the current $E/2k$ starts out on 1/2 and at the point of mutual impedance unbalance it becomes $Ee^{-\gamma x}/2k$. Circuit 1/2 therefore behaves like a primary of impedance $2k$ (for the two ends of the line are in series) containing e.m.f. $Ee^{-\gamma x}$. The circuit

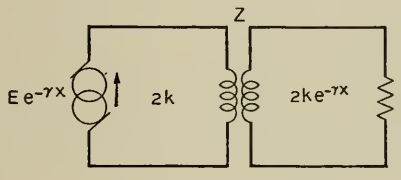


Fig. 4.

3/4 acts as a secondary of impedance $2k$. If the mutual impedance between the two lines is now made Z without any other change the current on 3/4 at x is

$$-\frac{EZe^{-\gamma x}}{4k^2 - Z^2},$$

which is attenuated by the factor $e^{-\gamma x}$ in reaching the transmitting end of the line, making the crosstalk

$$\Delta I_z = -\frac{EZ}{4k^2} e^{-2\gamma x}\left[\frac{1}{1 - \frac{Z^2}{4k^2}}\right]$$

or approximately

$$\Delta I_z = -\frac{EZ}{4k^2} e^{-2\gamma x} \text{ as } Z \text{ is small.}$$

At the farther end of the listening circuit, the approximate expression is

$$\Delta I_z' = -\frac{EZ}{4k^2} e^{-\gamma l}.$$

The total crosstalk due to the admittance and mutual impedance unbalance Y and Z is thus:

$$\Delta I = -E\left[\frac{Y}{16} + \frac{Z}{4k^2}\right] e^{-2\gamma x} \text{ at the transmitting end,}$$

$$\Delta I' = +E\left[\frac{Y}{16} - \frac{Z}{4k^2}\right] e^{-\gamma l} \text{ at the receiving end,}$$

from which we see that the ratio of the crosstalk due to the two kinds of unbalance is independent of the location of the point on the circuit at which these unbalances are introduced, and the two will be numerically equal in case:

$$Y = \frac{4}{k^2} Z$$

or

$$Y = \frac{4}{(667)^2} Z = 9 \cdot 10^{-6} Z, \text{ on non-loaded open wires.}$$

Appended tables give the direct capacity unbalances and the mutual impedance unbalances for a 40-wire lead of No. 12 wire. The relative importance of the two in producing crosstalk is shown below for a few typical cases:

	Unbalance per Mile		Ratio $\frac{9 \cdot 10^{-6} Z}{Y}$
	Direct Capacity Unbalance mmf.	Mutual Inductance Unbalance mh.	
Pairs adjacent on same cross-arm			
1/2 and 3/4	−1014	− 93	0.8
3/4 and 5/6	−1469	−143	0.9
Pairs adjacent in vertical plane			
1/2 and 11/12	602	72	1.1
3/4 and 13/14	409	72	1.6
5/6 and 15/16	704	147	1.9
Pairs separated by two steps			
1/2 and 5/6	− 116	− 31	2.4
3/4 and 7/8	− 84	− 19	−2.0
1/2 and 13/14	− 93	2	−0.2
3/4 and 15/16	− 150	4	−0.2
1/2 and 21/22	136	20	1.3
5/6 and 25/26	46	43	8.4

It will be noted that the two are of about equal importance for the cases of most severe static exposure. For 5/6 and 15/16, which is quite a severe exposure, the magnetic is twice as important as the static. For the pairs which are so far removed as to bring in considerable static shielding from the other wires the magnetic crosstalk may be still more important relatively. Thus for 5/6 and 25/26 it is eight times as great as the static.

At the sending end of the system the static and the magnetic crosstalks combine, while at the other end they tend to cancel each other. In practice, therefore, the summation is the more important case.

If both unbalances are pure reactances, the one being pure capacity and the other pure inductance, and the line has approximately the same impedance at all frequencies, the character of the crosstalk will be the same whether it is produced by capacity or mutual inductance. This will be approximately the case on well insulated open-wire lines. On non-loaded cable circuits the line impedance decreases as the frequency rises. On cables, therefore, the crosstalk due to mutual impedance will have the higher frequencies more strongly pronounced than the crosstalk due to capacity.

For a transposed line we find the total crosstalk by integrating the crosstalk throughout the entire length of the line. We will assume that the lines are infinitely long and that the transpositions give the system a periodic structure of lengths. Let x be the distance to the first transposition, a, b, c, $d \cdots s$, the distances from the transposition to the others in the periodic section.

$$I = -E\left(\frac{Y}{16}+\frac{Z}{4k^2}\right)\int_{0,\ x+a,\ x+a,\ \ldots}^{x,\ x,\ x+b,\ x+b,\ \ldots} e^{-2\gamma x}dx$$

$$= -\frac{E}{2\gamma}\left(\frac{Y}{16}+\frac{Z}{4k^2}\right)(1 - 2e^{-2\gamma x} + 2e^{-2\gamma(x+a)} \cdots - 2e^{-2\gamma(x+s)} + \cdots),$$

since the periodic section s must contain an even number of transpositions.

$$I = -\frac{E}{2\gamma}\left(\frac{Y}{16}+\frac{Z}{4k^2}\right)\left(1 - 2e^{-2\gamma x}\frac{1 - e^{-2\gamma a} + e^{-2\gamma b} - \cdots}{1 - e^{-2\gamma s}}\right)$$

or approximately

$$I = -\frac{E}{2\gamma}\left(\frac{Y}{16}+\frac{Z}{4k^2}\right)\left\{1 - 2[1 - 2\gamma x + 2\gamma^2 x^2]\right.$$

$$\times \frac{2\gamma(a-b+c\cdots)-2\gamma^2(a^2-b^2+\cdots)+\frac{4\gamma^3}{3}(a^3-b^3+\cdots)}{2\gamma s-2\gamma^2 s^2+\frac{4}{3}\gamma^3 s^3}\Bigg\}.$$

But the sum of alternate intervals in the transposition periodic section s must equal $s/2$. Therefore $a - b + c\cdots = a + (c - b) + (e - d) + \cdots = s/2$, and to the same approximation,

$$I = -\frac{E}{2\gamma}\left(\frac{Y}{16}+\frac{Z}{4k^2}\right)\Bigg\{1-[1-2\gamma x+2\gamma^2x^2]\times\frac{1-\frac{2\gamma}{s}(a^2-b^2+\cdots)+\frac{4\gamma^2}{3s}(a^3-b^3+\cdots)}{1-\gamma s+2\gamma^2s^2/3-\cdots}\Bigg\}$$

$$=\frac{E}{2}\left(\frac{Y}{16}+\frac{Z}{4k^2}\right)s\left[\left\{1-\frac{2x}{s}-\frac{2(a^2-b^2+\cdots)}{s^2}\right\}+\left\{\frac{1}{3}-\frac{2x}{s}+\frac{2x^2}{s^2}-2\left(1-\frac{2x}{s}\right)\frac{a^2-b^2+\cdots}{s^2}+\frac{4}{3s^3}(a^3-b^3+\cdots)\right\}\gamma s\right].$$

It will be noticed that the crosstalk varies linearly with the distance from the first transposition, approximately, and that by a suitable choice for this distance the crosstalk may be reduced to zero to the first approximation. This is, however, not a matter of especial practical importance, for incidental irregularities contribute to the crosstalk in practice. As soon as the crosstalk due to the regular transposition system is reduced to the order of that due to the accidental irregularities further reduction of this crosstalk is not a matter of commercial importance. The accidental irregularities in the distribution of the wires therefore set a limit to the extent to which it is worth while to reduce the length of the transposition sections.

If there are but two transpositions in the periodic interval $a = s/2$, $b = c = \cdots = 0$

$$I=\frac{E}{2}\left(\frac{Y}{16}+\frac{Z}{4k^2}\right)s\left[\left(\frac{1}{2}-\frac{2x}{s}\right)-\frac{x}{s}\left(1-\frac{2x}{s}\right)\gamma s\right]\text{ approximately.}$$

If $x = s/4$, I vanishes as to first order terms. If $x = s/2$ or 0, I has its maximum value,

$$\frac{E}{2}\left(\frac{Y}{16}+\frac{Z}{4k^2}\right)\frac{s}{2}.$$

II. Crosstalk Formulæ for Phantom Circuits *

There may be crosstalk between two phantom circuits, between a phantom and a distinct two-wire circuit, or between a phantom and one of its own side circuits.

In the first case we may assume that each side of both phantoms is perfectly balanced with respect to every part of the system, as we are not here concerned with the crosstalk on the side circuits. The two wires forming the side of a phantom may then be treated as a single conductor. The sources of crosstalk will therefore be direct admittance unbalance and mutual impedance unbalance exactly as for ordinary pairs. The capacity unbalance in formulæ [1] is found from the 16 direct capacities between the two sets of four wires, and the mutual impedance unbalance is found from the 16 mutual impedances between the same wires, but this is to be divided by four in order to allow for the division of the current between the two wires on each side of both phantom circuits.

In the second case each side of the phantom circuit may be assumed perfectly balanced. In the computation of the capacity unbalance, 8 direct capacities enter. There are also 8 mutual impedances involved in the mutual impedance unbalance, and these must be divided by 2 in order to allow for the division of the phantom circuit current between the two wires.

The crosstalk between a phantom circuit and one of its side circuits differs materially from the others, as the use of the same conductors to form the side circuit and one side of the phantom circuit introduces two additional sources of unbalance. These are: unbalance in the impedance of the two conductors forming the side circuit, and unbalance in the direct admittance from the two conductors forming the side circuit to the system outside of the phantom conductors.

The assumed distribution of unbalances is shown by Fig. 1. Insert

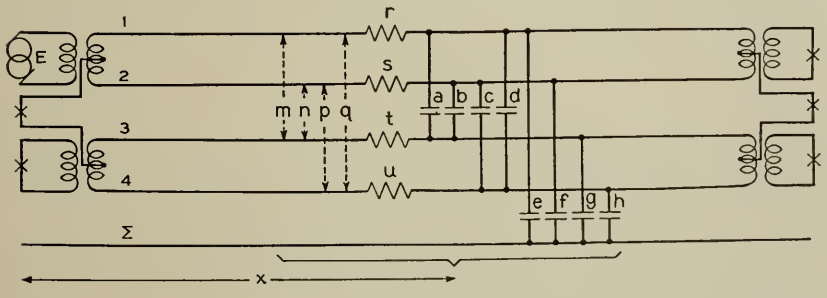


Fig. 1.

* Memorandum dated October 31, 1907.

[1] See September 14 memorandum.

an e.m.f. (E) at the sending end of circuit 1–2. Then at distance x conductor 1 will be at the potential $Ee^{-\gamma x}/4$ and carry the current $Ee^{-\gamma x}/2k$ before the unbalances are introduced. The potential of 2 and the current carried by 2 will be the same with sign reversed. Conductors 3, 4 and all others (Σ) in the system will be at potential 0 and carry no current. It follows at once that the impedances in 3 and 4 (t, u) and the admittances between these conductors and the conductors (Σ) of the system (g, h) will contribute nothing towards the crosstalk between 1–2 and the phantom. As equal impedances inserted in 1 and 2 will not unbalance the side circuit, the crosstalk must depend upon the difference between r and s and in consequence we may substitute $(r - s)/2$ for r in 1 and the negative of this in 2 without altering the crosstalk. Similarly, the effect of e and f depends entirely upon their difference, and we may substitute $\pm (e - f)/2$ for e and f. The direct capacities (a, b, c, d) may be resolved into four components,[2] and of these only the third unbalances both 1–2 and the phantom. The same applies to the mutual impedance unbalance. It follows that the crosstalk depends solely upon

$$\begin{aligned} X &= (r - s), \\ Y' &= (e - f), \\ Y'' &= (a - b - c + d), \\ Z &= (m - n - p + q), \end{aligned}$$

as indicated in Fig. 2.

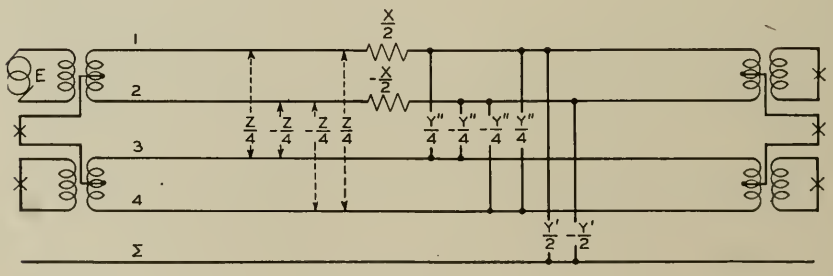


Fig. 2.

Substituting e.m.f. and currents for these in accordance with the rules for small changes, we have Fig. 3. This system is perfectly symmetrical with respect to the two wires in each side of the phantom and we may now treat the two wires on each side as one conductor, as in Fig. 4. The total e.m.f. around the phantom is $Ee^{-\gamma x}(X + Z)/4k$ and the impedance of the two ends of the phantom in series is $2K$.

[2] See September 14 memorandum.

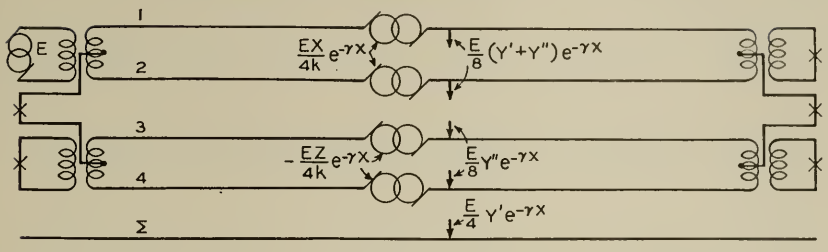


Fig. 3.

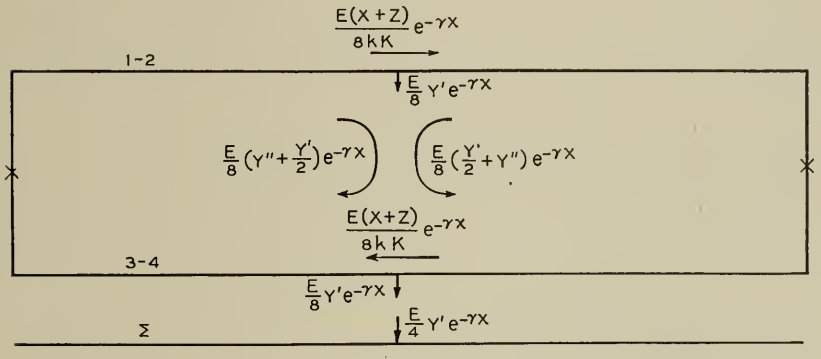


Fig. 4.

whence the current produced is $Ee^{-\gamma z}(X + Z)/8kK$. The currents entering and leaving the phantom may be resolved into $E(Y'/2 + Y'')e^{-\gamma z}/4$, leaving 1, 2 and entering 3, 4 and $EY'e^{-\gamma z}/4$, leaving 1, 2 and 3, 4 in parallel and entering Σ. The precise distribution of this last component depends upon the structure of the system, but as it flows symmetrically down the two sides of the phantom, it cannot introduce crosstalk. The first component divides equally between the two ends of the line, and this is the capacity unbalance current. The currents are therefore completely shown in Fig. 4. Allowing for the attenuation of the currents in reaching the ends of the line, we have the formulæ given below:

If a phantom (1, 2–3, 4) and its side circuit (1–2) have the transmission constants (K, Γ) and (k, γ) and the length (l) and are symmetrical throughout, then the crosstalk which will be introduced between them by the addition, at the distance (x) from the transmitting end, of small impedances (r, s, t, u) in the conductors, mutual impedances (m, n, p, q) between the conductors of the two side circuits and direct admittances (a, b, c, d; e, f, g, h) between the conductors of the two side circuits and between these conductors and the remainder

of the system is:

$$\Delta I = E\left(\frac{X+Z}{8kK} + \frac{Y}{16}\right) e^{-(\gamma+\Gamma)x} \qquad \text{at the sending end,}$$

$$\Delta I' = E\left(\frac{X+Z}{8kK} - \frac{Y}{16}\right) e^{-[(\gamma-\Gamma)x+\Gamma l]} \quad \text{at the distant end,}$$

where $X = (r - s)$,

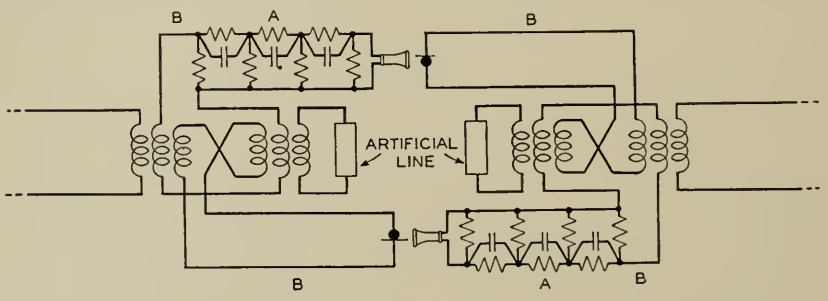
$$Y = 2(a - b - c + d) + (e - f),$$
$$Z = (m - n - p + q),$$

and the positive direction in both circuits is the same as in conductor 1.

III. Repeater Circuits *

The following points seem to merit an experimental trial.

1. A Two-way Repeater Circuit Including Two Repeaters, Each Operating as a One-way Repeater Only, as Illustrated by the Following Sketch



With this circuit the allowable unbalance is about double that with our present standard circuit. In addition to this, singing will not be introduced by any possible unbalance, however large, in either of the lines, provided the unbalance of the other line does not exceed a certain critical magnitude. Furthermore, the two lines connected together may differ radically in character since each is balanced separately against its own artificial line.

The present standard circuit or any one of several other repeater circuits may be substituted in place of the basic circuit shown in this sketch.

Although the circuit requires that all of the repeating apparatus be duplicated and that two artificial lines (of which at least one must be a close copy of the corresponding actual line as regards telephonic

* Dated March 7, 1912.

sending end impedance) be added, it still seems to me probable that the improved performance of the circuit will prove in this extra equipment.

2. The Use of Repeaters of Small Amplification at Periodic Intervals along the Line

Theoretically, a given total amplification can be secured with a larger singing margin if it is distributed among a number of properly spaced points along the line rather than concentrated at a single point. For example, four equally spaced repeaters each giving an amplification of five miles might be substituted for a single repeater giving twenty miles. If the circuit suggested by the above sketch were employed this would mean a total of eight repeater elements of which four would be used, one after another, as one-way repeaters in each direction. This raises the old question as to whether equally good quality can be obtained when several repeaters are used in securing a given amplification. This point seems worth further direct experimental investigation; one step in the right direction has probably been made by raising the natural period of the diaphragm.

3. The Use of a Compensating Device Such as an Artificial Line to Reduce the Amplification at the Resonant Frequencies to the Level of the Amplification at Other Telephonic Frequencies

In the sketch, equalizing artificial lines are shown at AA; obviously the same result may be secured by introducing them at any of a number of other points in the circuit. In this way the singing margin can be increased and the quality be somewhat improved, without materially reducing the telephonic amplification. But on general principles it would seem desirable to carry the equalization as far as possible in the repeater itself. The variability of the repeater sets a limit to what may be accomplished by any compensating device which reduces the total amplification by an invariable amount at each frequency and thereby increases the percentage variation. If it became necessary merely to eliminate certain frequencies lying outside of the range required for telephony, the use of an artificial selecting circuit would seem to present no difficulty.

The variation as well as the average amplification obtained from repeaters should be investigated. When these data have been obtained for the best type of repeater it will be possible to determine whether any material benefits can be derived by the introduction of compensating circuits.

4. Use of Two Lines for a Portion of the Route

The circuit shown above enables us to switch two-way transmission from a single line to a pair of lines and vice versa, since any amount of line may be inserted at *BBBB*. In connection with these lines any number of additional one-way repeaters may be inserted. The operation of the system is left unchanged beyond the change in the effective amplification which is equal to the difference between the repeater amplification and the attenuation of the inserted line. In case the total attenuation of the pair of lines exceeds the total amplification of all the repeaters at every frequency the system cannot sing whatever lines be connected at the ends. Practical applications will hinge upon the possibility of securing good quality from a number of repeaters used in sequence.

Suggestions 1 and 2 seem sufficiently promising to warrant some experimental work at an early date.

I am preparing a discussion of the general repeater circuit including any number of repeating elements and shall present the theoretical deductions applicable to the above suggestions in that memorandum.

Some Aspects of Low-Frequency Induction Between Power and Telephone Circuits *

By H. R. HUNTLEY and E. J. O'CONNELL

This article discusses the phenomena involved in low-frequency induction between power and telephone circuits and describes a demonstration which has been developed to illustrate certain of them.

Introduction

IN the practical problem of inductive coordination between power and telephone circuits, there are two aspects to be considered:

1. Induction within the frequency range used in transmitting speech which may result in disturbing noise in telephone receivers. This phenomenon is usually associated with the normal operation of power and telephone systems although abnormal conditions in either system may result in a large increase in the noise.
2. Induction at the fundamental frequencies used in the transmission of power. This is commonly referred to as "low-frequency induction" and in some cases may reach such magnitudes as to interrupt telephone service, constitute a hazard to telephone employees and produce other detrimental effects. Induced voltages of magnitudes sufficient to cause operating difficulties in telephone circuits occur usually only under abnormal power circuit conditions which produce large currents in the earth. Under normal circuit conditions, three-phase power circuits are so nearly balanced with respect to ground at their fundamental frequency that induction at this frequency is rarely sufficient to seriously affect well balanced telephone circuits.

Both types of induction have been and are being intensively studied cooperatively by the power and telephone industries. Much of this work has been handled through the Joint General Committee of the National Electric Light Association and Bell Telephone System which was formed in 1921 and it is now being carried forward by the Joint General Committee of the Edison Electric Institute and Bell Telephone System.

* This paper appeared in somewhat different form in *Amer. Railway Assoc. Proc.*, June, 1934, under the title "Demonstration of Low-Frequency Induction Between Power and Telephone Circuits" by H. R. Huntley.

Since the inductive coordination of power and telephone plants inherently involves the characteristics of both systems as well as the physical relations between them, problems can be effectively handled only by joint consideration in each specific case. As pointed out in a previous article dealing with noise induction,[1] effective cooperative action depends upon an adequate mutual understanding of the principles involved in coordination. Many of these principles can be demonstrated using comparatively simple apparatus.

This article describes a demonstration that has been developed to illustrate some of the more important factors concerned with low-frequency induction between power and telephone circuits, together with a discussion of this subject along the lines which would be followed in presenting the demonstration.

Demonstration Apparatus

The demonstration apparatus consists of two separate arrangements as follows:

(*a*) For many of the demonstrations, a miniature inductive exposure, consisting of a three-wire power line and a two-wire telephone line can be used. The power line is energized at a comparatively low voltage from a three-phase bank of transformers. The telephone line can be grounded at either or both ends and by means of a voltage measuring device, consisting of an amplifier and a projecting meter, a qualitative indication of the voltage along it or between it and ground can be obtained.

(*b*) For other demonstrations, a fairly high voltage in the telephone circuit is required, but since it is impracticable to secure this voltage using the miniature inductive exposure, it is necessary to use an iron core transformer. In order to improve the safety conditions when using this higher voltage, the miniature lines are not used and the circuit is entirely separate from that used in the low-voltage demonstrations.

A power supply frequency of 60 cycles per second is used. The phenomena illustrated are, however, applicable for all other frequencies commonly encountered in power transmission and distribution circuits.

In a demonstration of this kind, where the exposure is compressed into a small space and where the amount of power available is limited, it is obvious that the results can have no quantitative significance. This demonstration, therefore, is designed only to provide qualitative illustrations of some of the principles involved.

[1] "Some Theoretical and Practical Aspects of Noise Induction," by R. F. Davis and H. R. Huntley, published in *Bell System Technical Journal*, October, 1933.

Fundamentals of Problem—Magnetic Induction

Induction arises due to the fact that any wire transmitting electricity is surrounded by electric and magnetic fields which may cause voltages to appear on other wires in these fields. The relative strengths of the electric and magnetic fields depend on the characteristics of the circuit, the former being a function of the voltage on the circuit and the latter a function of the current in it. Induction due to electric fields is commonly called "electric induction" while that due to magnetic fields is called "magnetic induction."

When a ground occurs on a power line there are two factors which influence the induction into neighboring telephone circuits:

(*a*) The residual voltage is increased, which increases the electric induction.

(*b*) The residual current is increased, which increases the magnetic induction.

Both from theoretical analyses and experience it is known that magnetic induction is more important than electric induction in most cases of low-frequency induction. Consequently, the demonstration is concerned only with magnetic induction, i.e., induction due to the power system currents.

The magnetic field about a wire faulted to ground and carrying fault current is shown in Fig. 1. This magnetic field varies in proportion

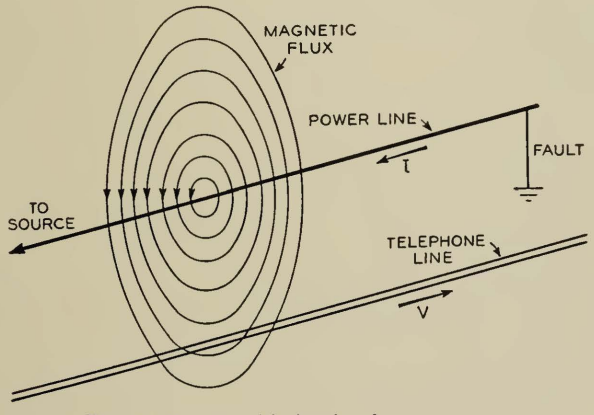


Fig. 1—Process of induction from currents.

to the current in the wire causing it. If other conductors, such as the pair of telephone wires shown in Fig. 1, lie within this field there are induced along them voltages proportional at every instant to the time rate of change of the magnetic flux which links the wires.

Consider first that the inducing circuit carries direct current. In such a case, as long as the magnitude of the direct current does not change, the magnetic flux is constant and no voltage will be induced in a paralleling telephone wire. However, if the current in the inducing wire changes suddenly, the flux about it changes in the same proportion and voltage is induced momentarily in the paralleling wire. The magnitude of this impulse of voltage will be proportional to the rate at which the flux changes and the voltage will last only as long as the flux is changing. Rapid changes in the current in direct current circuits, with consequent voltage impulses on paralleling telephone circuits, may occur when power apparatus is turned off or on. Also, of course, if a short circuit occurs, the current may rise very rapidly and then fall very rapidly as the circuit breaker operates. While the consideration of these direct current phenomena are important in some situations, they are not included in this demonstration and will not be further considered.

In the alternating current case, the current in the inducing wire is continually alternating so that the flux about it is continually alternating. Consequently, there will be induced in a paralleling wire, an alternating voltage proportional to the inducing current. It should be noted particularly that the induced voltage acts *along* the wire rather than between the wire and ground.

General Nature of Phenomena

In applying these principles of magnetic induction to the low-frequency induction problem, only the conditions which exist when a power circuit is faulted to ground and before the current is interrupted (usually by the operation of circuit breakers) need be considered. The current of interest during this time is that which flows out over the power line wires and returns through the ground, called "residual" current. The voltage induced is along the telephone wires in parallel. Since the telephone circuits are metallic, the talking paths over them are usually not seriously affected by the fundamental frequency voltage unless this voltage causes the telephone protectors to operate. Service over grounded telegraph circuits may, however, be impaired even if the induced voltage does not reach values high enough to operate protectors, and the telephone circuits under this condition may be made noisier than usual.

It can be seen that the electrical phenomena in which we are interested will be affected by three basic factors. The first of these is concerned with the "magnetic coupling" between the power and telephone lines, considered with ground return for the reasons pointed

out above. This coupling is a function of the strength and frequency of alternation of the flux set up at the location of the telephone line by a given amount of ground return current in the power line. The second factor is concerned with the amount of ground return current in the power circuit at the time of a ground fault since, for a given coupling, this will determine the strength of the magnetic field. The third factor is concerned with the conditions in the telephone plant which determine its reactions to a given induced voltage. Each of these three factors is taken up individually in the following discussion.

Since low-frequency induction between power and telephone circuits involves a series of separate and distinct occurrences, it is evident that

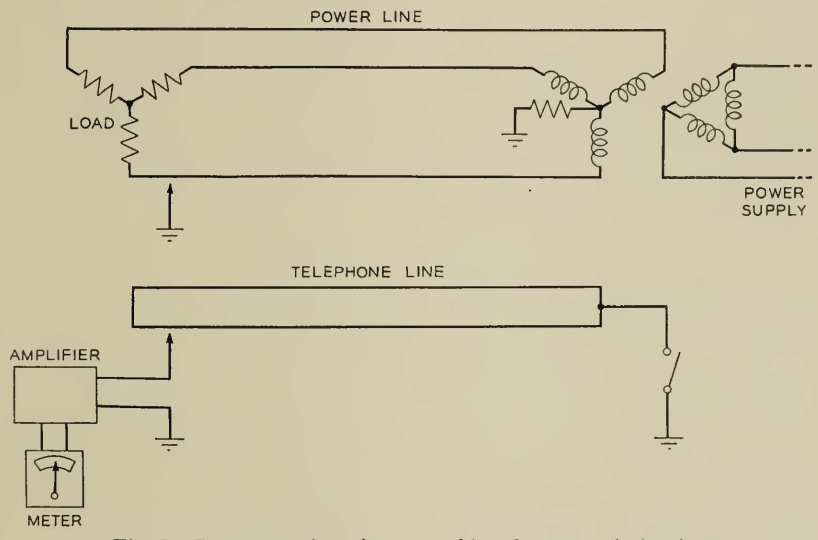


Fig. 2—Demonstration of nature of low-frequency induction.

the frequency with which ground faults occur on the power lines, the locations and circumstances of the various faults, the particular conditions in the telephone system at the time of such faults, etc., are also important. These matters are all subject to fortuitous variations so that there are many probability factors that must be considered in the study of any low-frequency induction problem. These probability factors cannot be demonstrated with the apparatus available.

In order to illustrate the general nature of the electrical phenomena, Fig. 2, which shows parallel power and telephone systems with the power line supplying a load, has been simulated using the miniature

lines. The load current may be substantial in magnitude but is normally confined to the line wires and is commonly called "balanced current." Under this condition, the induced voltage is small. If, now, one of the power conductors is grounded, current will flow to ground at that point. This current is the "residual current" mentioned previously. When this is done in the demonstration in such a way that the residual current flows through the exposure, the voltage induced along the telephone circuit rises very materially. Also, since the induced voltage acts along the telephone circuit, opening the ground connection at the far end of the telephone line reduces the voltage shown by the meter to a very small amount.

That the power current causing the induction is unaffected by transpositions in the power circuit, can be shown by transposing the power circuit in the set-up. No appreciable change in the induced voltage occurs when this is done. Likewise the induced voltage, since it is induced along the telephone wires in parallel, is unaffected by telephone circuit transposition, as can be shown by transposing the telephone circuit. Consequently, the matter of power or telephone circuit transpositions can be neglected in the further analysis.

Coupling Factors

Using the demonstration arrangements, some of the basic factors in coupling can be observed. For example, since the voltage is due to magnetic induction and accumulates along the telephone circuit, the coupling should be proportional to the length of the (uniform) exposure through which the fault current flows. This can be observed by noting the reduction in induced voltage as the fault on the power line is moved from the end of the exposure toward the supply end. (In the demonstration, the fault current is the same regardless of the location of the fault.)

Likewise, if the voltage accumulates along the telephone circuit, the longitudinal voltage measured should be proportional to the length of telephone circuit exposed. This can be observed by again placing the fault on the power line at the end of the exposure and moving the measuring point along the telephone line. As this point is moved toward the grounded end, the indicated voltage goes down.

Another basic factor in coupling is its relationship to the separation between the lines. Generally speaking, the greater the separation, the smaller the coupling. How much coupling will exist for a given separation depends on a number of factors, one of which is the structure of the earth. This effect will be discussed first.

In the type of problem we are considering the telephone wires comprise one side of a long loop, the other side of which is the earth. Likewise, the power wires comprise one side of a loop, the other side being the earth. It is a well known fact that the magnetic coupling between two parallel loops at a given separation increases as the sizes of the loops increase. The sizes of the loops in the case being considered are determined by the distribution of the return current in the earth.

A great deal of theoretical and experimental work has been done in connection with the analysis of the distribution of current in the earth, and it has been found that one of the important factors is the "resistivity" of the earth. The effect of resistivity of the earth can be briefly summarized as follows:

(*a*) Considering the outgoing and return paths for residual current on a power line, the mutual induction between the current in the wires and the return current in the earth tends to pull the earth currents together and to concentrate them under the line as near the surface as practicable. This action tends to decrease the coupling to an adjacent circuit by decreasing the effective separation of the sides of the loop.

(*b*) The resistance which the current encounters in flowing through the earth tends to make it spread out because, by so doing, the current density is reduced and the voltage drop is consequently reduced. This spreading out tends to increase the coupling to an adjacent circuit.

(*c*) The net distribution of the current in the earth is a balance between these two opposing tendencies and this distribution will be different for different resistivities of the earth. Generally speaking, the greater the resistivity of the earth, the more the current will spread and the greater will be the coupling to an adjacent circuit.

Figure 3 is a graphical representation of how the return current in the earth tends to spread with an increase in earth resistivity. While this figure shows only the vertical spread, a similar spreading also takes place horizontally.

The effect of the sizes of the primary and secondary loops on the coupling is greater when the loops are widely separated than when they are close together. For this reason, the effect of earth resistivity on coupling is much greater for wide separation exposures than for exposures where the lines are close together. Consequently, with high resistivity earth, the coupling not only is higher at all separations

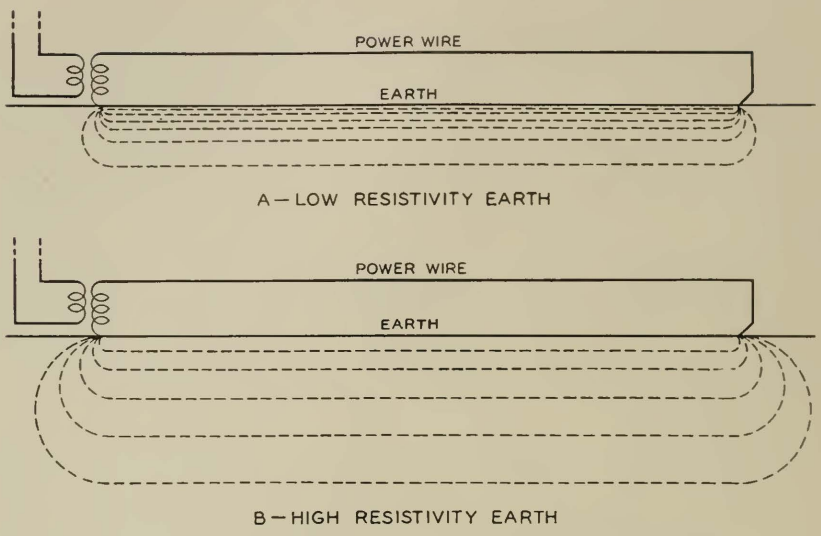


Fig. 3—Current distribution in earth of different resistivities.

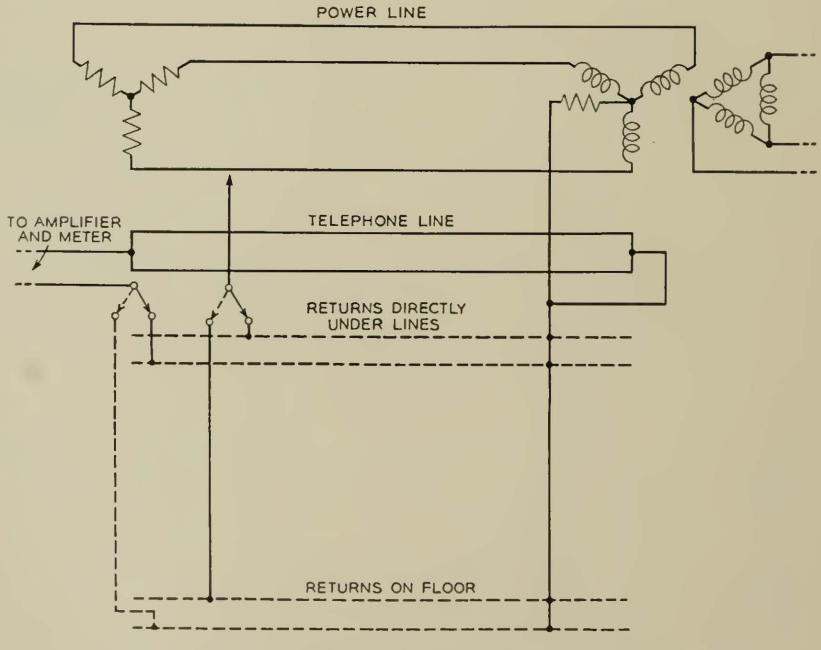


Fig. 4—Demonstration of effects of depth of return current in earth.

than with low-resistivity earth, but (except for very wide separations) the percentage reduction secured by increasing the separation a given amount is smaller.

Figure 4 shows a schematic of a demonstration set-up to show some of these effects. It is, of course, impracticable to employ earth of different resistivities in a demonstration of this kind, but the difference between return current which concentrates near the line and that which is more remote can be roughly indicated. In order to illustrate the fact that when the currents are concentrated closely under the line, the coupling falls off rather rapidly as the separation is increased, returns immediately under the lines are used and the telephone line is moved to change the separation. To illustrate that a wider distribution of current in the earth tends to increase the coupling and to make it less affected by separation, returns on the floor are used for both the power and telephone lines. It can be shown that:

(*a*) The induced voltage increases when the connections are changed from the upper to the lower returns.

(*b*) When the upper returns are used, *the percentage reduction* in induced voltage when the separation is increased, is greater than the percentage reduction when the lower returns are used and the separation is increased by the same amount, i.e., when the telephone line is moved between the same positions of minimum and maximum separation.

If the earth is not homogeneous, that is, has strata of different resistivities, the distribution of the earth current is distorted and varying effects are noted. Where local irregularities exist, marked and sometimes erratic changes in coupling may occur within comparatively short distances. An "effective" earth resistivity can usually be determined by test even where the earth is stratified.

Another important factor in determining the net coupling between power and telephone circuits is the effect of grounded wires or other linear grounded metallic structures along the inductive exposure. Voltages are induced in such grounded metallic structures in the same way as voltages are induced in telephone wires and these voltages cause currents. The magnetic fields accompanying these currents generally oppose those from the power wires and reduce the induction in the telephone circuit. The effect of such currents in grounded structures is generally spoken of as "shielding."

The amount and phase of the current in a grounded conductor in a given location and hence the shielding provided by it depend on the impedance of the conductor with earth return. Hence the shielding is

increased when the resistance of the conductor and its ground connections is reduced.

In order to illustrate these effects, the demonstration shown in Fig. 5 has been set up. With the shield wire on the power line, the shielding effect can be shown under two conditions as follows:

(*a*) When the switch directly grounding the shield wire is closed the induced voltage in the telephone circuit goes down materially due to the shielding effect of the current in the shield wire.

(*b*) When, instead of grounding the shield wire directly, it is grounded through a small resistance, the reduction in induced voltage is much smaller because the resistance limits the current in the shield wire.

In order that a conductor may exert a shielding effect, it must have a substantial coupling to either the power or telephone line; i.e., it must be fairly close to one or the other. By moving the wire shown in Fig. 5 it can be demonstrated that:

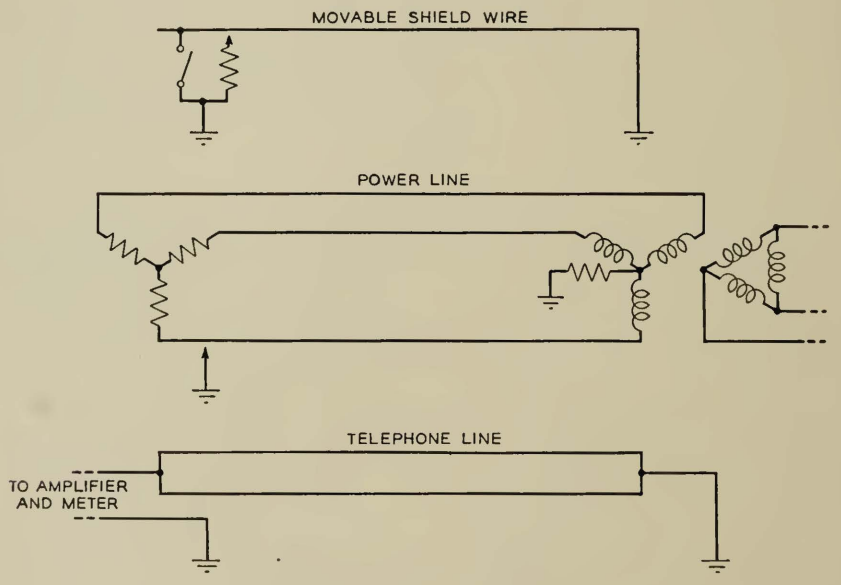


Fig. 5—Demonstration of shielding effect of grounded wire.

(*a*) With the shield wire on the power line or on the telephone line, substantial shielding is secured.

(*b*) If the wire is moved outside the exposure, the shielding is reduced to a small value.

While the demonstration shows only the effect of a grounded wire on or near the power or telephone lines, similar effects in varying degrees may be caused by such grounded metallic structures as underground pipe lines, railroads where the rails are bonded in long lengths, trolley lines, etc. In many situations the shielding effects of such structures may be substantial.

Another type of grounded conductor which may give substantial shielding is the metallic sheath of a telephone or power cable. A cable sheath will effect some shielding on conductors which are not enclosed by it in the same way as any other grounded metallic conductor, but the major shielding effect is experienced on conductors within the sheath. The shielding effect of a cable is, as in the case of a shield wire just demonstrated, determined to a considerable extent by its impedance with ground return.

Shielding due to a telephone cable can be demonstrated using the set-up shown in Fig. 6 and it is noted that:

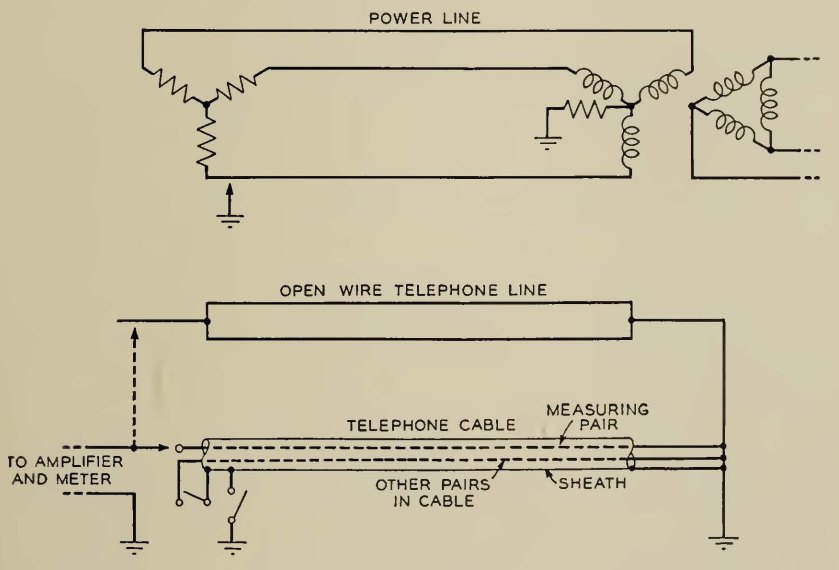


Fig. 6—Effect of cable shielding.

(*a*) The voltage along a conductor inside the cable is reduced when the sheath is directly grounded at both ends.

(*b*) If the effective resistance of the sheath is reduced by paralleling with it some of the conductors inside, the shielding is increased (i.e., the reduction in voltage is greater).

(*c*) When resistance is added in one of the sheath-to-ground connections, the shielding effect of the sheath and conductors is reduced.

(*d*) If the shield wire is grounded its effect is cumulative [2] with that of the cable sheath and conductors.

(*e*) The voltage along the open wires on the line is reduced when the cable is grounded at both ends. Here again, of course, the effect of the shield wire is cumulative [2] with that of the cable sheath and conductors.

The same shielding effects could be shown if the power instead of the telephone circuit were in cable. Also, of course, if there is more than one power or telephone cable, the shielding is increased. Iron armoring also tends to increase the shielding.

Power Circuit Conditions

Having illustrated some of the factors affecting coupling let us now review briefly the factors affecting the current in the power line at

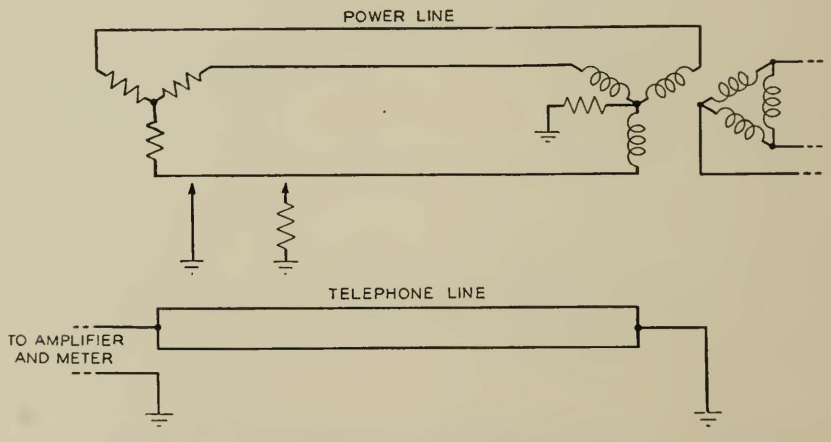


Fig. 7—Comparison of induced voltage with different amounts of residual current.

times of ground fault. Not only is the current flowing to ground under fault conditions on a power line usually greater than the current under normal operating conditions, but it is also a fact that a given current in a ground return circuit induces a larger voltage along a paralleling telephone circuit than the same amount of current if confined to the phase conductors. That the induced voltage depends on the amount of residual current can be demonstrated as shown in Fig. 7. Here a

[2] While the benefits are "cumulative," the individual effects are not directly additive due to mutual reactions between the different shielding conductors, the net effect being less than the total of the effects of each shielding conductor acting alone.

comparison is made between the induced voltage for two residual currents, one greater than the other. With larger current, the voltage indicated on the meter is larger. The resistance inserted in the fault to reduce the current in this demonstration might be thought of as simulating added impedance anywhere in the ground return circuit through the fault. For example, it might be thought of as simulating the effect of the line impedance which would be added if the fault occurred at some distance beyond the end of the exposure. Also, its effect is the same as would be produced by an increase in the reactance of the supply transformers or in the neutral-to-ground connection; or by an increase in the local resistance at the fault itself.

In analyzing the impedances further, there are two general types of power systems which must be considered: the "grounded neutral" and "isolated neutral" systems. These are illustrated in Fig. 8. In the grounded neutral system, the neutrals of one or more transformer banks are grounded directly or through impedance so that in the event of a fault, a path for current is established from the fault through the earth and back to the system through the neutral-to-ground connections. In the isolated neutral system there are normally no grounds on the system so that in the event of a fault the only path for fault currents is through the capacitances of the unfaulted phases to earth or through a second fault if one exists. Hence for a single fault, the fault current is limited to the charging and leakage current.

Figure 8-*D* shows for a grounded neutral system, the equivalent single-phase circuit for residual currents. In an actual line, the circuit conditions are, of course, usually much more complex than those shown. In even the simplest situations, there are usually other lines, generator points, or grounding points which supply some fault current. However, for the purpose of examining the fundamental phenomena, the simplified diagram can be used. As can be seen, the impedances which control the residual current are those associated with the fault, the line impedance, those in the transformer and generating equipment and the impedance, if any, in the neutral-to-ground connection. Impedances in any of these places tend to limit the fault current.

Figure 8-*C* shows a simplified diagram of the equivalent single-phase circuit of an isolated neutral system with a single-phase fault-to-ground. For this condition, it is evident that the fault-current path includes the capacitances to ground of the unfaulted phase conductors. In a small system these capacitances will be small and the fault current will, therefore, also be small, particularly if the voltage is not high. In extensive systems or systems having much

cable, the capacitances and hence the fault current may be fairly large, particularly if the voltage is high. Of course, in an actual system the capacitances to ground are distributed throughout the system so that the amount of residual current in the lines will vary from location to location, being a maximum at the fault and tapering to zero at the end of each branch of the system.

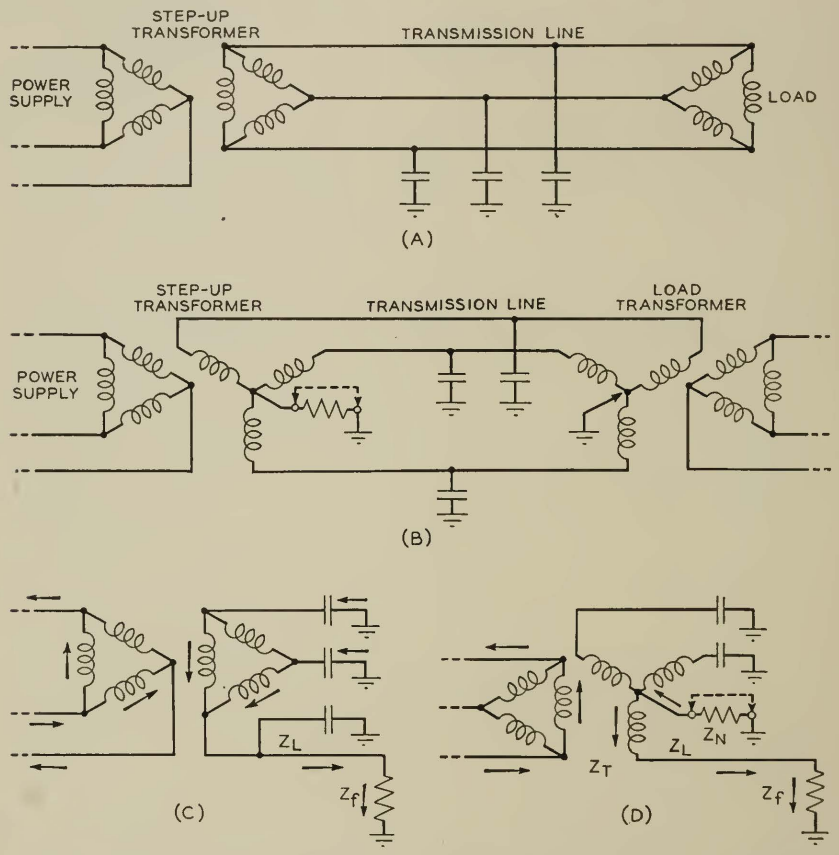


Fig. 8—Types of power systems. (*A*) Isolated neutral system under normal conditions. (*B*) Grounded neutral system under normal conditions. (*C*) Isolated neutral system with single fault. (*D*) Grounded neutral system under fault conditions.

If, in a power system, a second fault-to-ground occurs on another phase while the first persists, a large residual fault current will exist in the line between the faults even if the neutrals are isolated. Simultaneous faults on two phases at different points may occur on any type of system, but are more likely to occur on an isolated neutral system than on one in which the neutral is solidly grounded. This is

due to the fact that for the isolated system, full phase-to-phase or possibly higher voltage is impressed between the unfaulted phases and ground, thus increasing the voltage stress on the insulation of the entire system during the time of fault.

Figure 9 shows a demonstration set-up to illustrate the effects of faults on an isolated neutral system which is small enough so that the capacitances are negligible. It will be noted that when a single fault

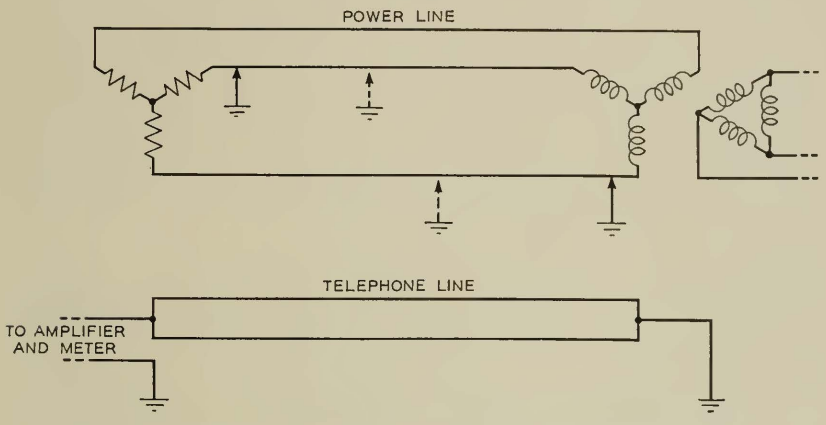


Fig. 9—Demonstration of effect of faults on isolated neutral system.

is put on the line at the far end, no appreciable rise in voltage on the telephone line occurs. However, when a second fault is put on another phase at the near end, the induced voltage immediately rises. In order to illustrate that the current is residual only between the faults, the faults can be moved closer together. When this is done, the induced voltage decreases until, with the faults at the same point, it practically disappears. The small remaining voltage is largely due to balanced current induction due to the heavy load on the power line when two of the phases are shorted.

A system grounded through a neutral impedance, such as when resistance or reactance is included in the neutral-to-ground connections or when a high reactance grounding bank is used, partakes of some of the characteristics of an isolated neutral system. Generally speaking, the addition of neutral impedance tends to reduce the fault current, this effect being proportionately larger for faults near the neutral grounding points. This reduction in fault current tends to reduce the voltage induced on nearby telephone lines and in some cases may reduce the "shock" to the power system and the damage at the point of fault. On the other hand, increasing the neutral impedance tends

to increase the difficulty of securing adequate selectivity in power system relay operation and to make a more complex relay system necessary. It also tends to increase over-voltages on the power system and to reduce the factors of safety for lightning arrestors.

In addition to the magnitude of the residual current, its duration is of importance since the length of time that the induced voltage persists on a telephone circuit has important reactions on its effects. For example, the chance of permanently grounding telephone protectors, with consequent interruption of service until the protector blocks are replaced, depends not only on the amount of current through the blocks but upon its duration. Likewise, many of the other effects, which are described later, are materially affected by the duration of the induced voltage. Since, except for self-clearing faults, the duration of fault current is determined by the time of operation of the relaying system, the reliability and speed of operation of the latter is an important factor. There are many types of relaying systems and it is not practicable to go into a discussion of them here except to point out that rapid and reliable relaying is usually simplest on a solidly grounded neutral system. For systems with large impedances in the neutrals, it may be difficult to secure rapid fault clearance, particularly if the system layout is complicated. For isolated neutral systems, rapid relaying on ground faults may be very difficult or impracticable.

Telephone Circuit Conditions

The voltage due to magnetic induction accumulates *along* the telephone circuit and can be represented as a voltage in series with the telephone wires. Figure 10 shows schematically how this voltage acts. The two sides of the metallic telephone circuit are assumed to have the same induced voltage and impedance and are shown here replaced by a single equivalent conductor. The total voltage which is equal to the product of power line fault current and coupling is represented by a number of generators connected in series through impedances representing, in total, the line impedance inside the exposure. At the ends of the exposure are connected impedances representing those in the line and between line and ground outside of the exposure. The longitudinal induced voltage acting through the series and shunt impedances of the telephone line will produce the following conditions of interest:

(*a*) Voltages between the telephone wires and ground at various places along the telephone line.

(*b*) Current in the longitudinal telephone circuit.

(*c*) Voltages between different wires on the telephone line.

Telephone circuits are supplied with protective devices. The part of the telephone protective system of most interest in connection with low-frequency induction is the carbon-block protector. This device provides a small air gap between carbon surfaces one of which is connected to the telephone conductor and the other to ground or cable sheath. When an excessive voltage is impressed on the telephone

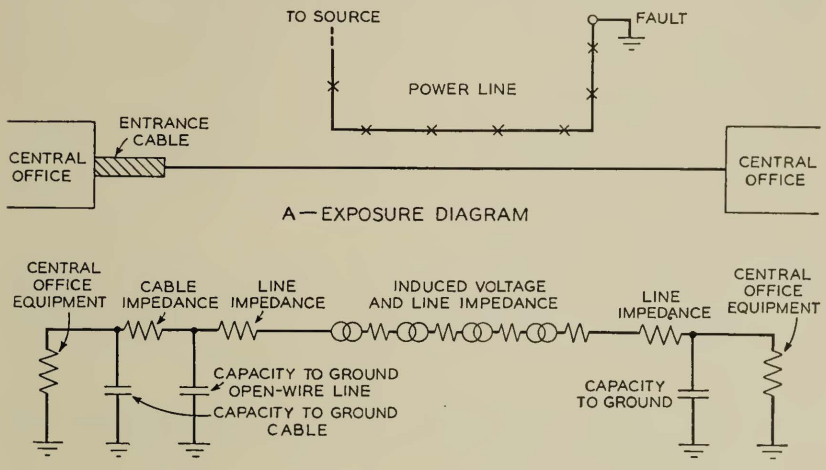


Fig. 10—Distribution of voltage to ground before protectors operate.

conductor an arc is established in the air gap thereby grounding the conductor. Protectors are located at central offices and at other points, such as at junctions of open wire and cable where it is desired to limit voltages on telephone wires due to lightning, contacts with power systems, induction, and other extraneous voltages.

In analyzing the distribution of induced voltage between a telephone circuit and ground assume first that no protectors are operated. Under this condition, the voltages to ground on the telephone wires at various points are determined by the impedances between the wires and ground along the line and at central offices where equipment is con-

nected to them. The voltage to ground at either end of the exposure is equal to the longitudinal current times the impedance-to-ground seen looking away from the exposure at that end. Figure 10-*C* illustrates how the voltages may distribute due to the distribution of impedances between the wires and ground along the line and in the central office equipment. Of course, in practice, the variety of impedance distributions encountered is almost infinite and the corresponding voltage distributions vary over a wide range.

If voltage-to-ground at any point where protectors are located exceeds the operating voltage of the protector, the protector operates and three things happen:

(*a*) The voltage-to-ground at the place where the protector operates is reduced to a low value. This makes the longitudinal voltage pile up at the protectors at the opposite end, and in most cases, they will also operate.

(*b*) The operation of the protectors at the two ends completes a loop consisting of the telephone circuit and ground so that the induced voltage will cause current to flow through both protectors.

(*c*) The voltages-to-ground on the circuits on which protectors have operated are changed and redistributed and the voltages on the other telephone circuits are also changed and redistributed due to shielding, as discussed later.

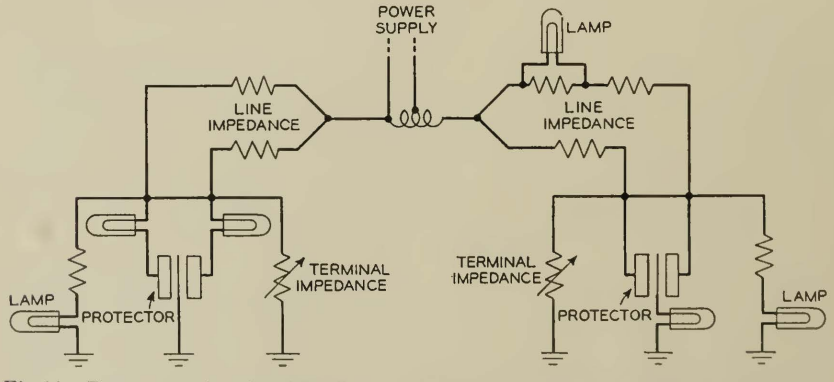


Fig. 11—Demonstration showing effect of terminal impedance on voltage distribution.

All of these effects take place within a very short time after the longitudinal voltage is applied so that for all practical purposes they can usually be considered as being instantaneous.

In order to illustrate these phenomena, the demonstration shown in Fig. 11 can be used. In this demonstration the longitudinal voltage

is impressed by means of a power transformer in order to secure a sufficiently high voltage. A general idea of the relative magnitudes of the voltages at the two ends can be had by observing the brilliancy of the voltage measuring lamps. By varying the slide wires which control the terminal impedances the proportions of the total voltage which appear at either end are changed and an idea of the changed distribution can be obtained by observing the changing glow of the measuring lamps. Finally the voltage at one end can be increased enough to cause the protectors at that end to operate, whereupon the measuring lamp goes out and the small protector lamp lights. Immediately the other protectors operate, as evidenced by the voltage lamp going out and the protector lamp lighting, and the line current increases as evidenced by the brilliance of the line current indicating lamp.

An important factor in the further analysis is the characteristics of the telephone protector. The arc takes place between two carbon surfaces. The gap between these two surfaces has a very high breakdown speed and a very low impedance after it is broken down. Another important characteristic from the standpoint of low-frequency induction is its tendency to become permanently grounded if heavy currents are discharged or if the discharge continues for some time. Consequently, the amount of current in the longitudinal circuit in the event of a breakdown and its duration are important factors in determining the chance of permanently grounding the protectors and causing the circuit to become inoperative until the blocks are changed. Duration is, of course, ordinarily a function of the duration of fault current on the power line as pointed out previously.

The amount of current through operated protectors is determined by the longitudinal voltage and the longitudinal impedance of the telephone circuit. If, for the moment, it is considered that only one wire is present, this current is simply the total longitudinal voltage divided by the total series impedance of the wire plus any resistance in the protector grounds. This can be seen from Fig. 11.

Ordinarily there are numerous circuits on an open-wire telephone lead or in a telephone cable. If the protectors on a number of these circuits break down, the current in each wire will be less than that which would exist were only one wire present as in the above illustration. This is due to the mutual impedance between the different telephone wires which causes the current in any one wire to reduce the current in the remaining wires. The *total* current in all of the wires of course increases as the number of wires on which protectors have operated is increased but not in direct proportion. Figure 12 illus-

trates this effect. The actual circumstances concerned in this phenomenon are, of course, that the wires which become grounded at their terminals through the operation of the protectors exert a shielding effect in exactly the same way as any other grounded conductor.

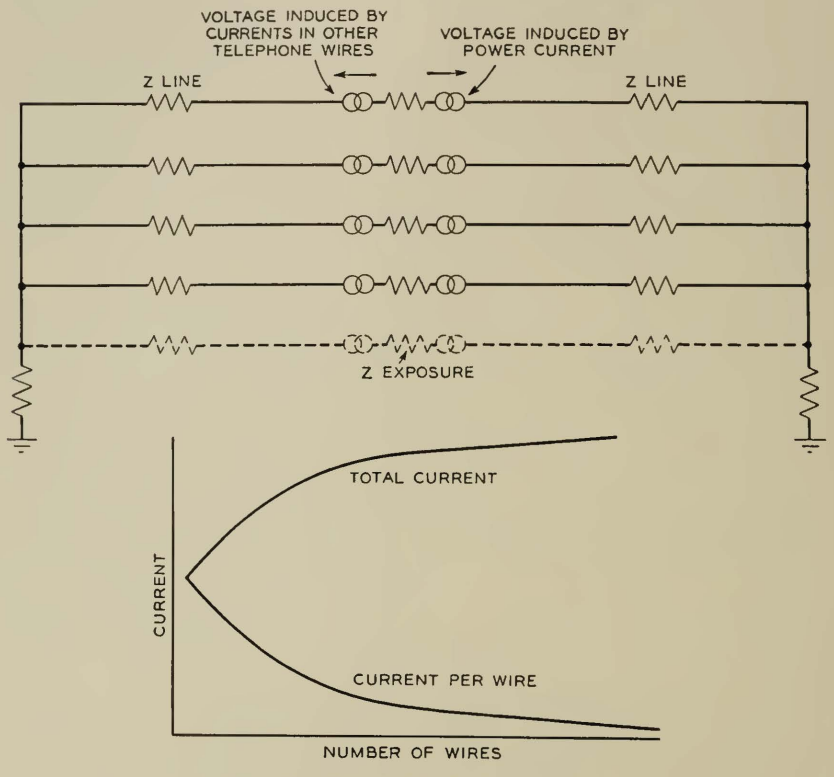


Fig. 12—Current in telephone wires.

The resistance of an individual wire is relatively high but if numerous protectors operate, the shielding may become fairly great, due to the closeness of the wires to each other and to the fact that a substantial amount of copper may be involved. Of course, this shielding is obtained at the expense of at least momentary interruption of the circuits on which protectors operate.

The shielding effect of current in grounded telephone wires is exerted on all telephone wires on the line regardless of whether the protectors on them have or have not operated. Consequently, what may happen on a large telephone line with a moderate induced voltage on it is that enough telephone protectors on different circuits operate to give

a shielding effect on the remaining wires sufficient to reduce the voltages on them to values lower than will operate the protectors.

Another important factor is the voltage-to-ground at various places along the telephone circuit after protectors operate. With the protectors operated the voltage-distribution-to-ground can be evaluated from the longitudinal induced voltages, the longitudinal currents, and the series impedances in the circuit. As the simplest and perhaps most striking case, consider a telephone circuit which is solidly grounded due to operated protectors at the ends of a uniform exposure with a fault on the power line at the end of or beyond the exposure. This situation is illustrated in Fig. 13. The distribution of induced

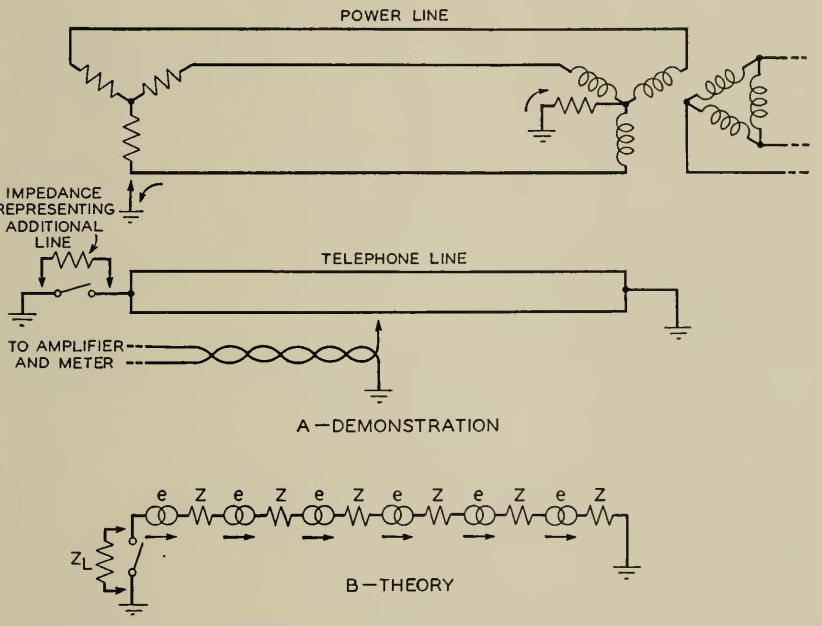


Fig. 13—Voltage to ground for telephone line grounded at ends where fault is outside of exposure.

voltage along the line is uniform and the longitudinal current is equal to the total longitudinal voltage divided by the total series impedance. If the net drop in voltage is taken from either end to any point along the circuit, it will be found that the induced voltage accumulated over this distance is equal and opposite to the voltage drop over this same distance due to the current flow through the impedances in this section. Consequently, under these conditions the voltage-to-ground is zero at all points along the circuit. This is true regardless of the

magnitude of the induced voltage. Figure 13 also shows the set-up by which this fact can be demonstrated. It will be noted that, while there is a fairly high longitudinal voltage, the voltage between the wires and ground with both ends of the circuit grounded is negligible at all points along the line.

If the telephone line extends beyond the exposure, the effect of this portion of the line is to add impedance between the exposure terminal and the protector without adding a corresponding induced voltage. If a power line fault occurs at the end of the exposure, a voltage-to-ground will exist at this point equal to the current in the telephone

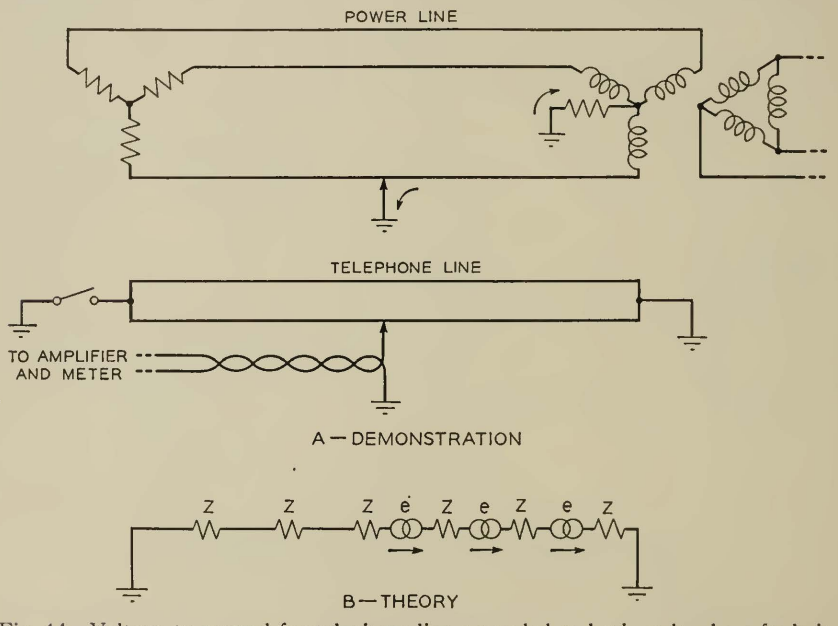


Fig. 14—Voltage to ground for telephone line grounded at both ends where fault is inside of exposure.

line times the impedance outside the exposure. This situation can be illustrated in the set-up of Fig. 13 by leaving the power fault at the end of the exposure and adding a small impedance in the ground connection to the telephone line at one end. When this is done it will be noted that a voltage-to-ground exists at the end where the impedance is added and that the voltage-to-ground decreases as the measuring point is moved toward the other end.

To illustrate that voltage-to-ground may occur under other conditions even though the telephone line is solidly grounded at the exposure terminals, consider the situation shown in Fig. 14. This

represents the same exposure conditions as in the preceding set-up and the telephone circuit is solidly grounded at both ends, but the fault current instead of flowing through the entire exposure flows through only half of it. In this case the telephone circuit impedances are the same as in the preceding case, but only half of the induced voltage is present. Consequently, the amount of current through the longitudinal circuit is only half of that in the preceding case. Now if the net voltage drop from either end to the middle is taken, it will be found that it is equal to one-half of the total longitudinal voltage induced under the conditions shown. Figure 14 also shows the set-up for demonstrating this condition. In this case the longitudinal voltage is smaller than in the preceding demonstration, but if both ends are grounded and the voltage measuring device is moved along the line, the voltage-to-ground increases from one end to the middle and then falls off from the middle to the opposite end.

In the last two demonstrations, the fault current on the power line was fed from one end only, i.e., "single-end feed." It sometimes happens that the fault current may be supplied to a power line, at least during the initial stage of a disturbance, from both ends, i.e., "double-end feed." The double-end feed condition tends to reduce the overall longitudinal induction when the fault occurs inside the exposure. In the demonstration shown in Fig. 15 it may be observed that for the set-up with a fault at the middle of the exposure the symmetry is so good that the total longitudinal voltage is very small. However, with the telephone circuit grounded at both ends, a substantial voltage-to-ground exists at a point in the telephone circuit opposite the fault and this voltage-to-ground reduces to zero at the ends. As the fault is moved toward either end of the exposure, there is a tendency for the longitudinal voltage to increase and for the voltage-to-ground, with both ends of the line grounded, to decrease until the limiting condition brought out in Fig. 13 is reached.

The analysis of voltage-to-ground can be carried out for any combination of impedances and induced voltage distributions by totaling vectorially the voltage drops (including any voltage drop over protector ground resistance) and the induced voltages between a grounded point and the point at which the voltage-to-ground is desired. The same analysis can also be carried out regardless of whether one or numerous wires are involved as long as all of the wires are grounded directly or through arrestors at the same points. If some of the wires on a line are not grounded, i.e., the protectors are not operated, the analysis for these wires must be carried out on the basis of their admittance-to-ground as discussed previously. In such a case, the

longitudinal voltage to be employed would be that remaining after correction for the shielding effect due to wires on which protectors have operated.

As mentioned previously, mutual shielding of the telephone wires on a line may prevent the operation of some protectors, particularly on a large line. Consequently, the exact analysis of the distribution of voltage-to-ground of all of the wires on a large line becomes very complex. Moreover, as is often the case, if impedance conditions are not uniform, such as where circuits are not coterminous, the complete analysis of the voltage-to-ground becomes even more complicated. Under such conditions it will generally be found that the distribution of voltage-to-ground along the different circuits is different, and consequently, voltages exist between different wires due to these differences in the voltage-to-ground. Likewise, voltages may exist between wires on which the protectors have operated and wires on which the protectors have not operated.

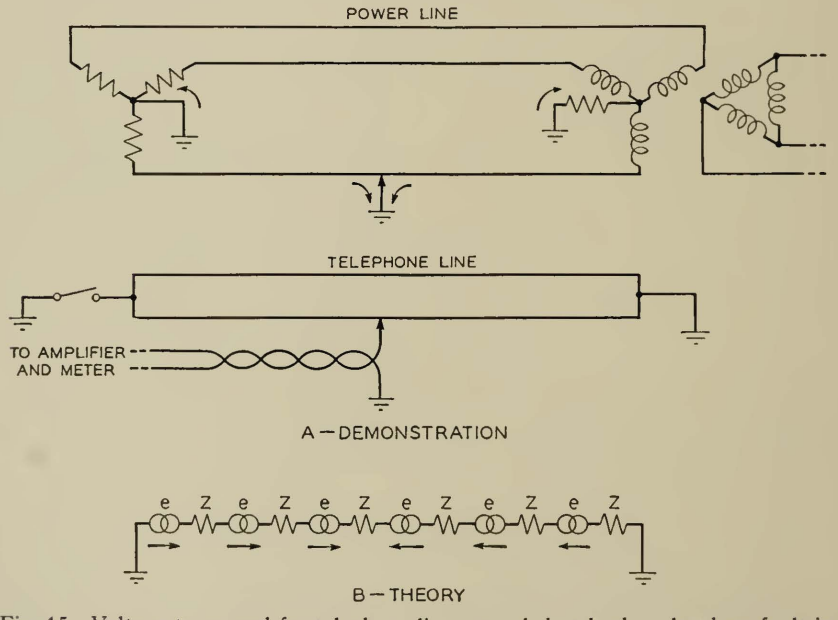


Fig. 15—Voltage to ground for telephone line grounded at both ends where fault is inside exposure—double-end feed.

All of the above analyses have been made on the assumptions of continuous telephone wires. If a wire is opened at one point the longitudinal voltage, reduced by shielding from any currents which exist in other continuous wires, will appear across the "open." On a large

line this shielding may, as pointed out previously, be so large that the voltage across the open is reduced to a fraction of the induced voltage.

Using the high voltage equipment which was used in connection with the demonstration of protector operation, "acoustic shock" can be demonstrated. Although strictly the term "acoustic shock" should be used only with reference to the effect on a person subjected to an abnormally loud sound, the term has also come to be used to designate a noise (usually transient) in a telephone receiver, the intensity of which is considerably higher than that of speech. It is produced by an excessive voltage across the terminals of the receiver.

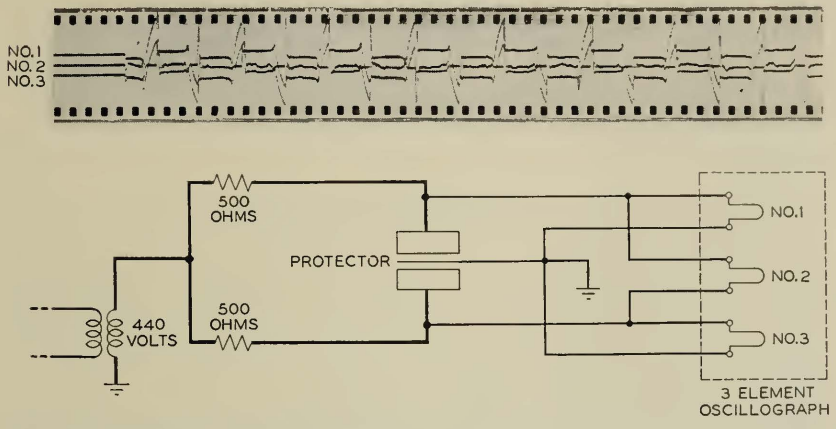


Fig. 16—Oscillograms showing voltages across protector blocks and resulting voltage across circuit, and schematic of test circuit.

The disturbances of this nature which are of primary interest in inductive coordination work are those which are liable to be experienced when a voltage high enough to cause the breakdown of protectors appears on a telephone circuit. This may be the result of low-frequency induction or may be produced by other causes, such as lightning or contacts between power and telephone circuits. Although induced voltages usually appear in equal magnitudes on the two sides of a circuit, the protector gaps on the two sides of the circuit discharge in an unsymmetrical manner with the result that a voltage higher than normal appears across the circuit. When this occurs a loud noise or rattle is produced in the receiver of a telephone set bridged across the circuit.

Figure 16 shows oscillograph traces of voltages measured across

operating protector blocks. Each outside trace shows the voltage across one of the two blocks. It will be noted that the two traces are not identical. The middle trace shows the resulting voltage across the circuit. It is this voltage which may cause acoustic shocks. The very jagged outline of this trace indicates that many frequencies other than 60 cycles are present.

The demonstration of Fig. 17 can be arranged to produce acoustic

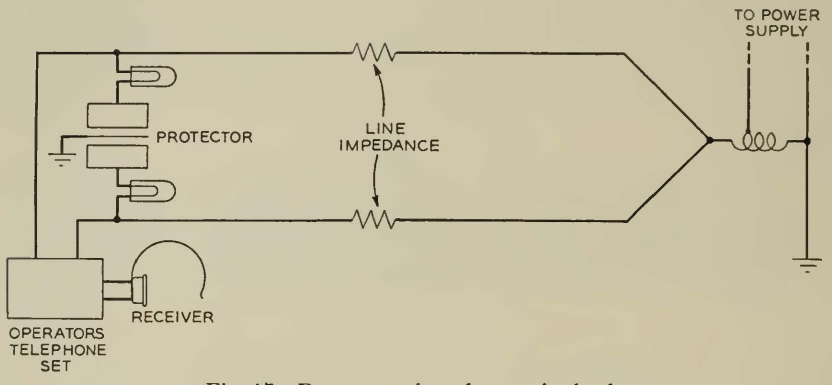


Fig. 17—Demonstration of acoustic shock.

shock. In this demonstration, sufficient voltage is impressed on the circuit to operate the protector blocks at one end. An operator's telephone set and a receiver are connected across the circuit at this end. When the voltage rises high enough to operate the blocks a relatively loud sound is emitted by the receiver.

Other effects may accompany the unsymmetrical discharge of the protector blocks. For example, the signals which are connected at the ends of the telephone circuits may operate and give what is commonly called a "false signal."

Probability Factors

In the preceding discussion, a number of factors were mentioned which may vary between different occurrences in the same inductive exposures. Among these may be mentioned the following:

(*a*) The impedances in the faulted circuit may vary between occurrences due to variations in the location of faults, variations in the effective fault resistance, etc. The effect of the variation in location of the fault, of course, is to change the line impedance in the faulted circuit and hence the fault current.

(*b*) The duration of the fault current may vary between occurrences due to variations in conditions which affect the speed of opera-

tion of the power circuit relays and circuit breakers. In some cases, of course, faults may clear themselves without circuit breaker operation and this introduces additional variations.

(*c*) Large variations in longitudinal voltage, voltage-to-ground, and current through protectors may occur with relatively small variations in the locations of faults when they occur inside inductive exposures.

(*d*) The shielding effects due to the operation of protectors on telephone circuits may vary considerably between different occurrences.

The variations in induced voltage duration, etc., between different occurrences are, of course, only part of the story. Obviously, the *total* number of faults which may occur on a power line in an exposure over a given period is equally important. This will be affected by numerous factors such as type of line, severity of lightning and other hazards, etc.

In addition, there are variations in the reactions on the telephone circuits. For example, the protector blocks used do not all break down at the same voltage and the fortuitous variations in the breakdown voltage may have an important bearing on the number of protectors which operate and consequently on the total shielding, current through protectors, etc. From the standpoint of possibilities of acoustic and electric shock, there are of course many other probability factors involved.

All of these factors are under investigation and our knowledge of them is increasing from day to day. It is probable, however, that low-frequency induction will always remain a subject in which quantitative analyses can tell only a part of the story.

Circulating Currents and Singing on Two-Wire Cable Circuits

By LEONARD GLADSTONE ABRAHAM

One of the important factors limiting the working net losses of two-wire cable circuits is the possibility of excessive circulating currents or actual singing.

A theory is developed for the computation of the distribution of singing margins on groups of two-wire circuits from the known gains and losses and known functions of the deviations of loading coils, loading coil spacing, cable capacitance and office equipment. The distribution functions of circulating current margins, of active return losses and of active singing points are also derived.

The possible application of these methods to specific problems is discussed and an example of the computations involved is given.

The theory herein involves certain approximations and empiricisms in determining the singing limitations but it is believed to give an answer which approaches the exact answer rather closely.

Introduction

AMONG the considerations which limit the minimum working net losses [1] of two-wire cable circuits, one of the most important is the desirability of avoiding excessive circulating currents. These circulating currents may manifest themselves as a quality impairment due directly to frequency and phase distortion or as sustained oscillation (singing).

In a given two-wire cable circuit, if the exact location and nature of each irregularity were known, it would be possible to compute exactly whether sufficient singing margin is available. The practicable method, however, is to compute the singing margins which will be exceeded on various percentages of a large group of such circuits, from the information which is available about the irregularities on a distributional basis. This paper first derives theoretical distributions of circulating current margins and singing margins without regard to various practical considerations such as the effect of repeating coils and other apparatus. In the second main division are discussed various considerations which are involved in applying the theory. In the third main division detailed computation methods are illustrated. The attached appendices cover the mathematical derivation of certain quantities.

[1] "Certain Factors Limiting the Volume Efficiency of Repeatered Telephone Circuits," L. G. Abraham, *Bell Sys. Tech. Jour.*, October, 1933.

Theoretical Distributions

Circulating Current Margins

A two-wire loaded cable circuit will be considered which consists of a number of repeater sections with 22-type repeaters. As shown on Fig. 1 there are various circulating paths in such a two-wire circuit

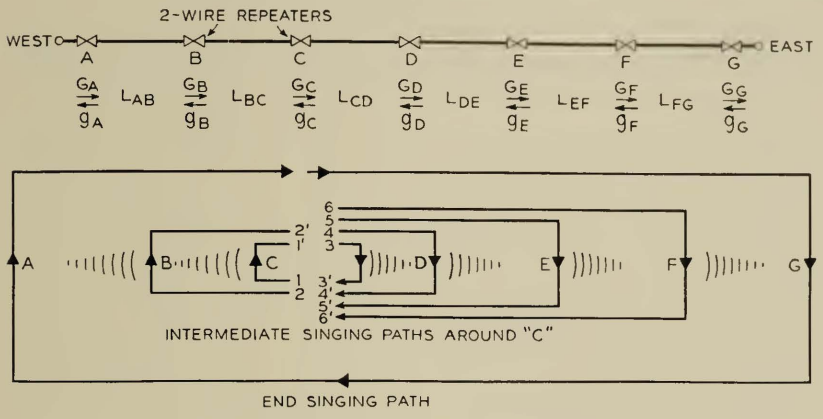


Fig. 1—Singing paths in a two-wire circuit.

which might cause objectionable circulating currents at any given frequency. Each repeater section consists of a large number of loading sections which have approximately the same capacitance and the same inductance per mile. Practically, however, there are deviations of the capacitance and inductance in a given loading section from the nominal average, each of which introduces an impedance irregularity which prevents the balancing network from exactly balancing the line. To determine the amount of current returned from each of these irregularities in a given case and the phase at which it is returned is, in general, impracticable. It is possible, however, to determine in what percentage of circuits the circulating current at any given frequency will exceed a certain percentage of the original current or, in other words, the percentage of cases in which the loss in the circulating current path will be less than a certain amount. The loss in the circulating current path at a given frequency is called the circulating current margin at that frequency.

The distribution of return losses at any given frequency in a cable section without repeater has been determined by G. Crisson in a paper entitled " Irregularities in Loaded Telephone Circuits," *Bell System Technical Journal*, October, 1925. This derives the distribution of the return losses which would be measured at a given frequency on a

large number of such cable circuits in terms of certain functions of the capacitance and inductance deviations. From this paper the distribution function of the return losses may be expressed in decibels as follows:

$$S = S_H + S_w - S_A + S_F = S_1 + S_F$$

(See Appendix VI for Nomenclature).

In this equation S_F is the distribution function of the return losses and S_1 is the return loss at the frequency in question which is exceeded by

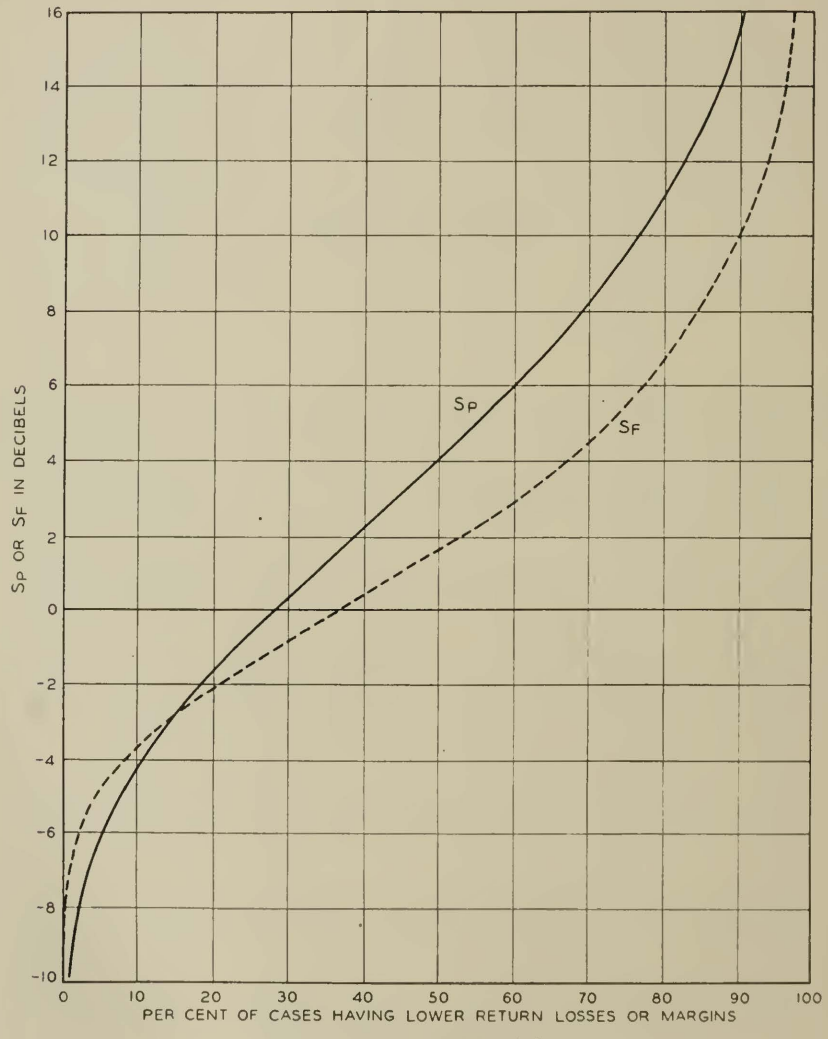


Fig. 2—Values of S_P and S_F.

63 per cent of the total return losses measured. S_H is determined entirely from the fractional deviations of the capacitances and inductances. S_w is determined by the proximity of the frequency in question to the cutoff frequency of the line facilities. S_A is determined by the attenuation of the cable circuit and is, in effect, the summation function of the different sources of irregularity.

Figure 2 shows the distribution curve S_F from the paper referred to above. In a similar manner the distribution curve of several repeater sections in tandem or in parallel [2] may be determined. This is known as an active return loss. Referring to Fig. 1 the return loss measured across the west hybrid coil of repeater C will be determined not only by the irregularities in the immediately adjacent repeater section but also by the irregularities in the other repeater section to the west of repeater C as seen through the intervening losses and gains of the circuit. Strictly speaking, the return losses on the east sides of repeaters A and B will also affect the active return loss because circulating paths around each repeater and around various combinations of repeaters will exist. However, these circulating paths have so much loss in any practical field circuit in which the other requirements are satisfied, that the return losses on the east side of A and B need not be considered.

Appendix I derives the formula for the distribution curve of an active return loss when the passive return losses of which it is made up are of the form given above, assuming no returned currents from beyond the terminal repeater. This distribution function is:

$$S_1 - F(N, T) + S_F.$$

In this case it is assumed that the value of S_1 is the same for the passive return losses of each repeater section, but from the appendix, the more general case where there is a different value of S_1 for each section may be determined.

Figure 3 gives values of $F(N, T)$ for the specific case where each repeater section has the same loss and each intermediate repeater has the same gain. The value of this function is also derived in Appendix I.

Referring again to Fig. 1 there will be an active return loss which may be measured on the west side of repeater C and also an active return loss which may be measured on the east side of repeater C. Assuming that the distribution function of the active return loss toward the west is $S_{11} + S_F$ (not including path A) while the dis-

[2] While this paper develops the theory specifically for the case of repeater sections in tandem, e.g., in a single two-wire circuit, it is generally applicable also to the case of repeater sections in parallel, e.g., in a toll conference connection.

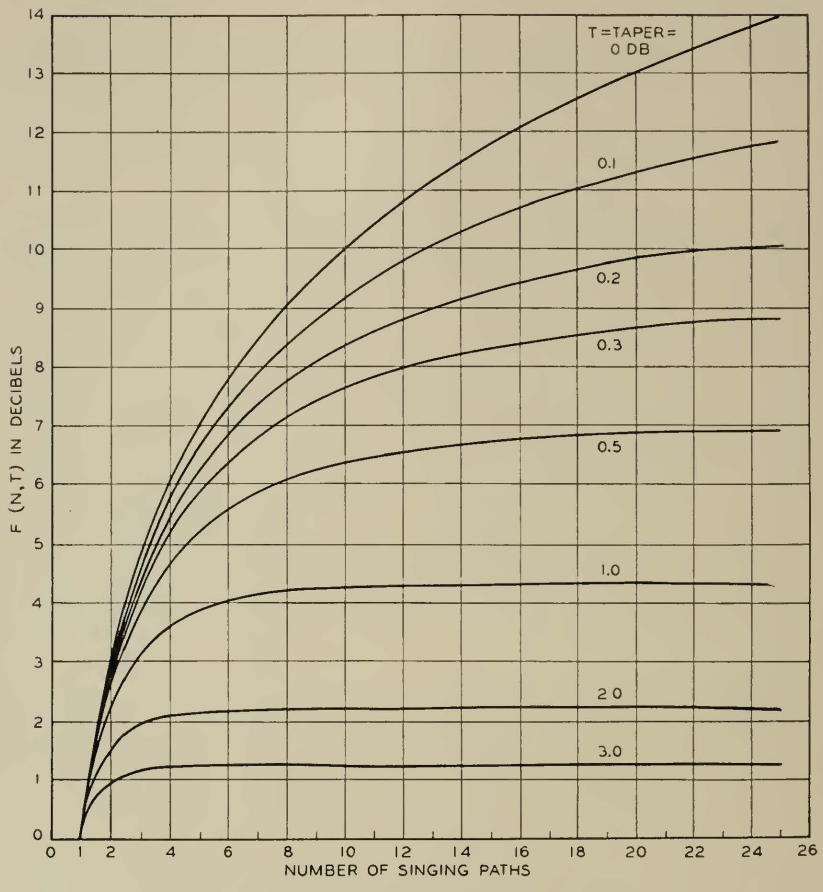


Fig. 3—$F(N, T)$ in perfectly regular circuit.

tribution function of the active return losses toward the east is $S_{12} + S_F$ (not including path G), the distribution function of the two return losses in series around the repeater will be $S_{11} + S_{12} + S_P$ as derived in Appendix II. In the general case, there will be gain in series with the active return losses, which may be considered as part of the return losses. For convenience, define S_{11} as follows:

$$S_{11} = S_1 - F(N_1, T) - G_C. \tag{1}$$

S_1 and $F(N_1, T)$ are as discussed above; N_1 is the number of repeater sections towards the west end of the circuit; and G_C is the gain of the repeater in the west to east direction, say. Similarly,

$$S_{12} = S_1 - F(N_2, T) - g_C. \tag{2}$$

The various quantities are like those enumerated above except that they are for the active return loss in the other direction and include the gain of the repeater in the east to west direction.

Figure 2 shows the value of S_P as derived in Appendix II plotted with S_F. By comparison of these curves, it may be seen that the spread of active return losses around a repeater will be somewhat larger than the spread of active return losses in one direction. The S_P curve approaches more nearly to a so-called normal law than the S_F curve does.

The previous discussion has been confined to the returned currents obtained from intermediate points in the circuit and also at a single pre-selected frequency. In addition to these currents, there is at each end of the circuit a current returned through the path called the terminal return loss. In a four-wire cable circuit this is the only path for circulating currents. From field data it appears that these terminal return losses approximately follow the S_F distribution curve also. The end path toward the west may be written as:

$$S_{21} + S_F = 7 + E_1 + S_F, \tag{3}$$

where 7 db is assumed as the terminal return loss for $S_F = 0$. In this case E_1 is the net loss at the particular frequency from the repeater in question (from the west side of repeater C) to the far end of the circuit and from that end of the circuit back to the repeater minus the gain of the receiving repeater on the latter side of the circuit (the W–E repeater). The end path toward the east is

$$S_{22} + S_F = 7 + E_2 + S_F. \tag{4}$$

If the losses of each cable section are equal and the gains of each intermediate repeater are equal, the repeater in question is at the center of the circuit, and $E_1 = E_2$, and at 1000 cycles E_1 would be equal to the nominal circuit net loss.

The active return loss at a given frequency toward the west which includes the end path (both including the west to east repeater gain) will be

$$S_{31} + S_F = (S_{11} \underset{p}{\mathrm{X}} S_{21}) + S_F \tag{5}$$

and toward the east will be

$$S_{32} + S_F = (S_{12} \underset{p}{\mathrm{X}} S_{22}) + S_F. \tag{6}$$

Where $\underset{p}{\mathrm{X}}$ means that the quantities so connected are combined as

if their powers added directly; e.g., $S_{11} \underset{p}{\mathrm{X}} S_{21} = S_{31}$ means that

$$10^{-S_{11}/10} + 10^{-S_{21}/10} = 10^{-S_{31}/10}.$$

The circulating current margin (M_c) around this repeater at a given frequency will therefore be $S_{31} + S_{32} + S_p$, as derived in Appendix II. Written in the more general form, the circulating current margin is

$$M_c = [S_1 - F(N_1, T) - G_C] \underset{p}{\mathrm{X}} [E_1 + 7] + [S_1 - F(N_2, T) - g_C] \underset{p}{\mathrm{X}} [E_2 + 7] + S_P. \quad (7)$$

Singing Margins

The objectionable effects of too low circulating current margin are to cause poor quality in transmission over the circuit and to cause the circuit to " ring " or sound " hollow " to the talker or listener. When a circuit oscillates or " sings," conversations over it become difficult or impossible, voice-operated devices on connecting circuits are locked up, parts common with other circuits such as common " C " batteries may be adversely affected, and other circuits are made noisy through crosstalk in the cable or in the repeater station. It is therefore important that the percentage of cases in which singing occurs shall be very small.

In general, the tendency will be for most of the loaded cable circuits to sing within a fairly limited frequency range which is usually near the upper frequency point at which the overall circuit loss begins to be appreciably greater than the loss at 1000 cycles. Figure 4 shows a

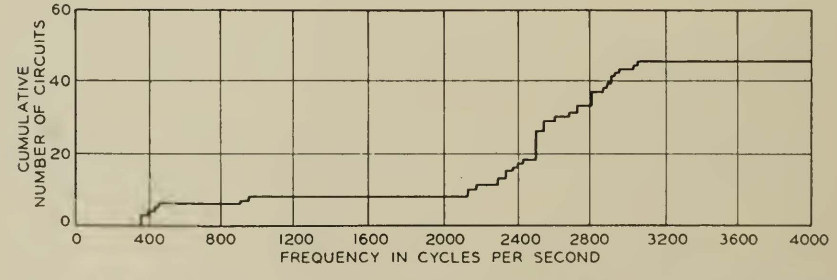


Fig. 4—Singing frequency in 22 tests on 46 19-gauge B & H–88–50 two-wire cable circuits.

cumulative plot of the singing frequencies during 22 tests [3] on 46 19-gauge B & H–88–50 two-wire cable circuits. It may be seen that

[3] A 22 test is a singing test made by increasing the gain of a normal working repeater in a two-wire circuit until singing begins.

over 80 per cent of these frequencies were between 2200 and 3100 cycles, while the nominal transmitted band on these facilities is from 250 to 3000 cycles.

Consider the case of the return loss of one circuit in a group of cable sections. A typical curve of the passive return loss of such a circuit is shown on Fig. 5. It will be seen that it consists of a set of "wabbles"

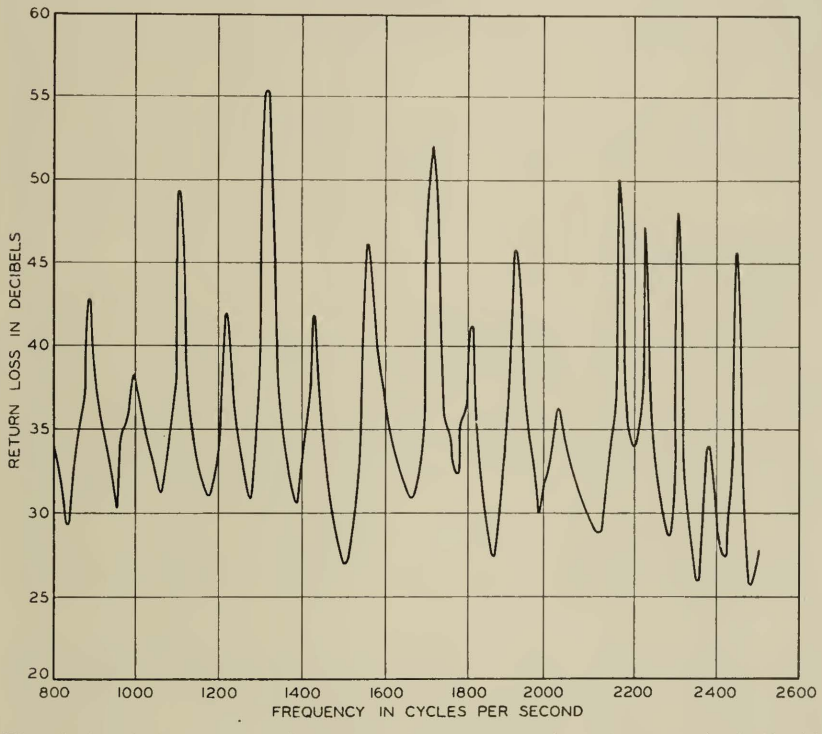


Fig. 5—Typical return-loss-frequency characteristic of a 19-gauge H–88–50 side circuit repeater section.

around a general trend line. When a measurement of return loss is made at a single pre-selected value, the return loss obtained is determined by what may be considered as two components, (1) the general quality of the repeater section as determined by the trend line, and (2) the particular part of the " wabble " on which the particular frequency measurement happens to fall (i.e., the bottom or top, or in between point of a given " wabble "). In measuring the return losses of a large group of lines at, say, 2900 cycles, the lower values of return loss will tend to be those which happened to be measured at the bottom of a " wabble." The higher the return loss of a given

circuit, the more chance there is that the return loss of that line happened to be measured at the top of a " wabble."

When singing points [4] are measured, however, it may be said that when singing takes place above, say, 2000 cycles, very nearly the minimum value of return loss minus gain is obtained. The distribution of singing points on a large group of lines, therefore, will be lower (in decibels after correction for gain characteristic) for a given percentage than the distribution of single-frequency return losses even at the frequency which has the lowest computed values of any within the transmitted band (say, 2900 cycles for a repeater cutting off about there). Following the reasoning concerning the " wabble " given above, the amount by which the singing point will be lower will be small for the lower return losses and singing points of a given group of lines, and will tend to increase more and more as the value of return loss becomes higher.

Figure 6 shows the S_F curve plotted on a scale which makes all cumulative normal law distributions come out as straight lines. It may be seen that the differences just described will tend to make the singing point distribution more nearly a normal law than the S_F curve. Field measurements show that such singing point measurements do approximate a normal law much more closely than they approximate the S_F curve. Measurements of singing points on about 900 19-gauge H–172–63 side circuits at 18 places during completion tests gave standard deviations about the average of the group at each place which, when added together as the weighted root mean square, gave a general standard deviation of 2.02 db. About 400 similar measurements on 16-gauge H–44–25 side circuits at 16 points gave a standard deviation of 2.05 db. Similar measurements on 233 19-gauge H–88–50 side circuits and 77 19-gauge H–88–50 phantom circuits gave standard deviations of 2.13 and 2.03 db, respectively. It therefore seems reasonable to conclude that a standard deviation of about 2 db is substantially correct for the distribution curve of singing points at a given place. It should be realized that if singing points for a given type of facility from a large number of places are grouped together and a standard deviation of the entire group obtained around the average of the entire group, it may be considerably larger than 2 db due to the differences in average values of the different groups. Such computations on about 7500 measurements on one type of facilities showed a standard deviation of about 3 db for the entire group.

[4] The singing point of a given line is the gain which must be connected between the two sides of a hybrid coil to just cause singing, when the line terminals of the hybrid coil are connected to the line in question and the network terminals are connected to the normal balancing network circuit. Unless otherwise specified, the singing point is expressed in terms of the 1000-cycle gain.

The other important point in determining the singing point distribution is to fix the amount by which its average is less than say, the reference 63 per cent point on the S_F curve; i.e., if the distribution

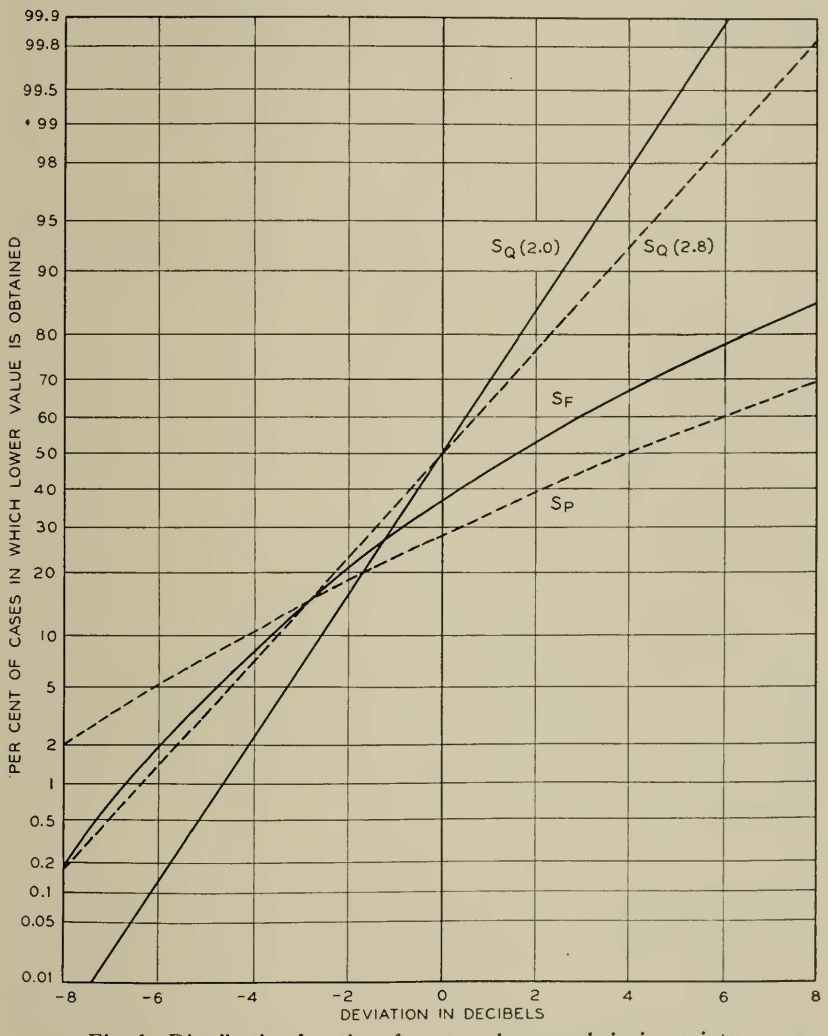


Fig. 6—Distribution functions for return losses and singing points.

curve of single frequency return losses for, say, the highest frequency having as much gain in the repeater as the gain at 1000 cycles, is $S_1 + S_F$, the distribution of singing points will be $S_1 - B + S_Q$ (2.0) where B is S_1 minus the average of the singing points, S_Q (2.0) is a normal law with a standard deviation of 2 db, and Q is the percentage

of cases which will have lower singing point values. It was not possible to compute B from the measurements referred to above due to incomplete information.

From field measurements including a large group on 19-gauge B & H–88–50 facilities during a Newark-Philadelphia trial, however, it appears that B is about 2.5 db. As further evidence, the reasoning given above concerning the " wabble " in the return loss distribution when considered together with the standard deviation of 2 db leads to the conclusion that the value of B must lie within a range from about 2 to say, 3 or 4 db for all facilities.

It is accordingly considered reasonable to assume that the distribution of singing points on passive repeater sections at a given place will be a normal law of the form $S_1 - 2.5 + S_Q(2.0)$.[5] The S_Q (2.0) law is shown on Fig. 6.

When the singing points of several such repeater sections in tandem are measured, it seems reasonable to say that the distribution of singing points will be a normal law of the form $S_1 - F(N, T) - 2.5 + S_Q(2.0)$, by analogy with the distribution functions of single-frequency return losses.

When an end path is measured by itself, the singing point in the 2000–3000-cycle range will generally be a little less than would be indicated by single-frequency return-loss measurements. However, the difference between the two sets of measurements will probably be less on the average than the 2.5 db indicated above, since the " wabbles " in measurements of terminal return losses are usually much less than in measurements of repeater section return loss, particularly for the lower terminal return losses which generally are on non-loaded loops. It is accordingly estimated that the distribution of singing points of the end paths is

$$S_{21} - 1 + S_Q(2.0) = 6 + E_1 + S_Q(2.0). \tag{8}$$

When an end path through a terminal return loss is added to the intermediate paths, tests have shown that the singing point of the resultant will in each case be approximately as if the currents of the singing points of the intermediate paths and the end path added directly. A group of such measurements is shown in Appendix III. The singing point of the resultant will therefore have a distribution of

[5] It is interesting to note here that in an article in *Electrical Communication* for July, 1934, entitled "The Prediction of Probable Singing Points on Loaded Cable Circuits," the law is given as effectively $S_1 - 2.5 + S_Q(1.75)$. The difference between these equations is generally within the limits of experimental error. It may be a real difference, however, since there are certain differences in adjustments of building-out condensers which might be expected to affect the standard deviation.

approximately

$$(S_{11} - 2.5) \underset{i}{\mathrm{X}} (S_{21} - 1) + S_Q(2.0), \tag{9}$$

where $\underset{i}{\mathrm{X}}$ means that quantities so connected are combined as if their currents added directly, e.g., $S_{11} \underset{i}{\mathrm{X}} S_{21} = S_{31}$ means that

$$10^{-S_{11}/20} + 10^{-S_{21}/20} = 10^{-S_{31}/20}.$$

When two end paths are connected in series, say around a four-wire cable circuit, it is believed that singing points in the 2000–3000-cycle range will add directly, because (1) due to the large amount of phase shift around the loop, there will be a large number of frequencies which will have the proper phase shift to permit singing and, (2) the return loss in the frequency range of interest on a given line will change slowly with frequency, which will cause the gain required to produce singing to change only slowly with frequency. On this assumption and assuming the normal law distribution given above, it follows directly from the mathematics of such functions that the distribution of the singing points of two such end paths in series is

$$12 + E_1 + E_2 + S_Q(2.8) = S_{21} - 1 + S_{22} - 1 + S_Q(2.8). \tag{10}$$

On the other hand, if only intermediate paths were present on the two sides of a two-wire repeater, the singing points in the two directions would not, in general, add directly, because they would generally be at different frequencies and when singing occurs at one of these frequencies (or at some new frequency) the sum of the return losses will generally be greater than would be indicated by the sum of the singing points. Tests show that the internal singing margin (the 22-test value corrected for frequency characteristic of the repeater, with only intermediate paths) is about 2.5 db higher on the average than would be indicated by the sum of singing points in the two directions. Similar tests show that when the end path is added so that it is very important compared with the intermediate paths, this average difference becomes about zero.

Also, the average one-way singing point on intermediate paths is about $B = 2.5$ db below the reference single-frequency return loss, as outlined above. Considering the internal singing margin as effectively a 21-test [6] measurement through two return losses in series,

[6] A 21-test is a singing test made by increasing the gain of a repeater until singing occurs, with a line and network connected to the line and network terminals, respectively, of one hybrid coil and a fixed known return loss connected to the other hybrid coil.

it seems reasonable to expect this singing point to be about 2.5 db on the average below the reference single-frequency return loss of the two in series.

From either or both of the above two paragraphs, it may be assumed that the internal singing margin is about

$$M_I = S_1 - F(N_1, T) - G_C + S_1 - F(N_2, T) - g_C - 2.5 + S_Q(2.8). \quad (11)$$

The distribution function $S_Q(2.8)$ is an assumption which seems reasonable from general considerations.

From the above discussion the singing margin of the entire circuit will be a function of all the intermediate and end paths which meets the following conditions:

1. When the end paths are unimportant compared to the intermediate paths, the singing margin is

$$M_I = S_{11} + S_{12} - 2.5 + S_Q(2.8). \quad (12)$$

2. When the intermediate paths are unimportant compared to the end paths, the singing margin is

$$M_E = S_{21} - 1 + S_{22} - 1 + S_Q(2.8). \quad (13)$$

3. When both kinds of paths are fairly important, the singing margin is some compromise between

$$(S_{11}) \underset{p}{\mathrm{X}} (S_{21}) + (S_{11}) \underset{p}{\mathrm{X}} (S_{21}) + S_Q(2.8),$$

which would have an average value equal to the reference value of the circulating current margin and about 2.5 db less than this. The singing margin is obviously greater than

$$(S_{11} - 2.5) \underset{i}{\mathrm{X}} (S_{21} - 1) + (S_{12} - 2.5) \underset{i}{\mathrm{X}} (S_{22} - 1) + S_Q(2.8)$$

because the active singing point in one direction is almost certain to be at a different frequency from that of the active singing point in the other direction.

A reasonable empirical compromise which satisfies all of these conditions and preserves the symmetry of the equation is to say that the total singing margin of the circuit is

$$M_s = (S_{11} - 1.25) \underset{p}{\mathrm{X}} (S_{21} - 1) + (S_{12} - 1.25) \underset{p}{\mathrm{X}} (S_{22} - 1) + S_Q(2.8). \quad (14)$$

This may be written in more general fashion as

$$M_s = (S_1 - F(N_1, T) - G_C - 1.25) \underset{p}{\mathrm{X}} (6 + E_1) + (S_1 - F(N_2, T) - g_C - 1.25) \underset{p}{\mathrm{X}} (6 + E_2) + S_Q(2.8) \quad (15)$$

or still more generally,

$$M_s = (S_H + S_w - S_A - L + T + q - F(N_1, T + q) - 1.25) \underset{p}{\mathrm{X}} (6 + E_1' + 2N_1q) + (S_H + S_w - S_A - L + T + q - F(N_2, T + q) - 1.25) \underset{p}{\mathrm{X}} (6 + E_2' + 2N_2q) + S_Q(2.8). \quad (16)$$

Criterion of Satisfactory Performance

The above discussion derives methods of determining the circulating current margin and the singing margin that will be obtained (on a distributional basis) for a cable circuit under a given set of conditions. As a practical matter, the overall net loss of the circuit will sometimes vary, regulating repeaters will change in gain setting, temporary troubles will occur and, in some instances, toll circuit terminations will be removed while connections are being set up. Each of these factors will reduce the singing margin that the circuit had under average conditions. In the Bell System toll circuits are usually designed so that a 10 db singing margin or more is obtained in 90 per cent of the loaded two-wire cable circuits under average conditions. This, of course, is the same thing as saying that 12 db singing margin should be exceeded under average conditions in 71 per cent of the cases, or 8 db in 97.7 per cent of the cases (see $S_Q(2.8)$ curve on Fig. 6). The following table shows the percentages of circuits which will have various lower singing margins under average conditions than the indicated values, for various different assumptions as to the design requirement (i.e., the singing margin which must be exceeded in 90 per cent of the cases).

Design Requirement in Db	Singing Margin in Db =	Percentages Having Lower Singing Margins					
		14	12	10	8	6	4
12		29	10	2.3	0.23	0.03	0.002
10		57	29	10	2.3	0.33	0.03
8		84	57	29	10	2.3	0.33
6		95	84	57	29	10	2.3

The following table shows the same information for circulating current margins, computed on the assumption that $S_{11} - 1.25 = S_{12} - 1.25 = (S_{21} - 1) = (S_{22} - 1)$. The design requirements in this case are also that 12 db, 10 db, 8 db or 6 db singing margin (not circulating current margin) shall be exceeded in 90 per cent of the cases under average conditions.

Design Requirement in Db	Circulating Current Margin in Db =	Percentages Having Lower Circulating Current Margins					
		14	12	10	8	6	4
12		11.3	5.55	2.25	0.78	—	—
10		19.4	11.3	5.55	2.25	0.78	—
8		29.4	19.4	11.3	5.55	2.25	0.78
6		40.2	29.4	19.4	11.3	5.55	2.25

It is estimated that the reduction in circulating current margin due to regulation, variation in circuit net loss, and temporary troubles will seldom exceed $4\frac{1}{2}$ db by very much but about this amount of reduction will occur fairly frequently. Similarly, the reduction in singing margin due to these causes and also to the removal of a termination while a circuit is being set up will seldom exceed $5\frac{1}{2}$ db. It is believed that about 4 db circulating current margin at the critical frequency is as small as can be obtained without very objectionable quality distortion.

When $5\frac{1}{2}$ db reduction in singing margin is obtained on the average, we may read from the first table above at $5\frac{1}{2}$ db margin, the percentage of cases in which actual singing will take place. For example, with a design requirement of 8 db, interpolating between the vertical columns headed 6 and 4 db, respectively, we find that about $1\frac{1}{2}$ per cent of cases will result in actual singing. Similarly, by reading at $4 + 4\frac{1}{2} = 8\frac{1}{2}$ db circulating current margin in the second table, we find the percentage of cases in which the quality will be objectionably distorted during the period in which a $4\frac{1}{2}$ db reduction in circulating current margin is obtained. The following table shows these two different percentages for the design limits given above. From this table it

Design Limit in Db	Percentage of Cases With Less Than 4 Db Circulating Current Margin Under Extreme Conditions	Percentage of Cases Which Will Sing Under Extreme Conditions
12	0.97	0.015
10	2.8	0.18
8	6.8	1.45
6	13.1	7.1

seems reasonable to require a 10 db singing margin under average

conditions in order to keep the percentage of circuits which will have objectionable circulating currents and the percentage which will sing occasionally low enough that serious service reactions will not be obtained.

Practical Considerations

Equipment and Terminating Effects

When actual cable sections are investigated, the effect of the equipment must be considered. The effect of adding, say, repeating coils in the line and network sides of the repeater, is to introduce some reflected current and to add a certain loss in the line which must be made up by increasing the repeater gain. The cable return loss is increased by twice the loss introduced, so that the only effect on circulating current paths is due to the differences in the line and network equipment.

Appendix IV derives an expression for the distribution curve of an active or passive single-frequency return loss of a line with equipment, on certain assumptions. The complication of using this method to determine circulating current margins or singing margins is large, however, and an approximate method is illustrated below.

Figure 7 shows an example of the combination of the equipment-return loss with the return loss of a cable section without equipment. It is assumed in this case that the equipment-return loss is 13 db higher than the value of cable-return loss for which $S_F = 0$. For example, the equipment return loss might be 40 db while the line return loss was $27 + S_F$ db. A generally similar curve would be obtained from the combination of the repeater section terminating effect with the cable return losses.

Since the principal interest in connection with these return losses is with the lower values, it seems reasonable to say that the effect may be approximated generally by shifting the S_F curve a certain part of a decibel, depending upon the difference between S_E and the cable-return losses. Strictly speaking, of course, it not only shifts the curve but also changes the standard deviation, but the complication of taking account of this does not seem worth while with present lines and equipment.

Since we are interested in the singing margin which is exceeded 90 per cent of the time under average conditions, it would be possible to select a value of cable return loss from the distribution curve which, if it occurred in each repeater section of a particular circuit, would give the indicated singing margin when the various intermediate singing paths are combined together as the sum of their power ratios.

This would be roughly where $S_F = -2$ db. The percentage of cable section return losses which would exceed this value, however, would be about 80 per cent, so a line which gave such a low singing margin would, in general, tend to be made up of a few very low-return losses

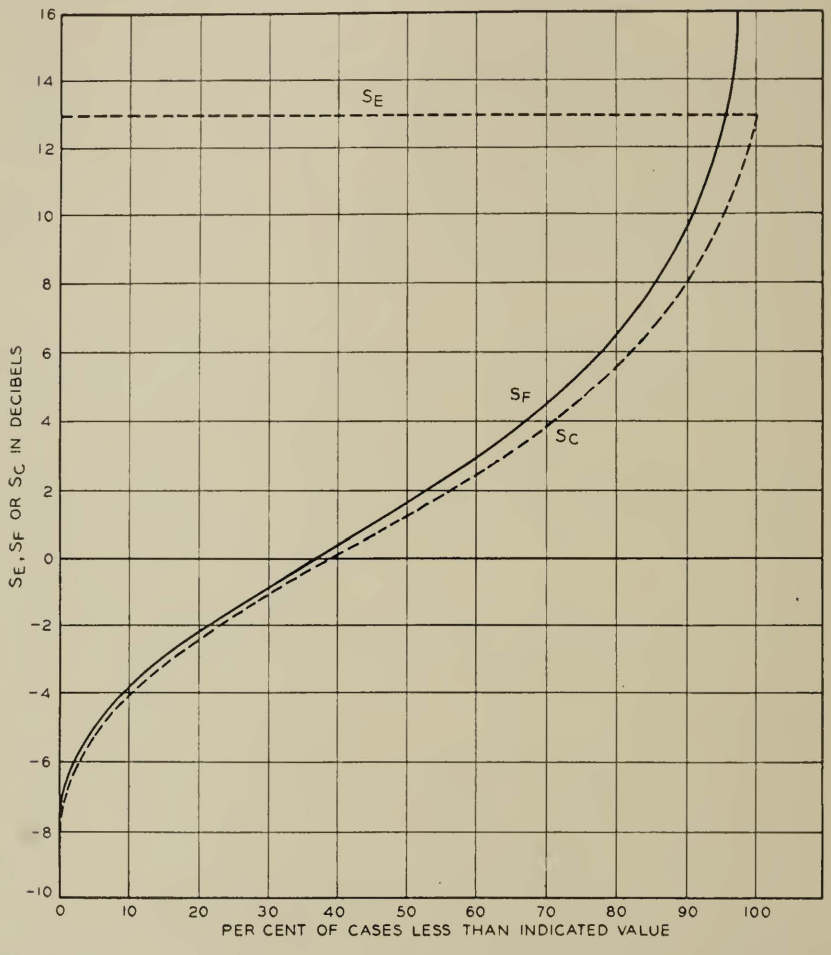


Fig. 7—Return loss distributions of cable pairs (S_F), equipment (S_E), and the combination (S_C).

and a considerable percentage of higher-return losses. It, therefore, seems that this value of cable-return loss, if combined with the equipment and terminating return loss, would not be sufficiently degraded to consider that degradation as a fair average degradation for the entire curve. On the other hand, if the cable return loss which is

exceeded in 50 per cent of the cases is taken (i.e., $S_F = +1.8$ db), the degradation would evidently be too large due to the equipment since the lines in which the singing margin might be expected to be low, would generally be those which happen to have more than the usual number of low line return losses.

As an approximation in the usual case, it seems reasonable to combine the line equipment return losses with the cable return losses as the sum of their power ratios, with a value of S_F of about -1 db (i.e., where 28 per cent of the return losses are lower than the value considered). One db added to the resultant will give the value of the resultant for which $S_F = 0$. The same thing may also be accomplished by combining the cable return loss for $S_F = 0$ with the near and far-end equipment return losses each increased by 1 db. The example which is given later shows the method used.

Taper

The taper of a regular two-wire circuit is the amount of decrease in repeater output level at a given frequency which occurs between succeeding repeaters from the transmitting to the receiving end of the circuit. For example, starting at the transmitting repeater, if the repeater output levels at 1000 cycles were, respectively, 3.0, 2.5, 2.0, 1.5, etc., the 1000-cycle taper would be 0.5 db.

An optimum taper from the standpoint of singing will be reached when the singing margin around the most critical intermediate repeater is equal to the singing margin around the terminal repeater. This is certain to occur for some taper under the present method of distributing gain since an increase of the taper causes the gain of the receiving terminal repeater to be increased by a multiple of the increase of taper. However, the optimum taper from the standpoint of singing is not generally used in the field, because the optimum taper from the standpoints of crosstalk and noise is zero db. A compromise is generally made on two-wire cable circuits; e.g., 19-gauge B and H–88–50 facilities use 0.5 db and 16-gauge H–44–25 facilities use 0.2 db taper.

Active Balances

In some cases, it may be more convenient to measure active return losses or active singing points rather than singing margins. The expected distribution of active return losses and active singing points including the end paths when the end paths are the normal working terminations, have been discussed above.

In practice, however, the termination when such an active singing point test is made will normally be a fixed one, generally consisting of

resistances with a return loss substantially the same at all frequencies. Appendix V derives the distribution of active singing points for these cases.

For exact results, the modification of the return loss due to the near-end and far-end equipment should be added before combining the end path. In the practical case, however, only the lower values of return loss are generally significant, and the approximate method outlined above may generally be used without serious error.

Minimum Value of a Group of Singing Points

When a large number of independent measurements are taken, say of singing points on different pairs at a given repeater station, it is often of interest to know whether the maximum or minimum in that group represents some special condition, say, definite circuit trouble, or is merely a case where the various components happen to have added up or subtracted by pure chance.

In the case of singing points, the question of interest may usually be stated as follows: In " n " similar measurements at a given place, what is the probability that the smallest value will be more than x db below the true average measurement, assumed known?

If the fraction of cases which will be greater than a value y db below the average is P, the probability that " n " values will all be greater than this value is P^n. In other words $P_m(n, y) = P_m = P^n$ is the probability that no one in " n " values will be as much as y db below the true average.

For example, if the singing point distribution is $25 + S_Q(2)$, the chance that any given value selected at random will be as low as $25 - 4.6 = 20.4$ db is one in a hundred. However, the chance that the lowest one of 20 measurements will be that low is $1 - (0.99)^{20} = 0.1819$ or almost one in five. If we substitute $V_m = 1 - P_m$ and $V = 1 - P$, we may write that $1 - V_m = (1 - V)^n$.

$$V_m = 1 - (1 - V)^n = 1 - 1 + nV - \frac{n(n-1)}{2!} V^2 + \cdots$$

$$= nV - \frac{n(n-1)V^2}{2!} + \frac{n(n-1)(n-2)}{3!} V^3 - \cdots.$$

And for small values of V, $V_m = nV$. In the above case, therefore, if we wanted to compute a value of singing point which would be lower than all the 20 measurements nine times out of ten, we may compute $V = (0.1/20) = 0.005$, and read on the normal distribution $S_Q(2)$ that this corresponds to about 5 db below the average; i.e., to $25 - 5 = 20$ db. The more exact formula would be that $0.9 = P^{20}$

or $20 \log P = \log 0.9$ or $\log P = -0.002288$ or $P = 0.99476$ or $V = 0.00524$. Reading on the normal law distribution for this value changes the 20 db limit as computed above to about 20.03 db, the difference being negligible compared to the usual accuracy of such measurements.

In general, measurements outside of the limits so determined are probably cases of trouble. However, in the case of active singing points, certain systematic changes in the measured values from the computed values must be allowed for before trouble is proved.

Example

Following is an example of the application of these various methods.

Given: The following losses at the critical frequency for various successive repeater sections shown in Fig. 1.

Repeater Section	Loss in Db
A to *B*	17
B to *C*	17
C to *D*	18
D to *E*	15
E to *F*	15
F to *G*	16
Total	98

The following gains and return losses at the critical frequency are assumed. The gains give a 9 db overall net loss at the critical frequency.

At Point	Repeater Gain at the Critical Frequency—db			Return Losses at the Critical Frequency for $S_F = 0$, i.e., Values Exceeded in 63 Per Cent of Cases—db							
				Bare Line **		Near-End Apparatus *		Far-End Apparatus †		Combination ‡	
	A to *G*	*G* to *A*	Sum	Toward *G*	Toward *A*	Toward *G*	Toward *A*	Toward *G*	Toward *A*	Toward *G*	Toward *A*
A	2.4	12.0	14.4	27.6	—	41	—	46	—	27.35	—
B	15.0	15.0	30.0	27.6	27.6	41	41	46	46	27.35	27.35
C	16.8	16.9	33.7	27.6	27.6	41	41	48	46	27.37	27.35
D	16.9	13.9	30.8	27.6	27.6	41	41	42	48	27.25	27.37
E	13.9	13.9	27.8	27.6	27.6	41	41	42	42	27.25	27.25
F	13.9	14.9	28.8	27.6	27.6	41	41	44	42	27.32	27.25
G	10.1	2.4	12.5	—	27.6	—	41	—	44	—	27.32
Total	89.0	89.0	178.0								

* This is 40 db nominal apparatus return loss plus 1 db allowance as described above.

† This is 11 db far-end equipment terminating return loss plus 1 db allowance as described above plus twice the loss from the measuring point to the far-end equipment.

‡ These are the combinations of the three preceding return losses as the sums of their power ratios.

** Twice the loss of the near-end equipment is included.

Considering C as the critical repeater:

Path

Toward A:

1	27.35
2	27.35 + 34 − 30 = 31.35
Total	25.90 Items 1 and 2 combined as the sum of their power ratios.

Toward G:

1	27.37
2	27.25 + 36 − 30.8 = 32.45
3	27.25 + 38 − 31.8 + 30 − 27.8 = 35.65
4	27.32 + 38 − 31.8 + 30 − 27.8 + 30 − 28.8 = 36.92
Total	25.42. Items 1, 2, 3 and 4 combined as the sum of their power ratios.

The internal circulating current margin then is 25.90 + 25.42 − 33.7 = 17.62 db, or more in 72 per cent of the cases and the internal singing margin is 25.90 + 25.42 − 2.5 − 33.7 = 15.12 db or more in 50 per cent of the cases, both under the average conditions specified.

Now 10 db internal circulating current margin would be exceeded under average conditions in the percentage of cases for which S_P is 10 − 17.62 = − 7.62 db or greater; i.e., in about 97.5 per cent of the cases, reading from the S_p curve on Fig. 6. Or, similarly, 10 db internal singing margin will be exceeded in about 97 per cent of the cases, reading at $S_Q(2.8) = 10 - 15.12 = -5.12$ from the $S_Q(2.8)$ curve on Fig. 6.

The single-frequency loss in the end path toward A, before the gain of repeater C is introduced is, for a net loss at the critical frequency of 9 db, $34 + 34 - 30.0 - 14.4 + 7 + S_F = 30.6 + S_F$. In the other direction the corresponding function is $36.1 + S_F$. When the net loss of the circuit is expressed as E rather than 9 db, these functions become respectively $21.6 + E + S_F$ and $27.1 + E + S_F$. The corresponding singing point distributions are $20.6 + E + S_Q(2)$ and $26.1 + E + S_Q(2)$, respectively.

The desired value of 10 db or more singing margin obtained in 90 per cent of the cases under average conditions will be obtained when the following equation is satisfied:

$$10 = (25.90 - 1.25) \underset{p}{\mathrm{X}} (20.6 + E) + (25.42 - 1.25) \underset{p}{\mathrm{X}} (26.1 + E) - 33.7 + S_Q(2.8), \quad (17)$$

where $Q = 10$ per cent, $S_Q(2.8) = -3.5$ db. Substituting in the

equation and simplifying,

$$-1.62 = 0 \underset{p}{\times} (E - 4.05) + 0 \underset{p}{\times} (E + 1.93), \qquad (18)$$

which is satisfied if $E = 8.8$ db.

The active balance of the circuit with a 9 db net loss at the critical frequency measured from the " A " end may be computed as follows: The successive paths from the line side of repeater " A " are (using values for $S_F = 0$):

1. 27.35
2. $27.35 + 34 - 30 = 31.35$
3. $27.37 + 34 - 30 + 34 - 33.7 = 31.67$
4. $27.25 + 34 - 30 + 34 - 33.7 + 36 - 30.8 = 36.75$
5. $27.25 + 34 - 30 + 34 - 33.7 + 36 - 30.8 + 30 - 27.8 = 38.95$
6. $27.32 + 34 - 30 + 34 - 33.7 + 36 - 30.8 + 30 - 27.8 + 30 - 28.8 = 40.22.$

Combining these paths as the sum of their power ratios gives 24.33 db, which is the computed single-frequency return loss without the end path from the line side of the terminal repeater for $S_F = 0$ (i.e., which will be exceeded in 63 per cent of the cases). The active singing point from this point will be $24.33 - 2.5 + S_Q(2) = 21.83 + S_Q(2)$. The corresponding active singing point referred to the drop side of the repeater will be $21.83 - 14.4 + S_Q(2) = 7.43 + S_Q(2)$ db. With a fixed termination giving a return loss of 5 db at the circuit terminal, the end path from the drop side of the repeater will be $18 + 5 = 23$ db. From Appendix V, for $d = 23 - 7.43 = 15.57$ db, the expected active singing point including the end path is the distribution curve $7.43 + S_9(15.57)$ under average conditions. For example, by interpolation in the table in Appendix V, two per cent of such circuits will have lower active singing points from the drop than about $7.43 - 5 = 2.43$ db and from the line side of the repeater than $2.43 + 14.4 = 16.83$ db, both under average conditions.

This answer may be computed by reading from the $S_Q(2.0)$ at two per cent which gives $S_2(2.0) = -4.1$ db. At this percentage, therefore, the active singing point without the end path is $7.43 - 4.1 = 3.33$ db. Combining this with the fixed and path of 23 db gives $3.33 - 0.9 = 2.43$ db.

APPENDIX I

Active Return Loss Of Intermediate Paths

Let L_k = the loss from the west side of repeater C to the output of repeater X which is the k th repeater to the west, plus the loss from the input of the W–E side of repeater X back to the input of repeater C (Fig. 1).

$$\text{Let } L_k = 20 \log_{10} t_k^{2k}. \tag{19}$$

Now from Equations (26) and (27) from " Irregularities in Loaded Telephone Circuits ": [7]

$$I_{1k} = \frac{I_0 R_L}{t_k^{2k}} \sqrt{\sum_{j=1}^{m} A_k^{4(j-1)} \cos^2 \theta_{jk}}, \tag{20}$$

$$I_{2k} = \frac{I_0 R_L}{t_k^{2k}} \sqrt{\sum_{j=1}^{m} A_k^{4(j-1)} \sin^2 \theta_{jk}}, \tag{21}$$

where I_{1k} is I' (from the paper) for the currents from the $(k+1)$ th repeater section from C (counting the adjacent repeater section as the first) and I_2k is I'' for the same section.

The total components will be

$$I_1 = I_0 R_L \sqrt{\sum_{j=1}^{m_k} \sum_{k=0}^{N-1} \frac{A_k^{4(j-1)}}{t_k^{4k}} \cos^2 \theta_{jk}}, \tag{22}$$

$$I_2 = I_0 R_L \sqrt{\sum_{j=1}^{m_k} \sum_{k=0}^{N-1} \frac{A_k^{4(j-1)}}{t_k^{4k}} \sin^2 \theta_{jk}}, \tag{23}$$

where N = the total number of repeater sections so considered.

Assuming (as in the paper) that $I_1 = I_2$

$$I_1 = I_2 = \frac{I_0 R_L}{\sqrt{2}} \sqrt{\sum_{k=0}^{N-1} \frac{(1 - A_k^{4m_k})}{t_k^{4k}(1 - A_k^4)}}. \tag{24}$$

By analogy, Equation (34) from the paper referred to may be used for the distribution of the total current I_F, with $I_1 = I'$; i.e.,

$$F = e^{-(I_F^2/2I_1^2)}. \tag{25}$$

Equation (42) from the paper may therefore be written

$$S = S_H + S_W + S_F - S_A - F(N, T) \tag{26}$$

where

$$S_A + F(N, T) = 10 \log_{10} \left(\sum_{k=0}^{N-1} \frac{(1 - A_k^{4m})}{t_k^{4k}(1 - A_k^4)} \right). \tag{27}$$

[7] G. Crisson, October 1925, *Bell Sys. Tech. Jour.*

Where all values of A_k are the same, and equal to A,

$$S_A = 10 \log_{10} \left(\frac{1 - A^{4m}}{1 - A^4} \right) \tag{28}$$

$$F(N, T) = 10 \log_{10} \left(\sum_{k=0}^{N-1} t_k^{-4k} \right). \tag{29}$$

Where all values of " t_k " are the same and equal to " t,"

$$F(N, T) = 10 \log_{10} \left(\frac{1 - t^{-4N}}{1 - t^{-4}} \right). \tag{30}$$

This is the case where all the gains are equal and all the losses are equal at all points.

In the case where $t = 1$, i.e., $T = 0$ db,

$$F(N, O) = 10_{10} \log N. \tag{31}$$

APPENDIX II

Distribution of Two Active Return Losses in Series

Assume two return losses in series, the two being selected at random from groups of return losses (active or passive) having distributions represented by $S_{11} + S_F$ and $S_{12} + S_F$, respectively. What is the distribution curve of the two in series?

Let

$$p_1 = \frac{i_1}{{k_1}^2} e^{-i_1^2/2k_1^2}, \tag{32}$$

$$p_2 = \frac{i_2}{{k_2}^2} e^{-i_2^2/2k_2^2}, \tag{33}$$

where p_1 is the relative probability of obtaining a returned current i_1 from the first line $(S_{11} + S_F)$ and k_1 corresponds to I_1 in Appendix I. The quantities with subscript 2 are the same for the second line $(S_{12} + S_F)$. Now if $i_1 i_2 = i_3$, the probability that i_3 exceeds a certain value I is

$$P(i_3 > I) = \int_{i_2=0}^{\infty} \int_{i_1=I/i_2}^{\infty} p_1 di_1 p_2 di_2. \tag{34}$$

Substituting from Equations (32) and (33) and letting

$$x = \frac{{i_2}^2}{2{k_2}^2}, \qquad a = \frac{I}{2k_1 k_2}, \tag{35}$$

$$P(1_3 > I) = \int_0^{\infty} e^{-(x + (a^2/x))} dx, \tag{36}$$

$$P(1_3 > I) = 2aK_1(2a), \tag{37}$$

where K_1 is a Bessel's function of the second kind (see " Theory of Bessel Functions " by G. N. Watson).

Now

$$a = \frac{I}{2k_1k_2} = \text{anti-log}_{10}\left(\frac{S_{11} + S_{12} - S_I}{20}\right), \tag{38}$$

where S_I is the return loss obtained from the two in series.

Let

$$S_p = -20 \log_{10} a = -S_{11} - S_{12} + S_I, \tag{39}$$

$$S_I = S_{11} + S_{12} + S_p. \tag{40}$$

This last expression is the distribution curve of the two return losses in series.

APPENDIX III

ADDITION OF END PATH TO ACTIVE SINGING POINTS
(16 Ca. H-44-25 2-Wire Cable Side Circuits)

(1) Number of Repeater Sections	(2) Active Singing Point to Terminal Repeater on Zero—db	(3) Loss of End Path—db	(4) Active Singing Point with End Path Added—db	(5) (2) and (3) Added as Sum of Current Ratios	(6) (2) and (3) Added as Sum of Power Ratios
4	29.8	20.4	17.1	17.9	19.0
4	31.2	20.4	18.5	18.2	20.2
4	32.3	23.8	21.3	21.0	23.2
4	31.2	23.8	21.2	20.7	23.1
4	33.1	23.8	20.9	21.2	23.3
4	29.3	20.8	17.8	18.0	20.2
4	35.7	20.8	18.7	19.3	20.7
4	27.0	20.8	18.0	17.3	19.9
8	25.6	22.2	19.0	17.7	21.6
8	24.2	19.9	16.1	15.8	18.5
18	17.9	19.1	12.7	12.4	15.4
18	17.7	19.1	13.5	12.4	15.3
18	20.1	25.2	15.5	16.2	18.9
18	18.7	25.2	15.1	15.3	17.8
18	20.0	19.0	14.5	13.5	16.5
18	16.9	19.0	12.0	11.9	14.8
18	19.1	25.1	16.2	15.6	18.1
18	19.9	25.1	17.2	16.1	18.8
18	21.0	20.9	14.0	14.9	17.9
18	20.1	24.3	14.1	14.5	17.5
18	19.0	24.3	15.7	15.2	17.9
18	17.8	24.3	15.0	14.4	16.9
		Average =	16.45	16.34	18.88

The different tests using the same number of repeaters are different circuits and different testing repeaters. Columns (2) and (4) are measured singing points. Column (3) is the sum of the measured losses from the measuring point to the drop side of the terminal repeater and back, plus the return loss of resistance termination applied at the drop side of the repeater.

Columns (5) and (6) show computed values corresponding to column (4), on different assumptions as to the method of addition of the intermediate and end paths. The values in column (5) are generally much closer to the values in column (4) than are those in column (6).

APPENDIX IV

Distribution of Active Return Losses with Intermediate Paths Including Equipment

From a consideration of the paper, " Irregularities in Loaded Telephone Circuits," by G. Crisson,[7] it appears that it is reasonable to say that the total current returned from a cable section with near-end equipment will be

$$i_t = \sqrt{i_1^2 + i_e^2},$$

where i_1 is the current returned from the line without equipment and i_e is the current returned from the equipment. (It is assumed here that i_t and i_e are referred to the line side of the coil and not the drop side.)

Then the probability that the current returned from the cable alone will exceed I is

$$F(I) = \frac{1}{I_1^2}\int_I^\infty i_1 e - \frac{i_1^2}{2I_1^2}\,di_1 = e - \frac{I^2}{2I_1^2}.$$

The probability that the total current will exceed $I_3 = \sqrt{I^2 + i_e^2}$ is

$$P(i_t > I_3) = e - \frac{(I_3^2 - i_e^2)}{2I_1^2} = e\,\frac{i_e^2}{2I_1^2}\,F(I_3)$$

or

$$P(i_t > I_3) = \text{antilog}_{10}\left(.434\ \text{antilog}_{10}\left(\frac{S_1 - S_E}{10}\right)\right)F(I_3) = F',$$

where

$$S_1 = -10\log_{10} 2I_1^2 \qquad \text{and} \qquad S_E = -20\log_{10} i_e.$$

This equation says that the probability of getting more than a given value of return current I_3 from the combined cable section and near end equipment is equal to the product of the probability of getting I_3 from the cable section alone by the factor $e(i_e^2/2I_1^2)$. Or by analogy with the method used in the paper referred to above, letting the distribution curve of the combination return loss be $S_1 + S_{F'}$,

$$S_{F'} = -\log_{10}\left(\log_e\frac{1}{F} + \frac{i_e^2}{2I_1^2}\right)$$

$$= -\log_{10}\left(\log_e\frac{1}{F} + \text{antilog}_{10}\,\frac{S_1 - S_E}{10}\right)$$

$$S_{F'} = S_F \underset{p}{X} (S_E - S_1).$$

[7] Loc. cit.

Or the distribution curve of the return losses from the repeater section with equipment is

$$S_1 + S_{F'} = (S_1 + S_F) \underset{p}{\mathrm{X}} (S_1 + S_E - S_1)$$

$$= (S_1 + S_F) \underset{p}{\mathrm{X}} (S_E)$$

$$S_{F'} = (S_F) \underset{p}{\mathrm{X}} (S_E - S_1).$$

Curves of $S_{F'}$ could be drawn for different values of $S_E - S_1$, since S_F is known for a given value of F. These would apply either for passive return losses of a single repeater section or for active return losses of several repeater sections in tandem.

For example, suppose $S_E = S_1$. For $S_F = 0$, the value of F is about 0.368 and the value of $S_{F'} = 0 \underset{p}{\mathrm{X}} 0 = -3.01$ db for $F' = 0.368$. Or the same point could be derived from the earlier formulas as follows: When $S_F = -3.01$ db, $F = 0.1353$. The factor antilog $\left(0.434 \text{ antilog } \left(\frac{0}{10}\right)\right)$ equals 2.718 and the value of F' for $S_{F'} = -3.01$ db is $F' = (0.1353)(2.718) = 0.368$.

APPENDIX V

Distribution of Active Return Loss Made up of Intermediate Paths and a Fixed End Path

In a practical case, when an active singing point toward a circuit terminal is measured for maintenance purposes, the termination at the circuit drop will be a fixed known value of return loss rather than one selected at random from a distribution curve. The distribution curve of active singing points toward the drop will then be a curve obtained at each percentage, by the current sum addition of the active singing point with no current returned from beyond the terminal repeater and the end path singing point. If the active intermediate path singing point has a distribution curve $S_{11} - 2.5 + S_Q(2) = S_4 + S_Q(2)$ and the end path has a fixed loss S_5, the following table may be used to find the distribution curve of the combination $S_4 + S_g(d)$, for different values of $d = S_5 - S_4$. The table was derived from the equation

$$(S_4 + S_Q(2)) \underset{p}{\mathrm{X}} S_5 = S_4 + S_g(d)$$

Percentage of Cases with Lower Singing Points	0.003	0.13	2.2	15.7	50.0	84.3	97.8	99.87	99.997	99.99997
$S_g(\infty)$ in db..	−8	−6	−4	−2	0	2	4	6	8	10
$S_g(15)$ in db..	−8.6	−6.7	−4.9	−3.2	−1.4	+0.2	+1.8	+3.4	+4.8	+6.1
$S_g(10)$ in db..	−9.0	−7.3	−5.6	−4.0	−2.4	−0.9	+0.5	+1.1	+2.9	+4.0
$S_g(6)$ in db...	−9.6	−8.0	−6.4	−4.9	−3.5	−2.2	−1.1	0	+0.9	+1.8
$S_g(3)$ in db...	−10.2	−8.6	−7.2	−5.9	−4.6	−3.5	−2.5	−1.6	−0.9	−0.2
$S_g(0)$ in db...	−10.9	−9.5	−8.2	−7.1	−6.0	−5.1	−4.2	−3.5	−2.9	−2.4
$S_g(-3)$ in db.	−11.9	−10.6	−9.5	−8.5	−7.6	−6.9	−6.2	−5.6	−5.2	−4.8
$S_g(-6)$ in db.	−13.1	−12.0	−11.1	−10.2	−9.5	−8.9	−8.4	−8.0	−7.6	−7.3
$S_g(-10)$ in db	−15.1	−14.2	−13.5	−12.9	−12.4	−12.0	−11.5	−11.3	−11.0	−10.8

or

$$(S_Q(2)) \underset{p}{\mathrm{X}} (S_5 - S_4) = S_g(d).$$

It will be noted that $S_g(\infty) = S_Q(2.0)$.

It should be noted that the values of $S_g(d)$ corresponding to the higher percentages may be considerably modified by the effect of the equipment return losses, particularly when d is large. However, this is usually of no very great importance because it is the lower percentages and the smaller values of d which are of the most practical importance.

APPENDIX VI

NOMENCLATURE

A = attenuation factor per loading section (current ratio).

A_k = attenuation factor per loading section (current ratio) of k th repeater section from critical repeater.

a = $\log_{10}{}^{-1}\left(\dfrac{S_{11} + S_{12} - S_I}{20}\right)$.

B = S_1 minus the average singing point on a group of lines.

b = the standard deviation of a group of singing point measurements.

d = $S_5 - S_4$.

E = net loss of circuit in db.

E_1 = loss in end path toward West for zero terminal return loss. When primed, it is the loss at 1000 cycles.

E_2 = same as E_1 but toward East.

$F(I_F) = F$ = probability of obtaining a greater returned current then I_F from a randomly selected one of a group of lines.

F' = probability of obtaining a greater returned current than I_3 from a cable section with near end equipment.

$F(N, T)$ = $10 \log_{10} \left(\sum_{k=0}^{N-1} t_k^{-4k} \right)$.

G_C and g_C = gains as defined herein.

I = current obtained in a particular case.

I_0 = current sent into the line at the point of measurement.

I_1 = $\text{anti} \log_{10} \left(\frac{-S_1 - 3}{20} \right)$ = current for active return loss corresponding to I' from " Irregularities in Loaded Telephone Circuits " by G. Crisson in the October 1925 *Bell System Technical Journal.*

I_2 = $\text{anti} \log_{10} \left(\frac{-S_1 - 3}{10} \right)$ = current for active return loss corresponding to I'' from paper referred to under I_1.

I_3 = current defined above under F'.

I_F = current defined under $F(I_F)$ above.

I_{1k} = I' (from paper referred to above) for the currents from the k th repeater section from the critical repeater.

I_{2k} = I'' (from paper referred to above) for the currents from the k th repeater section from the critical repeater.

i_1 and i_2 = currents returned from particular active or passive return losses.

i_3 = current returned from a particular pair of active or passive return losses in series.

i_e = $\text{anti} \log_{10} \left(\frac{-S_E}{20} \right)$.

i_t = total current returned from line with equipment.

j = number of the loading section measured from the repeater at which the return loss is of interest.

$K_1(x)$ = Bessel function of the second kind and order unity with argument x.

k = the number of the repeater or repeater section from the critical repeater, e.g., for the adjacent repeater section $k = 1$, for the next $k = 2$, etc.

k_1 $= \text{anti}\log_{10}\left(\frac{-S_{11}-3}{20}\right)$ = current corresponding to S_{11} in same way as I_1 corresponds to S_1.

k_2 $= \text{anti}\log_{10}\left(\frac{-S_{12}-3}{20}\right)$ = current corresponding to S_{12} in same way as I_2 corresponds to S_1.

L = loss of a cable section.

L_k = the loss in db from the West side of the critical repeater to the output of the k th repeater to the West, plus the loss from the input of the k th repeater W–E to the input of the critical repeater W–E.

M_E = the end path singing margin, i.e., the singing margin with no currents returned from intermediate paths (e.g., with a four-wire circuit).

M_I = the internal singing margin on an active line, i.e., the singing margin without any currents returned from the end paths (e.g., with the terminal repeaters turned down).

M_S = the singing margin of the circuit; i.e., the singing margin with currents returned both from intermediate paths and from end paths.

m_k = total number of loading sections in the k th repeater section.

m = value of m_k when all repeater sections have the same number of loading sections.

N = number of repeater sections in the part of the circuit for which the active return loss is to be computed.

n = number of cases considered as a group.

p_x = relative probability of obtaining a returned current i_x from one active or passive return loss.

$P(x > y)$ = probability that the value of x selected at random from a distribution curve is greater than a pre-selected value y.

P = probability that a singing point will have a lower value than a pre-selected value y.

$P_m(n, y) = P_m$ = probability that the minimum of n values will be as high or higher than a pre-selected value y db below the true average.

Q = probability that a lower value of singing margin will be obtained.

q = frequency taper in db per repeater section.

R_L = the representative reflection coefficient (see paper by G. Crisson referred to above).

S_H = the irregularity function (see paper by G. Crisson referred to above).

S_W = the frequency function (see paper by G. Crisson referred to above).

S_A = the attenuation function (see paper by G. Crisson referred to above).

S_F = the distribution function (see paper by G. Crisson referred to above).

S = $S_H + S_W - S_A + S_F$.

S_1 = $S - S_F = S_H + S_W - S_A = -10 \log_{10} 2I_1^2$.

S_2 = Same as S_1 but for terminal return loss.

S_4 = $S_{11} - 2.5$.

S_5 = end path loss with fixed termination.

S_{11} and S_{12} = values, respectively, toward the West and the East of active or passive single frequency return loss for $S_F = 0$, made up of intermediate paths only.

S_{21} and S_{22} = same as S_{11} and S_{12} but for end paths.

S_{31} and S_{32} = same as S_{11} and S_{12} but for combination of intermediate and end paths.

S_I = return loss which is actually obtained on a particular case.

S_C = return loss in a particular case with equipment.

S_E = equipment return loss.

S_P = $-20 \log a$ = distribution function of two return losses in series, where each such return loss has a distribution curve of the form $S + S_F$.

S_F' = function corresponding to S_F after equipment has been added.

$S_Q(b)$ = distribution function of singing points or singing margins; i.e., a normal law with an average of zero and a standard deviation of b.

$S_g(d)$ = distribution functions of active return losses with fixed terminations.

t = $10^{-(T/20)}$.

T = taper in db per repeater section.

t_k = $10^{-(L_k/40k)}$.

V = $1 - P$.

V_m = $1 - P_m$.

$\underset{i}{\mathrm{X}}$ = a symbol indicating that the two quantities in decibels are combined as if their currents added directly.

E.g.,

$S_{11} \underset{i}{\mathrm{X}} S_{21} = S_{31}$ means that $10^{-(S_{11}/20)} + 10^{-(S_{21}/20)} = 10^{-(S_{31}/20)}$.

$\underset{p}{\mathrm{X}}$ = a symbol indicating that the two quantities in decibels so connected are combined as if their powers added directly.

E.g.,

$S_{11} \underset{p}{\mathrm{X}} S_{21} = S_{31}$ means that $10^{-(S_{11}/10)} + 10^{-(S_{21}/10)} = 10^{-(S_{31}/10)}$.

x, y = variables as used.

θ_{jk} = phase angle of current returned from the j th loading section in the k th repeater section.

Operation of Ultra-High-Frequency Vacuum Tubes

By F. B. LLEWELLYN

Previous electronics analyses are extended by the introduction of more general boundary conditions. The results are applied to the calculation of the rectifying properties of diodes at very high frequencies and to the amplifying properties of negative grid triodes at both low and high frequencies. The effect of space charge on the various capacitances in triodes is discussed, and formulas for the amplification factor and plate impedance are presented in terms of the tube geometry. Finally, a discussion of the input impedance of negative grid triodes is given together with a comparison of the theoretical value with the results of measurements made by several well-known experimenters.

IN the study of the functioning of vacuum tubes at ultra-high frequencies it has been necessary to retrace the steps followed in the early history of vacuum tube performance but with the difference that a microscope for viewing the path at close range must be substituted for the telescope with which the original trail was mapped from afar off. As a result of this microscopic survey, formulas have been developed which are applicable to frequencies so high that the time of flight of the electron across the tube may occupy several cycles of the high-frequency oscillation. In addition to this result, several by-products of the study are found to have a useful application in the low-frequency field and to throw additional light on the multitudinous activities of the electrons and their effect upon the external circuit.

In this respect it is particularly interesting to see the way in which the geometry of the vacuum tube enters into the determination of the amplification factor of negative grid triodes and to compare the results now obtained with the earlier results of such workers as Abraham, King, Schottky, Lane and Van der Bijl. The effect of the negative grid on the transit time of the electrons also yields low-frequency relations in which certain new facts concerning the plate resistance are brought out.

Various papers [1, 2, 3, 4, 5] published within the last few years have dealt with the general problem of vacuum tubes in which the electron transit time is of importance and have derived results which are useful in several practical applications. However, in none of these have the initial relations been general enough to allow more than a very rough application to be made to the most widely used tube of all—the negative grid triode. A paper [2] by the present writer contains certain general conclusions concerning the negative grid tube at very high

frequencies and derives expressions for the complex values taken by the amplification factor and plate impedance. While these expressions represent the general trend of the variations, they are based on certain a priori assumptions, as was emphasized in the paper, and hence partake of the telescopic viewpoint of the low-frequency vacuum tube analysis rather than the microscopic viewpoint which is now necessary.

It was with the aim of overcoming this limitation that the work described in the present paper was undertaken. Its successful outcome was made possible by a very simple generalization of the methods described in the references, but one which has such far-reaching consequences that it appears worthwhile to start the analysis at the very beginning, and abbreviate only to the extent that intermediate algebraic steps are omitted because of their unwieldy length.

For convenience, the paper is divided into four parts. In Part I the mathematical analysis is outlined in its fundamental form and general working formulas are developed. In a way similar to low-frequency analysis these formulas may be divided into constant or d.-c. relations, first-order relations, second-order relations, and so on; where the first-order relations apply to a.-c. effects for small amplitudes only, the second-order relations contain rectification and distortion terms; and so on, in exact correspondence with the well known low-frequency relations. Part II contains the solution of the first-order relations expressed in appropriate form for later computations, while Part III does the same thing for second-order relations. Also in Part III, by way of illustration, the effect of frequency on the rectifying properties of parallel plane diodes is discussed. Part IV applies the general first-order and d.-c. solutions to the negative grid triode and shows how its various properties depend on frequency. A discussion of the important effect of active grid loss is included.

In certain cases the same formulas are expressed in several different ways. This is done because of the difficulty of determining the most useful method of expression before a large number of applications shall have been made. In most cases the general equations have been arranged to conform as far as possible with the most widely used modes of expression of the corresponding low-frequency equations. Where a choice of modes of expression is available, both modes have usually been given, it being left to future experience to determine the more useful one. While this procedure results in some repetition, the two modes of expression of the same equation are found in many instances to have their individual advantages, the one being more suitable for one type of application while the other is more particularly adapted to a different application.

Part I—General Analysis

In line with most previous electronic papers consideration is here directed to the behavior of electrons between two parallel planes of practically infinite extent. Differing, however, from earlier works, neither plane is to be regarded as constituting a thermionic emitter or cathode in the general sense. For certain special applications conditions may be chosen so that one of the planes coincides with, and assumes the properties of a zero potential cathode, but in developing the general relations this idea is strictly avoided. It will therefore be premised as a starting point that the velocity and the acceleration of the electrons at one of these two planes are given as initial conditions. It will be found that this generalization completely avoids certain ambiguities which were discussed in the February 1935 issue of the *Proceedings of the Institute of Radio Engineers.*[4] It also allows application of the results to be made to a much wider range of devices, including a fairly rigorous treatment of the negative grid triode.

In the following analysis, and differing from previous references, a change in the units employed has been made so that all quantities are expressed immediately in the practical system of engineering units (amperes, volts, ohms, coulombs, etc.) instead of in the electrostatic and electromagnetic systems. This change has been found to be of great advantage in the use of the equations, since it obviates all necessity for the continuous and irritating transformation of units that accompanies the electrostatic and electromagnetic systems.

The analysis starts with the expression for the total current density

$$I = \rho u + \epsilon \frac{\partial E}{\partial t}, \tag{1}$$

where I is current density, amperes per cm.^{2},
ρ is charge density, coulombs per cm.^{3},
u is electron, or charge velocity, cm. per sec.,
ϵ is permittivity of a vacuum, which is $1/36\pi 10^{11} = 8.85 \times 10^{-14}$ farads/cm.,
E is the electric intensity, volts per cm.

The equation of motion of an electron is

$$eE = kma, \tag{2}$$

where e is electronic charge, coulombs,
m is electronic mass, grams,

$e/m = -1.77 \times 10^8$ coulombs per gm.,*
$k = 10^{-7}$ is the ratio of dyne-cm. to joules,
a is electron acceleration, cm. per sec.^{2}

In this equation the effect of a magnetic field is disregarded. This is thoroughly justified until electron velocities approach that of light or the spacing between the two parallel planes becomes comparable with the wave-length of any alternating field considered. A more detailed discussion is given by Benham.[1]

A third fundamental equation is

$$\text{div } \epsilon E = \rho,$$

which, for the parallel planes now considered, becomes

$$\epsilon \frac{\partial E}{\partial x} = \rho. \tag{3}$$

From (1), (2) and (3) is obtained

$$\frac{eI}{km\epsilon} = \frac{da}{dt}. \tag{4}$$

The total current, I, may be considered to be composed of two parts, the first being a constant component and the second being a function of time only. On this basis we can write

$$\frac{eI}{km\epsilon} = K + \varphi'''(t), \tag{5}$$

where K is the constant part, and $\varphi'''(t)$ is the variable part, the primes denoting derivatives with respect to the argument in parentheses. Inserting (5) into (4) and integrating once with respect to time, we find:

$$a = K(t - t_a) + \varphi''(t) - \varphi''(t_a) + a_a + \alpha(t_a), \tag{6}$$

where $a_0 + \alpha(t_a)$ is the acceleration when $t = t_a$ and a_a is independent of t_a.

Another integration gives

$$u = K\frac{(t - t_a)^2}{2} + \varphi'(t) - \varphi'(t_a) - (t - t_a)\varphi''(t_a) + (t - t_a)a_a + (t - t_a)\alpha(t_a) + u_a + \mu(t_a), \tag{7}$$

where $u_a + \mu(t_a)$ is the velocity when $t = t_a$ and u_a is independent of t_a.

* This value is based on deflection measurements (which are applicable to vacuum tube analysis) rather than on spectroscopic measurements which give 1.76×10^8. Compton and Langmuir [6] use the spectroscopic figure. For a comprehensive discussion of values of physical constants, see Birge, *Phys. Rev. Supp.*, Vol. 1, July, 1929.

A third integration gives

$$x = K\frac{(t-t_a)^3}{6} + \varphi(t) - \varphi(t_a) - (t-t_a)\varphi'(t_a) - \frac{(t-t_a)^2}{2}\varphi''(t_a) + \frac{(t-t_a)^2}{2}a_a + \frac{(t-t_a)^2}{2}\alpha(t_a) + (t-t_a)u_a + (t-t_a)\mu(t_a), \quad (8)$$

where x is zero when $t = t_a$.

This choice of initial conditions allows one of the two parallel planes to be located at the position where x is zero and where the electron velocities and accelerations are given as above in terms of the time instant t_a when the electron crosses the plane, which may be referred to as the " a " plane. When these initial conditions are specified, then (6), (7), and (8) allow the acceleration, velocity and position, respectively, to be determined at any time t, thereafter.

These quantities are expressed in terms of t and t_a whereas it is more convenient in vacuum tube work to have the acceleration and velocity expressed in terms of t and x. Ideally, this could be done by solving (8) for t_a and thence eliminating t_a from (6) and (7). Practically, (8) cannot be solved directly for t_a because it is a higher order equation and involves φ, α and μ which may be (and usually are) transcendental functions of t_a. However, an indirect method can be employed.

If φ, α and μ were zero, then x would be given by the relatively simple equation

$$x = K\frac{(t-t_a)^3}{6} + a_a\frac{(t-t_a)^2}{2} + u_a(t-t_a). \quad (9)$$

Although this is a cubic, nevertheless $(t - t_a)$ may be obtained with relative ease in any particular case. The use of a new variable T to replace x is suggested by (9) and accordingly the defining equation of T under all conditions is taken to be:

$$x = K\frac{T^3}{6} + a_a\frac{T^2}{2} + u_aT, \quad (10)$$

which holds even when φ, α and μ are not zero. It is evident when φ, α and μ are small that T does not differ very much from $t - t_a$ as (10) must then become nearly equivalent to (9). It thus seems expedient to write in general

$$t - t_a = T + \delta, \quad (11)$$

and note that δ becomes very small when φ, α and μ are small.

On the basis of (11), functions of $(t - t_a)$ may be expanded into series in powers of δ as follows:

$$f(t-t_a) = f(T+\delta) = f(T) + f'(T)\delta + \frac{1}{2!}f''(T)\delta^2 + \cdots, \quad (12)$$

and similarly functions of t_a may be written

$$f(t_a) = f(t - T - \delta) = f(t - T) - f'(t - T)\delta + \frac{1}{2!}f''(t - T)\delta^2 - \cdots. \quad (13)$$

When (10), (11), (12) and (13) are used in conjunction with (8) the result is a relation between t, T and δ as follows:

$$\begin{aligned}
0 = {} & \frac{K}{6}[3T^2\delta + 3T\delta^2 + \delta^3] \\
& + \frac{a_a}{2}[2T\delta + \delta^2] + u_a\delta \\
& + \varphi(t) - \left[\varphi(t-T) - \varphi'(t-T)\delta + \frac{1}{2!}\varphi''(t-T)\delta^2 - \cdots\right] \\
& - (T+\delta)\left[\varphi'(t-T) - \varphi''(t-T)\delta + \frac{1}{2!}\varphi'''(t-T)\delta^2 - \cdots\right] \\
& - \frac{1}{2}(T^2 + 2T\delta + \delta^2)\left[\varphi''(t-T) - \varphi'''(t-T)\delta + \frac{1}{2!}\varphi''''(t-T)\delta^2 - \cdots\right] \\
& + \frac{1}{2}(T^2 + 2T\delta + \delta^2)\left[\alpha(t-T) - \alpha'(t-T)\delta + \frac{1}{2!}\alpha''(t-T)\delta^2 - \cdots\right] \\
& + (T+\delta)\left[\mu(t-T) - \mu'(t-T)\delta + \frac{1}{2!}\mu''(t-T)\delta^2 - \cdots\right]. \quad (14)
\end{aligned}$$

This equation may be written in the form of a power series in δ. It cannot yet be solved directly for δ. It has the advantage, however, that δ is not involved in the transcendental functions φ, α and μ, so that an indirect method may be used. Let δ, φ, α and μ each be split up into series as follows:

$$\left.\begin{aligned}
\delta &= \delta_1 + \delta_2 + \delta_3 + \text{etc.} \\
\varphi &= \varphi_1 + \varphi_2 + \varphi_3 + \text{etc.} \\
\alpha &= \alpha_1 + \alpha_2 + \alpha_3 + \text{etc.} \\
\mu &= \mu_1 + \mu_2 + \mu_3 + \text{etc.}
\end{aligned}\right\} \quad (15)$$

These are to be substituted into (14) and the resulting expression may then be expressed as an infinite series of separate equations such that the first equation includes all linear terms which have the subscript 1, but no other terms; The second equation includes all linear terms having the subscript 2 and also all quadratic terms having the subscript 1; the third equation includes all linear terms having the subscript 3, cubic terms with subscript 1, and also products of quadratic terms with subscript 1 and linear terms with subscript 2. The rules for succeeding equations are analogous, so that in general, the sum of the subscripts of each term of the nth equation is equal to n.

The first of these equations may be solved for δ_1, the second for δ_2, the third for δ_3, and so on, giving the following:

$$\delta_1 = -\frac{\left[\varphi_1(t) - \varphi_1(t-T) - T\varphi_1'(t-T) - \frac{1}{2}T^2\varphi_1''(t-T) + \frac{1}{2}T^2\alpha_1(t-T) + T\mu_1(t-T)\right]}{K\frac{T^2}{2} + a_aT + u_a}, \quad (16)$$

$$\delta_2 = -\frac{\left[\begin{array}{l}\varphi_2(t) - \varphi_2(t-T) - T\varphi_2'(t-T) - \frac{1}{2}T^2\varphi_2''(t-T) \\ \quad + \frac{1}{2}T^2\alpha_2(t-T) + T\mu_2(t-T) \\ \quad + \delta_1\left(\frac{1}{2}T^2\varphi_1'''(t-T) - \frac{1}{2}T^2\alpha_1'(t-T)\right. \\ \qquad \left. + T\alpha_1(t-T) - T\mu_1'(t-T) + \mu_1(t-T)\right) \\ \quad + \delta_1^2\left(K\frac{T}{2} + \frac{1}{2}a_a\right)\end{array}\right]}{K\frac{T^2}{2} + a_aT + u_a}, \quad (17)$$

$$\delta_3 = -\frac{\left[\begin{array}{l}\varphi_3(t) - \varphi_3(t-T) - T\varphi_3'(t-T) - \frac{1}{2}T^2\varphi_3''(t-T) \\ \quad + \frac{1}{2}T^2\alpha_3(t-T) + T\mu_3(t-T) \\ \quad + \delta_2\left(\frac{1}{2}T^2\varphi_1'''(t-T) - \frac{1}{2}T^2\alpha_1'(t-T)\right. \\ \qquad \left. + T\alpha_1(t-T) - T\mu_1'(t-T) + \mu_1(t-T)\right) \\ \quad + \delta_1\left(\frac{1}{2}T^2\varphi_2'''(t-T) - \frac{1}{2}T^2\alpha_2'(t-T)\right. \\ \qquad \left. + T\alpha_2(t-T) - T\mu_2'(t-T) + \mu_2(t-T)\right) \\ \quad + \delta_1^2\left(-\frac{1}{4}T^2\varphi_1''''(t-T) + \frac{1}{2}T\varphi_1'''(t-T)\right. \\ \quad \left. + \frac{1}{4}T^2\alpha_1''(t-T) - T\alpha_1'(t-T) + \frac{1}{2}\alpha_1(t-T)\right. \\ \qquad \left. + \frac{1}{2}T\mu_1''(t-T) - \mu_1'(t-T)\right) \\ \quad + \delta_1\delta_2\left(KT + a_a\right) + \delta_1^3\frac{K}{6}\end{array}\right]}{K\frac{T^2}{2} + a_aT + u_a}, \quad (18)$$

$$\delta_4 = \cdots. \quad (19)$$

These formulas together with (11) and (15) allow the acceleration (6) and velocity (7) to be written in terms of t and T, and hence effectively in terms of t and x, for T is a function of x only, as given by (10). The resulting equations for acceleration and velocity are conveniently broken up into a series of equations in accord with the same rules formulated for obtaining the δ's from (14). Thus, we write

$$a = a_0 + a_1 + a_2 + a_3 + \text{etc.}, \tag{20}$$

where a_0 comprises those terms having zero or no subscript, and obtain from (6):

$$a_0 = KT + a_a, \tag{21}$$

$$a_1 = K\delta_1 + \varphi_1''(t) - \varphi_1''(t - T) + \alpha_1(t - T), \tag{22}$$

$$a_2 = K\delta_2 + \varphi_2''(t) - \varphi_2''(t - T) + \varphi_1'''(t - T)\delta_1 + \alpha_2(t - T) - \alpha_1'(t - T)\delta_1, \tag{23}$$

$$a_3 = K\delta_3 + \varphi_3''(t) - \varphi_3''(t - T) + \varphi_1'''(t - T)\delta_2 + \varphi_2'''(t - T)\delta_1 - \frac{1}{2}\varphi_1''''(t - T)\delta_1^2 + \alpha_3(t - T) - \alpha_2'(t - T)\delta_1 - \alpha_1'(t - T)\delta_2 + \frac{1}{2}\alpha_1''(t - T)\delta_1^2, \tag{24}$$

$$a_4 = \cdots. \tag{25}$$

Treating the velocity in a similar way we write

$$u = u_0 + u_1 + u_2 + u_3 + \text{etc.} \tag{26}$$

and obtain from (7)

$$u_0 = K\frac{T^2}{2} + a_aT + u_a, \tag{27}$$

$$u_1 = KT\delta_1 + a_a\delta_1 + \varphi_1'(t) - \varphi_1'(t - T) - T\varphi_1''(t - T) + T\alpha_1(t - T) + \mu_1(t - T), \tag{28}$$

$$u_2 = KT\delta_2 + \frac{K}{2}\delta_1^2 + a_a\delta_2 + \varphi_2'(t) - \varphi_2'(t - T) - T\varphi_2''(t - T) + T\varphi_1'''(t - T)\delta_1 + T\alpha_2(t - T) + \alpha_1(t - T)\delta_1 - T\alpha_1'(t - T)\delta_1 + \mu_2(t - T) - \mu_1'(t - T)\delta_1, \tag{29}$$

$$u_3 = KT\delta_2 + K\delta_1\delta_2 + a_a\delta_3 + \varphi_3'(t) - \varphi_3'(t - T) + \frac{1}{2}\varphi_1'''(t - T)\delta_1^2 - T\varphi_3''(t - T) + T\varphi_1'''(t - T)\delta_2$$

$$+ T\varphi_2'''(t - T)\delta_1 - \frac{1}{2}T\varphi_1''''(t - T)\delta_1^2 + T\alpha_3(t - T)$$
$$- T\alpha_1'(t - T)\delta_2 - T\alpha_2'(t - T)\delta_1 + \frac{1}{2}T\alpha_1''(t - T)\delta_1^2$$
$$+ \alpha_1(t - T)\delta_2 + \alpha_2(t - T)\delta_1 - \alpha_1'(t - T)\delta_1^2 + \mu_3(t - T)$$
$$- \mu_1'(t - T)\delta_2 - \mu_2'(t - T)\delta_1 + \frac{1}{2}\mu_1''(t - T)\delta_1^2. \tag{30}$$

Aside from their length these equations are not complicated and in applications to special cases many terms are apt to vanish, leaving relatively compact expressions.

In circuit work the potential difference between the two parallel planes, " a " and " b," say, is more often required than the electron acceleration. This may be found from the definition of the potential difference, namely

$$V_a - V_b = \int_a^b E\partial x, \tag{31}$$

in which t remains constant during the integration. From (10)

$$\partial x = \left(K\frac{T^2}{2} + a_a T + u_a \right) dT = u_0 dT, \tag{32}$$

so that with the aid of this equation and (2) the potential difference is given by:

$$W_a - W_b = \int_a^b a dx = \int_0^T a\left(K\frac{T^2}{2} + a_a T + u_a \right) dT, \tag{33}$$

where the symbol, W, is used as an abbreviation for eV/km.

In the same way as the acceleration and velocity are divided into components, the potential difference may be split up into $(W_a - W_b)_0$, $(W_a - W_b)_1$, $(W_a - W_b)_2$, etc. which are defined by:

$$(W_a - W_b)_0 = \int_0^T a_0 u_0 dT, \tag{34}$$

$$(W_a - W_b)_1 = \int_0^T a_1 u_0 dT, \tag{35}$$

$$(W_a - W_b)_2 = \int_0^T a_2 u_0 dT, \tag{36}$$

$$(W_a - W_b)_3 = \int_0^T a_3 u_0 dT, \tag{37}$$

and where the δ's given by (16), (17), (18), etc., are to be inserted in the a's given by (21), (22), (23), etc., before the integration can be carried out.

The formal solution of the problem has been reached with the attainment of (34), (35), (36), etc. As soon as specific functions are chosen for the current, $[K + \varphi'''(t)]km\epsilon/e$, the initial accceleration, $a_a + \alpha(t)$, and the initial velocity, $u_a + \mu(t)$, the integrations can be performed to give the potential difference between two planes located respectively at $x = 0$ and $x = x$.

While the general relation between current and potential is non-linear, the first order current $\varphi'''(t)e/km\epsilon$ is linearly related to the first order potential difference $(W_a - W_b)e/km$, and the results for this case can be expressed conveniently by using complex functions in the manner usual with electrical engineers. This will be done in some of the following applications.

In the treatment of second and higher-order components, it is convenient to select the current components to correspond with powers of the potential rather than vice versa. For example, $(V_a - V_b)_1$ is taken to be the complete expression for the fluctuating component of potential, thus causing $(V_a - V_b)_2$, $(V_a - V_b)_3$, etc. to vanish. Equation (36) then yields an expression for the second-order current in terms of the first-order current, and (37) does likewise for the third-order current.

The general equations are, however, applicable to any converging method of selecting the components, and the proper one for any particular case is to be determined by considerations of simplicity and convenience.

Part II—First-Order Solution

This is the linear case, so that to each component of current, $A \sin \omega t$, say, there corresponds a potential component of the form $B \sin (\omega t + \eta)$. It follows that a current of the form Ae^{pt} will produce a corresponding potential difference Pe^{pt}, and that p may be taken to be a generalized exponent having the value $i\omega$ when sinusoidal currents are considered. In the latter case, P will usually be complex. The generalized exponent p results in a more compact symbolism than would be possible with the imaginary exponent, $i\omega$.

Thus, we write for the current

$$\frac{eI}{km\epsilon} = K + \varphi_1'''(t) = K + Je^{pt}. \tag{38}$$

In a similar way the initial fluctuating components of acceleration and

velocity are taken to be

$$\alpha_1(t - T) = Ge^{p(t-T)} = (Ge^{-pT})e^{pt}, \tag{39}$$

$$\mu_1(t - T) = He^{p(t-T)} = (He^{-pT})e^{pt}. \tag{40}$$

With this nomenclature, all terms appearing in the first-order equations will contain the factor e^{pt}, which may accordingly be omitted throughout. Thus, instead of writing Ge^{pt} for the acceleration at the " a " plane, the single symbol α_1 will be used, where the omission of the functional notation $\alpha_1(t - T)$ indicates that the acceleration is taken at the " a " plane where T is zero, and that the multiplying factor e^{pt} is understood. In a similar way the symbol μ_1 will indicate the first-order fluctuating component of velocity at the " a " plane with the factor e^{pt} understood.

A still further symbolism will be of assistance. The quantity pT will be denoted by β. When p is the imaginary $i\omega$ then ωT is the transit angle, θ, which has been defined in previous papers,[1, 2] and in the sinusoidal case $\beta = i\theta$.

With this nomenclature, the first-order potential difference, (35), acceleration, (22), and velocity (28) may be written:

$$(W_a - W_b)_1 = \frac{J}{p^4}\left[K\left(\frac{\beta^3}{6} - \beta - 2e^{-\beta} - \beta e^{-\beta} + 2\right) + pa_a\left(\frac{\beta^2}{2} + \beta e^{-\beta} + e^{-\beta} - 1\right) + p^2u_a(\beta + e^{-\beta} - 1)\right] - \frac{\alpha_1}{p^2}\left[a_a(\beta e^{-\beta} + e^{-\beta} - 1) + u_a p(e^{-\beta} - 1)\right] + \frac{\mu_1}{p^2}\left[K(\beta e^{-\beta} + e^{-\beta} - 1)\right], \tag{41}$$

$$a_1 = \frac{J}{p}\left[1 - e^{-\beta} - \frac{K}{p^2u_0}\left(1 - e^{-\beta} - \beta e^{-\beta} - \frac{\beta^2}{2}e^{-\beta}\right)\right] + \alpha_1\left(1 - \frac{KT^2}{2u_0}\right)e^{-\beta} - \mu_1\frac{KT}{u_0}e^{-\beta}, \tag{42}$$

$$u_1 = \frac{J}{p^2}\left[1 - e^{-\beta} - \beta e^{-\beta} - \frac{a_0}{pu_0}\left(1 - e^{-\beta} - \beta e^{-\beta} - \frac{\beta^2}{2}e^{-\beta}\right)\right] + \frac{\alpha_1}{p}\left(1 - \frac{a_0T}{2u_0}\right)\beta e^{-\beta} + \mu_1\left(1 - \frac{a_0T}{u_0}\right)e^{-\beta}. \tag{43}$$

An alternate method of procedure, which is particularly useful at moderately low frequencies where the transit angle θ is only a few radians, is to expand functions of $(t - T)$ into power series in T. The

same result may be obtained from (41), (42) and (43) above by writing $e^{-\beta}$ in series form. The result is:

$$(W_a - W_b)_1 = \sum_{n=0}^{\infty} (-\beta)^n \left[J\left(\frac{KT^4(n+2)}{(n+4)!} + \frac{a_a T^3(n+2)}{(n+3)!} + \frac{u_a T^2}{(n+2)!} \right) + \alpha_1 \left(\frac{a_a T^2(n+1)}{(n+2)!} + \frac{u_a T}{(n+1)!} \right) - \mu_1 \left(\frac{KT^2(n+1)}{(n+2)!} \right) \right], \tag{41a}$$

$$a_1 = \sum_{n=0}^{\infty} \frac{(-\beta)^n}{n!} \left[JT \left(\frac{1}{n+1} - \frac{KT^2}{2u_0} \frac{1}{(n+3)} \right) + \alpha_1 \left(1 - \frac{KT^2}{2u_0} \right) - \mu_1 \left(\frac{KT}{u_0} \right) \right], \tag{42a}$$

$$u_1 = \sum_{n=0}^{\infty} \frac{(-\beta)^n}{n!} \left[JT^2 \left(\frac{1}{n+2} - \frac{a_0 T}{2u_0} \frac{1}{(n+3)} \right) + \alpha_1 T \left(1 - \frac{a_0 T}{2u_0} \right) + \mu_1 \left(1 - \frac{a_0 T}{u_0} \right) \right]. \tag{43a}$$

These latter forms of expression clearly show what happens at low frequencies and they prove that the equations do not give infinite values for any of the components so long as $a_0 T/u_0$ and KT^2/u_0 remain finite. Now in general

$$\frac{a_0 T}{u_0} = \frac{(KT + a_a)T}{K\frac{T^2}{2} + a_a T + u_a}$$

and

$$\frac{KT^2}{u_0} = \frac{KT^2}{K\frac{T^2}{2} + a_a T + u_a},$$

and these remain finite for all finite positive values of T, K, a_a and u_a. Thus the difficulty at the origin which was discussed [4] in the *Proceedings of the Institute of Radio Engineers*, February, 1935 is overcome by the generalized definition of T in (10). It is true that δ_1 still tends toward infinity when T approaches zero if a_a and u_a are zero and μ_1 is different from zero. This means that the ratio of the variation in transit time, δ_1, to the transit time T tends toward infinity when T approaches zero, and when variations in initial velocity are still present with no constant initial acceleration. This is a logical and expected result, but it leads also to the conclusion that the electrons actually halt their forward

motion and reverse their direction. When this happens, (1) is no longer applicable because the velocity at a given point has become multi-valued. This restriction was pointed out by Müller [3] as limiting any analysis which starts with (1).

It may be concluded then that the inherent limitation on (41), (42) and (43) is a singly valued velocity rather than the behavior of the δ's in the neighborhood of a point where the acceleration is zero.

Part III—Second-Order Solution

The second-order solution has practical utility in the computation of distortion and of the modulating and detecting properties of ultra-high-frequency thermionic systems. Even in low-frequency applications such computations are long and tedious. The introduction of the added complication of appreciable transit angles causes further difficulty because of the unwieldy length of the equations. Accordingly, instead of a general exposition of second-order effects, a greatly simplified special case will be treated at the present time, leaving the details of a more general solution until the need for it has become more acute.

The rectifying properties of a parallel plane diode operating with complete space charge will be calculated. The complete space charge condition is defined by placing initial velocities and accelerations equal to zero. When this has been done in (16), (17) and (23) and the resulting values of the δ's have been substituted in (23), functions of $(t - T)$ may be expanded into power series to give the following:

$$a_2 = -2\sum_{n=0}^{\infty}\frac{(-T)^{n+1}\varphi_2^{(n+3)}(t)}{(n+3)(n+1)!} + \frac{1}{K}\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}\frac{(-T)^{n+m+1}\varphi_1^{(n+3)}(t)\varphi_1^{(m+3)}(t)}{(n+3)(m+3)n!m!}. \quad (44)$$

The second-order potential difference is given by (36) where $u_0 = KT^2/2$ for complete space charge. Thus, from (44)

$$(W_a - W_b)_2 = K\sum_{n=0}^{\infty}\frac{(n+2)}{(n+4)!}(-T)^{n+5}\varphi_2^{(n+3)}(t) - \frac{1}{2}\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}\frac{(-T)^{n+m+4}\varphi_1^{(n+3)}(t)\varphi_1^{(m+3)}(t)}{(n+m+4)(n+3)(m+3)n!m!}. \quad (45)$$

The second-order potential difference $(W_a - W_b)_2$ will now be taken as zero, which implies that the first-order potential difference only is

impressed on the diode. This gives from (45)

$$K\left[\frac{1}{12}T^4\varphi_2'''(t) - \frac{1}{40}T^5\varphi_2^{iv}(t) + \frac{1}{180}T^6\varphi_2^{v}(t) - \cdots\right]$$
$$= \frac{1}{2}\left[\frac{T^4\varphi_1'''(t)\varphi_1'''(t)}{36} + \frac{T^6\varphi_1^{iv}(t)\varphi_1^{iv}(t)}{96} + \frac{T^8\varphi_1^{v}(t)\varphi_1^{v}(t)}{800} + \cdots\right]$$
$$+\left[-\frac{T^5\varphi_1'''(t)\varphi_1^{iv}(t)}{60} + \frac{T^6\varphi_1'''(t)\varphi_1^{v}(t)}{180} - \frac{T^7\varphi_1'''(t)\varphi_1^{vi}(t)}{756}\right.$$
$$\left.+\frac{T^8\varphi_1'''(t)\varphi_1^{vii}(t)}{4032} - \frac{T^7\varphi_1^{iv}(t)\varphi_1^{v}(t)}{224} + \frac{T^8\varphi_1^{iv}(t)\varphi_1^{vi}(t)}{1152} - \cdots\right]. \quad (46)$$

If the first-order voltage which is impressed on the diode is taken to be a single sine wave, it follows from the linearity of the first-order relations that the first-order current is likewise a single sine wave of the same angular frequency, ω. Thus, let

$$\varphi_1'''(t) = A \sin \omega t. \quad (47)$$

Then

$$\varphi_1^{iv}(t) = A\omega \cos \omega t,$$
$$\varphi_1^{v}(t) = -A\omega^2 \sin \omega t,$$
$$\varphi_1^{vi}(t) = -A\omega^3 \cos \omega t.$$

When these are substituted into (46) it is seen that the right hand side of the equation contains a d.-c. term and a double-frequency term. The left hand side must accordingly contain terms of the same frequencies. Hence the most general form which can be assumed for the second-order current $\varphi_2'''(t)e/km\epsilon$ is:

$$\varphi_2'''(t) = (a_0 + a_1\theta + a_2\theta^2 + \cdots)$$
$$+ (b_0 + b_1\theta + b_2\theta^2 + \cdots) \sin 2\omega t$$
$$+ (c_0 + c_1\theta + c_2\theta^2 + \cdots) \cos 2\omega t, \quad (48)$$

where $\theta = \omega T$ is the transit angle, and the a's have no reference to the symbols previously used for acceleration.

When (47) and (48) are substituted into (46) it will be seen that the coefficients of the d.-c. term, the sin $2\omega t$ term and the cos $2\omega t$ term respectively may be equated on the two sides of the equation. This gives three equations and in each of these, coefficients of corresponding powers of the transit angle may be equated, thus providing the values of all of the coefficients in (48).

Without carrying out this procedure in detail it is possible to find the d.-c. component directly from (46) and (47) by noting that time derivatives of the d.-c. component of $\varphi_2'''(t)$ are zero. Hence the left

hand side of (46) reduces to its first term. Substitution of (47) into the right hand side and selection of d.-c. components then gives:

$$\varphi_2'''(t)_{\text{d.-c.}} = \frac{1}{12}\frac{A^2}{K}\left[1 - \frac{\theta^2}{40} + \frac{\theta^4}{2800} - \cdots\right]. \tag{49}$$

This indicates that the rectified current decreases when the transit angle θ becomes appreciable. In the second part of his first electronics paper,[1] Benham reached the conclusion that the rectified current increases with frequency. To reconcile his result with (49) it is only necessary to note that (49) indicates a decrease in rectified current with frequency provided that the amplitude A of the first-order current remains constant, whereas Benham's result was based on a constant amplitude of the first-order voltage. A direct comparison therefore necessitates the computation of the first-order voltage. From (41a) with zero initial acceleration and velocity we have:

$$(W_a - W_b)_1 = \frac{JKT^4}{12}\left[1 - \frac{3}{10}i\theta - \frac{1}{15}\theta^2 + \frac{1}{84}i\theta^3 + \frac{1}{560}\theta^4 \cdots\right]. \tag{50}$$

Taking the amplitude of the current factor $\varphi_1'''(t)$ to be A as in (47) and the amplitude of the voltage factor $(W_a - W_b)_1$ to be B, we find from (50) that

$$B^2 = A^2\frac{K^2T^8}{144}\left[\left(1 - \frac{1}{15}\theta^2 + \frac{1}{560}\theta^4 - \cdots\right)^2 + \left(-\frac{3}{10}\theta + \frac{1}{84}\theta^3 - \cdots\right)^2\right]$$

or

$$B^2 = A^2\frac{K^2T^4}{144}\left[1 - \frac{13}{300}\theta^2 + \frac{11}{12600}\theta^4 - \cdots\right]. \tag{51}$$

The A^2 may thus be eliminated between (51) and (49) giving

$$\varphi_2'''(t)_{\text{d.-c.}} = \frac{12B^2}{K^3T^8}\left[1 + \frac{11}{600}\theta^2 + \frac{39}{140000}\theta^4 + \cdots\right], \tag{52}$$

which is in agreement with the result obtained by Benham.

The physical explanation of the increase in rectified current which occurs when the transit angle becomes appreciable follows directly when it is realized that as the frequency is made higher, the fundamental component of the alternating current increases when the a.-c. voltage is maintained at constant amplitude. This is because of the capacitance of the diode. It is this increase in fundamental current

which causes the increase in rectified current, and not the effect of transit angle directly.

Part IV—Negative Grid Triodes

The general aspect of the negative grid triode is shown in Fig. 1. The upper diagram shows a plan view of the electrode arrangement and

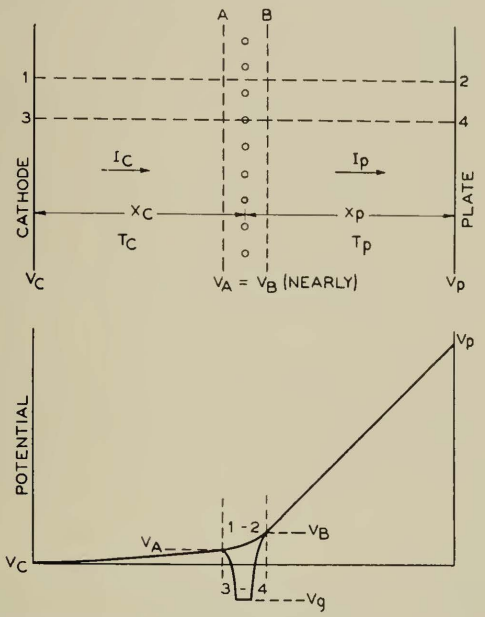


Fig. 1.—Nomenclature and potential distribution in negative grid triodes.

indicates the nomenclature which will be adopted. The planes at V_c and V_p constitute the cathode and plate, respectively, and the grid wires are indicated by the small circles. The planes A and B are imaginary planes located on opposite sides of the grid and only far enough away so that the irregularities in the potential distribution caused by the individual negative grid wires have practically disappeared. An analysis of the potential distribution in the immediate neighborhood of a shielding screen has been made by Maxwell (Treatise on Electricity and Magnetism Vol. I) and this applies to the grid of the triode in the absence of electron flow. The presence of electrons between the negative grid wires will tend to decrease the irregularities shown in Maxwell's analysis and therefore allow the planes A and B to be placed somewhat closer together than his figures would indicate. As far as the writer is aware, exact relations for the potential distribution near the grid wires in the presence of electron flow have not been

worked out so that Maxwell's analysis is our only basis at the present time for determining how near to the grid we can imagine A and B to be fixed.

The distance from cathode to grid will be called x_c and from grid to plate will be called x_p. Corresponding to x_c there will be an electron transit time T_c, and to x_p there will be the transit time T_p. The density of the total current leaving the cathode is I_c so that the current factor $eI/km\epsilon$ will be J_c. The factor for the density of the total current reaching the plate will be J_p.

The general scheme for analysis of the triode is to regard the structure as composed of two parallel plane diodes. The first of these comprises the cathode and the imaginary plane A, while the second comprises the imaginary plane B and the plate or anode. In accord with normal operation, complete space charge is postulated at the cathode, so that initial velocities and accelerations are zero, and the general parallel plane equations, (41), (42) and (43) may be applied directly to conditions between the cathode and the plane A.

The plane B and the plate also constitute a structure to which the general equations can be applied, but first the initial velocities and accelerations at B will have to be found, and finally a relation between J_c and J_p will be needed. Moreover, the relation of the potentials at A and B to the potential V_g of the grid wires themselves must be found, because it is only the latter that is available for external use or measurement.

As an aid to finding these various relations the lower diagram in Fig. 1 will be of assistance. This is a graph showing the general form of the d.-c. potential as a function of the distance from the cathode. Between cathode and plane A the potential curve is the same regardless of whether the line 1–2 in the upper diagram which passes between two grid wires is followed, or whether the line 3–4 which passes through a grid wire is followed. The same thing applies between plane B and the plate.

Between planes A and B, however, conditions become vastly different according to whether the potential curve is drawn for the line 1–2 or the line 3–4, and the general shape of the potential curve for the two conditions is marked 1–2 and 3–4 respectively in the lower diagram. Along 1–2 the potential is everywhere positive, as otherwise electrons would not be able to penetrate the grid mesh and reach the plate. On the other hand, the grid wires themselves are at a negative potential, and the curve 3–4 shows the way in which the potential surface forms into pockets surrounding the grid wires. The size of these pockets, and hence the location of the planes A and B is determined by the

relative potentials of grid and plate, the size of the grid wires, and the spacing between them.

In applying the general analysis for parallel planes to this type of structure it is possible to find all of the desired initial conditions at the B-plane in terms of conditions at the A-plane as well as to find the actual a.-c. potential of the grid wires themselves and the relation between plate current and cathode current provided that a very simple condition is fulfilled. This condition is merely that the size of the potential pockets surrounding the grid wires is so small that the planes A and B may be taken close enough together to cause the electron transit time between the two to be negligibly small compared with the transit time from cathode to A or from B to the anode. Without this condition, the problem appears almost hopeless. With it, the procedure is straightforward and simple.

For the factors affecting the fulfillment of this condition, Maxwell's analysis fortunately provides a guide that is at least safe, because if the planes can be located near together compared with the distances x_c and x_p on the basis of his analysis, then the smoothing effect of the electrons between grid wires will make the actual operating conditions still better. The adaptation of his equations gives the following expression for the potential distribution and for small diameter grid wires:

$$V = \left[\frac{V_g - \frac{x_c}{x_c + x_p} V_p}{\left(\frac{x_c x_p}{x_c + x_p} \right) \frac{4\pi}{a} - 2 \log_e \left(2 \sin \frac{\pi c}{a} \right)} \right] \left[\frac{4\pi x_p}{a} \left(\frac{x_c + d}{x_c + x_p} \right) - \log_e (1 - 2e^{2\pi d/a} \cos^{2\pi z/a} + e^{4\pi d/a}) \right] + V_p \left(\frac{x_c + d}{x_c + x_p} \right). \quad (53)$$

Here d is the distance from the grid to the place where the potential is to be computed, and is considered positive when directed from the grid toward the plate, z is distance measured parallel to the grid, a is the distance between centers of grid wires, c is the wire radius, and the cathode potential, V_c is taken to be zero. The only term here contributing to the potential pockets surrounding grid wires is the logarithmic one which alone involves z, the distance along plane A or B. The condition for the variations to be smoothed out with either positive or negative values of d is

$$e^{2\pi D/a} \gg 2,$$

where D represents the magnitude of d. Below are given some of the values of $e^{2\pi D/a}$

D/a	$e^{2\pi D/a}$
0.1	1.875
0.2	3.514
0.3	6.587
0.4	12.35
0.5	23.17
0.6	43.39
0.7	81.34
0.8	152.5
0.9	285.5
1.0	536.0

This shows that if the magnitude of the cosine term which involves z is to be limited to one per cent of the largest of the three terms constituting the logarithm, then D/a should be greater than 0.84. On the other hand if a ten per cent variation is tolerated, then D/a may be as small as 0.48.

There is a further consideration that tends to smooth out the corrugations in the potential caused by the grid wires. The relation (53) was derived on the assumption that the cathode and plate were quite distant from the grid. When this is not the case, the fact that both are equipotential surfaces tends to crowd the irregularities in toward the grid, and hence to decrease the area of the pockets.

The general conclusion to be drawn from this investigation into the potential distribution is that the electronics analysis can be applied with greatest accuracy to tubes where the grid wires are very close together relative to the distance between grid and either cathode or anode. When this is not the case, the analysis may be expected to show the general trend of the performance in most cases but to be unreliable for quantitative computation. In extreme cases, where the grid wires are very far apart, the grid action approximates more nearly to a change in the effective cathode area than a uniform action over the entire surface. The electronics analysis can be adapted to this extreme case when the change in effective cathode area occurs instantaneously with a change in grid potential, for then the diode analysis applies directly between the effective cathode area and the anode. Means for doing this will become evident when the theory for grids with fine mesh is understood so that the details will not be described.

The analysis for fine-mesh grids now follows, based on the assump-

tion applicable to them that the planes A and B in Fig. 1 are very near together compared with x_c and x_p. As a first consequence of this assumption, the electron velocities at B must be the same as those at A, and hence one of the initial conditions at the B-plane is provided for. A second consequence is that the potential at B is the same as that at A, and therefore the potential between plate and cathode is the sum of the potentials between cathode and A and between B and the plate.

The accelerations at the two planes are not the same. This can be seen from the lower diagram in Fig. 1 when it is remembered that the acceleration is proportional to the slope of the potential curve. The accelerations can be found, however, by a relatively simple calculation and this will be done in the course of the following analysis.

D.–C. Relations

As a preliminary to the treatment of first order effects in negative grid triodes, certain d.-c. relations will be determined.

The distance is related to the transit time by (10). At the cathode and for complete space charge the acceleration and velocity are zero, so that the cathode-grid transit time T_c is given by

$$x_c = K\frac{T_c^3}{6}. \tag{54}$$

At B the acceleration is yet to be found, but may be taken to be g times that at A. Hence, from (21)

$$a_B = gKT_c. \tag{55}$$

The velocity at B is the same as that at A, so that from (27)

$$u_B = K\frac{T_c^2}{2}. \tag{56}$$

These values, (55) and (56), may now be inserted as initial conditions in (10) to give a relation between x_p and T_p. Calling the ratio T_p/T_c by the symbol h and x_p/x_c by the symbol y we thus obtain:

$$gh^2 = y/3 - h - h^3/3. \tag{57}$$

In this way the acceleration is related to the transit times. These have now to be expressed in terms of quantities that can be measured directly, namely the current and the plate voltage. To do this, the general expression for potential difference is obtained by integration

of (34) which gives:

$$(W_a - W_b)_0 = \frac{1}{8} K^2 T^4 + \frac{1}{2} K a_a T^3 + \frac{1}{2} K u_a T^2 + \frac{1}{2} a_a^2 T^2 + a_a u_a T. \quad (58)$$

Putting in the initial conditions we have

$$(W_c - W_A)_0 = \frac{1}{8} K^2 T_c^4 \quad (59)$$

and

$$(W_B - W_p)_0 = \frac{1}{8} K^2 T_c^4 [4gh(1 + h^2) + 4g^2 h^2 + 2h^2 + h^4]. \quad (60)$$

From these, and placing the cathode potential equal to zero, we get

$$\sqrt{\frac{V_p}{V_A}} = 2gh + 1 + h^2.$$

With (57) the ratio g may be eliminated from this, giving

$$\sqrt{\frac{V_p}{V_A}} = \frac{2y}{3h} - 1 + \frac{1}{3} h^2. \quad (61)$$

When h is small, as it normally is, the last term may be disregarded, giving

$$h = \frac{2/3\ y}{1 + \sqrt{V_p/V_A}} \text{ (approximately).} \quad (61a)$$

Equations (57) and (61) allow g and h to be found in terms of V_p, which can be measured directly, and V_A which can be found in terms of the current and the distance x_c by combining (54) and (59) to give:

$$(W_c - W_A)_0 = \frac{1}{8} K^{2/3} (6x_c)^{4/3}. \quad (62)$$

This is the well-known Child's equation, and when W and K are expressed in terms of voltage and current appears in the usual form

$$I_0 = -2.34 \times 10^{-6} V_A^{3/2}/x_c^2 \text{ amperes/cm.}^2 \quad (63)^*$$

Several other expressions which are useful for computation purposes may be found from these d.-c. relations:

* The negative sign occurs because of the assumed current direction, from the cathode. The numerical factor is 2.34 instead of 2.33 given by Compton and Langmuir [6] because of the value of e/m which was used, q.v.

The transit angle $\theta_c = \omega T_c$ may be written in terms either of the voltage V_A or the current I_0, thus

$$\theta_c = \frac{9500 x_c}{\lambda \sqrt{V_A}} = \frac{-126}{\lambda} \left(\frac{x_c}{I_0} \right)^{1/3} \text{radians}, \tag{64}$$

where λ is the free-space wave-length in centimeters of an alternating current of angular frequency ω.

The slope of the static characteristic of a diode coinciding with the cathode and the plane A may be expressed

$$-\left. \frac{\partial V_A}{\partial I_0} \right|_{x_c} = r_c = -\frac{2}{3} \frac{V_A}{I_0} = \frac{285{,}000 x_c^2}{\sqrt{V_A}} = \frac{-3780 x_c^{4/3}}{I_0^{1/3}}, \tag{65}$$

where r_c is the low-frequency resistance in ohms of a square centimeter of area.

From (64) and (65) it can be seen that

$$r_c = 30 \lambda x_c \theta_c. \tag{66}$$

A further expression that occurs frequently in following equations is:

$$\frac{K T_c^4}{\epsilon} = 12 r_c. \tag{67}$$

Later we shall be able to show how the low-frequency plate resistance of the triode is related to r_c and the amplification factor, as well as how the inter-electrode capacitances of the "cold" tube are involved in these quantities. A simple approximation for the transit time ratio h will also be derived.

First-Order Relations

In the picture shown by Fig. 1 an alternating current is assumed to flow from the cathode to the plane A. This current I_c is related to the quantity J in the general equations (41) or (41a) by the expression $J_c = eI_c/km\epsilon$, and the current includes both conduction and displacement components. Complete space charge at the cathode allows initial velocities and accelerations to be placed equal to zero, so that the first order potential difference between cathode and plane A is given by (41a) as follows:

$$(W_c - W_A)_1 = J_c K T_c^4 \sum_{n=0}^{\infty} (-\beta_c)^n \frac{(n+2)}{(n+4)!}. \tag{68}$$

In a similar way, the potential between plane B and the plate can be written when the initial first-order acceleration at B has been found,

the d.-c. initial conditions being given in the previous section, and the first order initial velocity being the same as that at A. From (43a) we have

$$\mu_B = \mu_A = J_c T_c^2 \sum_{m=0}^{\infty} \frac{(-\beta_c)^m (m+1)}{(m+3)!} . \tag{69}$$

The computation of the a.-c. component of acceleration at B is based on the particular property possessed by negative grid tubes that no electrons reach the negative grid. Then, because the velocity at B is the same as that at A it follows that the conduction component of the current is the same at both planes. The total first-order current may be written from (1)

$$I_1 = (\rho u)_1 + \epsilon \frac{\partial E_1}{\partial t} .$$

Multiplying through by $e/km\epsilon$ we can write

$$J = Q + \frac{\partial a_1}{dt}, \tag{70}$$

where $Q = (\rho u)_1 e/km\epsilon$ measures the conduction current, and is the same at planes A and B. Hence at plane A, and since $\partial/\partial t = p$ for exponential currents and voltages:

$$J_c = Q + p a_{1A}. \tag{71}$$

At plane B we have similarly

$$J_p = Q + p\alpha_B. \tag{72}$$

From (71) and (72) there results

$$p\alpha_B = J_p - J_c + p a_{1A}. \tag{73}$$

The value of a_{1A} may be obtained from (42a) so that the acceleration at B is given by

$$p\alpha_B = J_p - J_c - 2J_c \sum_{m=0}^{\infty} \frac{(-\beta_c)^{m+1}(m+2)}{(m+3)!} . \tag{74}$$

All of the initial values at the plane B have been obtained and are expressed in (55), (56), (69) and (74). The potential between B and the plate is now obtained from (41a) and follows, where $h = T_p/T_c$:

$$\frac{(W_B - W_p)_1}{KT_c^4} = (J_p - J_c)\frac{h(gh+1)}{2\beta_c} + J_p h^4 \sum_{n=0}^{\infty} \frac{(-h\beta_c)^n (n+2)}{(n+4)!}$$

$$+ J_c\left[h^2 \sum_{n=0}^{\infty} (- h\beta_c)^n \left(\frac{2gh(n+2) + (n+3)}{2(n+3)!} \right) \right.$$
$$+ h \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} h^n (- \beta_c)^{n+m} \qquad (75)$$
$$\left. \times \left(\frac{2gh(n+1)(m+2) + (n+2)(m+2) - h(n+1)(m+1)}{(n+2)!(m+3)!} \right) \right].$$

The attainment of (68) and (75) does not completely solve the problem because it is the potential of the grid wires that can be measured rather than the potential at A and B. The transformation may be readily accomplished, however, by writing

$$(V_A - V_g)_1 = I_g Z_g, \qquad (76)$$

where Z_g is the effective impedance between the plane A (or B) and the grid wires. In the negative grid tube when A and B are close together, Z_g is a pure capacitance, C_g.

Writing (68) and (75) respectively in terms of V and I instead of W and J we now have in symbolic form

$$(V_c - V_A)_1 = I_c Z_c, \qquad (77)$$
$$(V_A - V_p)_1 = I_c(Z_3 - Z_1) + I_p(Z_1 + Z_2), \qquad (78)$$

where

$$I_c = I_g + I_p. \qquad (79)$$

The four equations, (76), (77), (78) and (79) are the basis of negative grid triode analysis. The impedances involved refer to a square centimeter of area and are as follows:

$$Z_g = \frac{1}{pC_g} = \text{impedance between plane } A \text{ (or } B\text{) and the grid wires,} \qquad (80)$$

$$Z_c = 12 r_c \sum_{n=0}^{\infty} (- \beta_c)^n (n+2)/(n+4)!, \qquad (81)$$

$$Z_1 = 12 r_c h(gh + 1)/2\beta_c, \qquad (82)$$

$$Z_2 = 12 r_c h^4 \sum_{n=0}^{\infty} (- h\beta_c)^n (n+2)/(n+4)!, \qquad (83)$$

$$Z_3 = 12 r_c h^2 \left[\sum_{n=0}^{\infty} (- h\beta_c)^n \left(\frac{2gh(n+2) + (n+3)}{2(n+3)!} \right) \right]$$
$$+ 12 r_c h \left[\sum_{n=0}^{\infty} \sum_{m=0}^{\infty} (- \beta_c)^{n+m} h^n \right.$$
$$\left. \times \left(\frac{2gh(n+1)(m+2) + (n+2)(m+2) - h(n+1)(m+1)}{(n+2)!(m+3)!} \right) \right]. \qquad (84)$$

The value of r_c is expressed by (65).

For large values of transit angle it is convenient to have these impedances expressed in closed instead of in series form. By using (41) instead of (41a) throughout the analysis the result may be obtained. The expressions for Z_g and Z_1 will not differ from those given above, but Z_c, Z_2 and Z_3 may be written in the following forms:

$$Z_c = \frac{12r_c}{\beta_c^4}\left[\frac{\beta_c^3}{6} - \beta_c(e^{-\beta_c}+1) - 2(e^{-\beta_c}-1)\right], \tag{81a}$$

$$Z_2 = \frac{12r_c}{\beta_c^4}\left[\frac{h^3\beta_c^3}{6} - h\beta_c(e^{-h\beta_c}+1) - 2(e^{-h\beta_c}-1)\right], \tag{83a}$$

$$\begin{aligned} Z_3 = \frac{12r_c}{\beta_c^4}\Bigg[& \beta_c^3\frac{1}{2}(h+gh^2) \\ & + \beta_c[(h-2gh-1)e^{-(h+1)\beta_c} + he^{-h\beta_c} + e^{-\beta_c}] \\ & + [2(h-gh-g)e^{-(h+1)\beta_c} + 2(1-h+gh)e^{-h\beta_c} + 2ge^{-\beta_c} - 2] \\ & + \frac{2}{\beta_c}(1-g)(e^{-(h+1)\beta_c} - e^{-h\beta_c} - e^{-\beta_c} + 1)\Bigg]. \end{aligned} \tag{84a}$$

Of the various impedances Z_3 is the only one that is really troublesome. The values of Z_c and Z_2 may be obtained from data given in published papers [1, 2] when it is noticed that Z_2/h^4 may be calculated from Z_c if β_c is replaced by $\beta_p = h\beta_c$. In the treatment of Z_3 there seems to be no easy road, although the series form (84) may be expanded with comparative ease.

Further steps consist in the transposition of these equations to obtain convenient forms and to show how they harmonize with low-frequency theory. A useful expression may be obtained from the fundamental relations (76), (77), (78), (79) by eliminating V_A, I_c and I_g. The result is

$$\begin{aligned} (V_c - V_p)_1 + \frac{(Z_1 - Z_c - Z_3)}{Z_c + Z_g}(V_c - V_g)_1 \\ = I_p\left[\frac{[(Z_c + Z_2 + Z_3)Z_g + (Z_1 + Z_2)Z_c]}{Z_c + Z_g}\right]. \end{aligned} \tag{85}$$

This begins to look like the familiar low-frequency equation

$$(V_c - V_p)_1 + \mu(V_c - V_g)_1 = I_pZ_p, \tag{86}$$

where μ is the amplification factor and Z_p is the internal plate impedance. The two are equivalent at all frequencies if I_p is interpreted as the density of the total plate current, and not the conduction com-

ponent, only. Then we have:

$$\mu = \frac{Z_1 - Z_c - Z_3}{Z_c + Z_g} = \frac{(Z_1 + Z_2) - (Z_c + Z_2 + Z_3)}{Z_c + Z_g}, \tag{87}$$

$$Z_p = \frac{(Z_c + Z_2 + Z_3)Z_g + (Z_1 + Z_2)Z_c}{Z_c + Z_g}. \tag{88}$$

The relations involved in these two equations may be made somewhat clearer by finding what they resolve into at low frequencies. This may be done by going to (80)–(84). The first thing to notice is that Z_1 and Z_g become very large because the frequency term $p = i\omega$ appears in their denominators. The other terms are relatively small at low frequencies and we have:

$$\mu_0 \rightarrow Z_1/Z_g, \tag{87a}$$

$$r_p = Z_p \rightarrow Z_c(1 + \mu_0) + Z_2 + Z_3. \tag{88a}$$

It will be shown below that this formulation for the amplification factor is in accord with that derived by Maxwell in his "Treatise on Electricity and Magnetism" for the shielding effect of a grid mesh.

Low-Frequency Relations in Negative Grid Triodes

In (87a) and (88a) the general form of the low-frequency triode relations is given. It is instructive to compute these in some detail so that the role played by the capacitance C_g between the planes A or B and the grid wires is demonstrated.

To do this, (82) which gives the impedance Z_1 may first be transformed by aid of (57) and (67) to give

$$Z_1 = \frac{x_p}{\epsilon p}(1 - h^3/y). \tag{89}$$

The impedance Z_g may be written as in (80) so that the low-frequency amplification factor is

$$\mu_0 = \frac{x_p}{\epsilon} C_g(1 - h^3/y). \tag{90}$$

It is of interest to note that ϵ/x_p is the capacitance per unit area between the plate and a solid plane at the grid. As expressed by Compton and Langmuir [6] from Maxwell's analysis the low-frequency amplification factor is

$$\mu_0' = \frac{x_p}{\frac{a}{2\pi}\log_e \frac{a}{2\pi c}}, \tag{91}$$

where a is the distance between centers of grid wires and c is the wire radius.

Comparing (90) and (91) we see that both are proportional to x_p, but that (90) contains a correction term, h^3/y which is normally very small as h is usually of the order of one-third to one-tenth. The presence of the correction term may be explained by remembering that (90) was derived for conditions holding when electrons are flowing while (91) applies strictly only to a cold tube in which no free electrons are present. This being the case, it is to be expected that the two equations would be equivalent if the correction term were omitted from (90). The term being small, this is very nearly true in any case and gives

$$C_g = \frac{2\pi\epsilon}{a \log_e \frac{a}{2\pi c}} \text{ farads in cm.}^2 \tag{92}$$

The capacitance C_g which was introduced as existing between the electron stream and the grid wires is thus shown to have a value which can be calculated with fair exactness under the conditions when Maxwell's equation (91) holds, namely when the grid wires are small compared with their separation.

Attention is now directed to the low-frequency value of the plate impedance, Z_p. From (88a) together with (81), (83) and (84) it may be shown that

$$r_p = r_c \left[\mu_0 + \frac{4}{3}(1 + y)(1 + h) - \frac{1}{3}(1 + h)^4 \right]. \tag{93}$$

This is thought to be the first instance of an expression for the plate resistance derived on strictly theoretical grounds. The formula as it stands contains r_c which is given by (65) and h which is given by (61), and both involve V_A. This latter may be found from (63) in terms of the direct current. A convenient approximation for V_A is obtained by making the assumption that the presence of electrons changes the d.-c. potential V_A by a small amount only, so that its value may be calculated from the static capacitances of a cold tube. The appropriate diagram is shown on Fig. 2.

From the figure, putting $V_c = 0$ we obtain

$$V_{A0} = \frac{V_{p0} + \frac{C_g}{C_p} V_{g0}}{1 + \frac{C_g}{C_p} + \frac{C_c}{C_p}}.$$

It was shown in (90) that C_g/C_p is the static, or "cut-off" amplification factor, μ_0' of the tube. Again, $y = C_c/C_p$ so that

$$V_{A0} = \frac{V_{p0} + \mu_0' V_{g0}}{1 + \mu_0' + y} \text{ (approximately).}$$

When the cathode is heated so that electrons flow, the potential V_A becomes depressed below this value which should be used only in forming rough estimates.

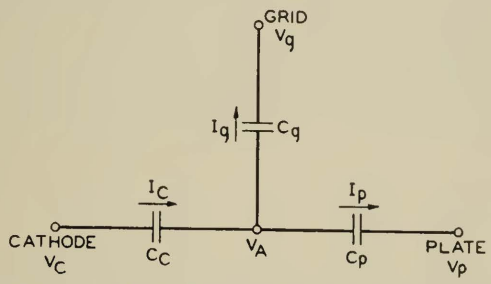


Fig. 2.—Equivalent network of negative grid triode in the absence of electrons.

A more accurate formula for V_A involves the transit time ratio h, but may be obtained as follows: The plate current of an ideal negative grid triode is proportional to $(V_{p0} + \mu V_{g0})^{3/2}$ so that

$$r_p = \frac{2}{3}\frac{(V_{p0} + \mu_0 V_{g0})}{I_0}.$$

Similarly

$$r_c = \frac{2}{3}\frac{V_{A0}}{I_0}.$$

From these two equations together with (93) is obtained

$$V_{A0} = \frac{V_{p0} + \mu_0 V_{g0}}{\mu_0 + \frac{4}{3}(1 + y)(1 + h) - \frac{1}{3}(1 + h)^4},$$

which may also be written

$$V_{A0} = \frac{V_{p0} + \mu_0 V_{g0}}{1 + \mu_0 + \frac{4}{3}y(1 + h) - \frac{1}{3}h^2(6 + 4h + h^2)}. \tag{94}$$

This is in a form for comparison with the approximate equation above and shows the modification produced by the presence of electrons.

General Relations in Negative Grid Triodes

The high-frequency values of the amplification factor and the plate impedance could be computed in detail from (87) and (88) together with the various expressions given for the impedances involved. To do this would require an enormous amount of computation, so that the details are deferred until such time as it becomes evident that (87) and (88) express the high-frequency properties of tubes in the most useful way. The fact that they are analogous to the ordinary low-frequency conventions does not at all assure their general utility in the high-frequency field, and the fundamental equations (76) to (79) may be arranged in a wide variety of forms. For example, a companion equation to (85) may be obtained from the fundamental equations by eliminating V_A, I_p and I_c and thus obtaining:

$$(V_c - V_g)_1 - \left(\frac{Z_c}{Z_2 + Z_3 + Z_c}\right)(V_c - V_p)_1 = I_g\left[\frac{(Z_2 + Z_3 + Z_c)Z_g + (Z_1 + Z_2)Z_c}{Z_2 + Z_3 + Z_c}\right]. \quad (95)$$

Just as (85) gives an equivalent circuit between cathode and plate involving the whole current reaching the plate, so does (95) give an equivalent circuit between cathode and grid involving the whole current reaching the grid. The two equations completely describe the tube performance when the external connections are known. Because of the way in which I_g and I_p are defined they include the so-called grid-plate path, which is treated as a separate circuit at low frequencies.

The impedance presented by a tube to an e.m.f. applied between cathode and grid may now be calculated. To carry out the computation in general, it would be necessary to know the impedance attached between plate and cathode in the external circuit so that $(V_c - V_p)_1$ could be obtained from (85). However, the high-frequency properties of negative grid tubes may be illustrated more directly by choosing for consideration a special case that avoids having to take this additional step involving (85). This special case is the one where such a large capacitance is connected between the plate and cathode that $(V_c - V_p)_1$ is zero for any of the frequencies to be considered. The result will therefore be particularly applicable to finding the input impedance of screen tubes where the requirement for the special case is fulfilled.

For this special case where $(V_c - V_p)_1$ is zero, (95) may be solved

directly for the impedance Z_a presented to an input applied between grid and cathode and gives

$$Z_a = Z_g + \frac{(Z_1 + Z_2)Z_c}{Z_2 + Z_3 + Z_c}. \tag{96}$$

The impedance Z_g is a pure capacitance, but the second term on the right of (96) contains both reactive and resistive components which latter account for the active grid loss which has been the subject of several investigations both of a theoretical and experimental nature.[2, 7, 8] At very low frequencies the capacitance represented by Z_1 predominates and the resistive component vanishes leaving for the input impedance Z_a merely the following

$$Z_a \to Z_g + Z_1\left(\frac{Z_c}{Z_2 + Z_3 + Z_c}\right) = \frac{(y - h^3)}{i\omega\mu_0 C_c}\left[\frac{r_p/r_c}{\frac{4}{3}(1 + y)(1 + h) - \frac{1}{3}(1 + h)^4}\right]. \tag{97}$$

This expression may be written in several different forms but in none of them does a simplification occur in the way in which the transit time ratio h enters the equation. Perhaps the best mode of expression is a comparison of the "hot" capacitance C_a given by (97) where $Z_a = 1/i\omega C_a$ with the capacitance C_0 of a cold tube with plate and cathode tied together. This latter may be written

$$C_0 = \frac{\mu_0' C_p(1 + y)}{1 + y + \mu_0'}, \tag{98}$$

where μ_0' is the "cold" amplification factor, (91), and is related to μ_0 as shown by (90) so that $\mu_0 = \mu_0'(1 - h^3/y)$. The ratio C_a/C_0 is the "dielectric constant" of the hot tube and is

$$\frac{C_a}{C_0} = \left(\frac{1 + y + \mu_0'}{1 + y}\right)\left[\frac{\frac{4}{3}(1 + y)(1 + h) - \frac{1}{3}(1 + h)^4}{\mu_0 + \frac{4}{3}(1 + y)(1 + h) - \frac{1}{3}(1 + h^4)}\right]. \tag{99}$$

For illustration suppose that a certain tube has the following values: $y = 1$, $\mu_0 = 10$, $h = 1/5$. Then from (99) the dielectric constant of the hot tube would be 1.19. This is somewhat less than the value of

4/3 obtained by Benham [1] in an analysis which considered effects between cathode and grid only. Benham's value applies strictly to the capacitance between cathode and the plane A in Fig. 1 while (99) applies to the parallel combination of cathode-grid and grid-plate capacitances. The constants of the individual capacitances may be calculated from the fundamental equations (76)–(79) but are omitted here because of space limitations. A series of experiments performed several years ago by Mr. A. J. Rack and the writer and covering a wide range of operating conditions with several vacuum tubes showed values greater than unity for the dielectric constant of the cathode-grid capacitance and less than unity for the grid-plate capacitance. The parallel combination had a dielectric constant which was greater than unity in the range investigated, being thus in accord with (99) which always gives constants greater than unity for normal values of h.

The input capacitance of detector- and voltmeter-tubes being thus a somewhat complicated function, it is to be expected that the calculation of the impedance at higher frequencies where transit times are appreciable will be similarly complicated. To avoid undue length only the first term contributing to the resistive component of the active grid loss will be computed in detail. It must be pointed out, however, that the series in powers of transit angle which represents the input impedance converges slowly so that the first term is useful only when the transit time is small.

Keeping this in mind, we go to (96) and write the impedance in series form, obtaining finally the following expression for the equivalent shunting resistance between cathode and grid:

$$\frac{1}{R_a} = \frac{\theta_c^2}{180}\frac{\mu_0}{r_p}\left[\frac{y^2A - yB + C}{(y - h^3)^2}\right] \times \left[\frac{\mu_0}{\mu_0 + \frac{4}{3}(1 + y)(1 + h) - \frac{1}{3}(1 + h)^4}\right], \quad (100)$$

where

$$A = 9 + 44h + 45h^2,$$
$$B = 51h^2 + 123h^3 + 55h^4 + 3h^5,$$
$$C = 45h^4 + 51h^5 + 24h^6 + 11h^7 + 3h^8.$$

As in previous cases, this may be written in several ways, depending on the mode of expression of θ_c and r_p. For example

$$\theta_c = 2r_c\omega C_c,$$

so that

$$\frac{1}{R_a} = \frac{(\omega C_c)^2}{45}\,\frac{r_p}{\mu_0}\left[\frac{y^2A - yB + C}{(y - h^3)^2}\right] \times \left[\frac{\mu_0}{\mu_0 + \frac{4}{3}(1 + y)(1 + h) - \frac{1}{3}(1 + h)^4}\right]^3. \quad (100a)$$

Comparison of (100) and (100a) shows that the transconductance μ_0/r_p appears in the numerator of the former, but in the denominator of the latter, and illustrates the care that must be taken in deriving sweeping conclusions concerning the effect of various tube parameters without taking all of the contributing factors into consideration. In the case of (100) the conclusion is that the loss may be reduced by decreasing the transconductance, but only if the transit angle θ_c is unchanged. On the other hand, (100a) says that the loss may be reduced by increasing the transconductance, but only if this is accomplished without change in the cathode-grid spacing, and without altering h by an amount large enough to affect materially the factors in square brackets.

Experimentally, it is found in many tubes that the loss increases when the transconductance is increased by changing the voltages applied to a given tube. This would seem to be at variance with (100a), for the cathode-grid spacing, and hence C_c, has not been altered by the voltage change. The explanation of the difficulty apparently lies in the departure of the static characteristics of many tubes from the 3/2 power law, which again may be explained in part by the presence of initial velocities and the large size of the potential pockets surrounding the grid wires. In a rough way the action of the latter is to vary the effective cathode area when the voltages are changed, producing an increased area with increase of current, and hence producing a current variation greater than the 3/2 power law.

In a recent paper, D. O. North [11] derives a formula for the active grid loss by neglecting space charge between grid and plate. His result is similar in many respects to (100) and both contain the factors $\theta_c^2\mu_0/r_p$. In an experimental check, W. R. Ferris [10] secures excellent results by obtaining the transconductance from the static characteristics of the tubes used, but computing θ_c by a formula which differs only slightly from (94). It can be shown that such a procedure would give a computed loss which increases with transconductance when the static characteristic is of the form $I = K(V_{p0} + \mu_0 V_{g0})^n$ and when n is greater than 2. The static characteristics are not given in Mr.

Ferris' paper so that it is not evident whether the exponent is greater than 2. The loss did, however, increase with transconductance and was checked by the computations in a satisfactory manner, which would imply either that the exponents were actually greater than 2 in Mr. Ferris' tubes, or that their cylindrical shape caused a decrease in the effect of the transconductance on the transit angle.

The equations in general indicate that the shunting resistance between cathode and grid is proportional to the square of the wave-length. This is in accord with the theory and experiments of Thompson and Ferris,[7, 10] and with the experiments of J. G. Chaffee [8] on tubes biased as class *A* amplifiers. However, when the tubes were biased as detectors, Chaffee [8, 9] found that the resistance varied more nearly as the first power of the wave-length. There are several factors which may contribute to this difference. With detector bias near cut-off the transit angles are large so that more terms of the fundamental equations may be needed. In Chaffee's work these were computed to be of the order of two or three radians which was scarcely enough to cause the entire effect observed by him. Another cause is thought to be an actual reversal in the direction of motion of electrons caused by the alternating potential operating in the vicinity of cut-off. Further study both of experimental and theoretical nature is required, however, before the point can be considered to be satisfactorily explained.

Conclusion

In general, the analysis presented in the foregoing pages is capable of serving as a guide to indicate the kind of results to be expected in the operation of vacuum tubes at ultra-high frequencies. In those cases where the physical structure of the tube complies with the conditions laid down for the theoretical treatment, a quantitative agreement can be anticipated. The importance of departures of the physical structure from this ideal can be evaluated in many instances by a careful comparison of the actual with the ideal structure.

Much yet remains to be done in the way of computing and tabulating the various factors involved in the equations and of investigating the effects of such things as initial electron velocities from a hot cathode, large size grid wires, coarse mesh grids, and cylindrical structures.

References

1. W. E. Benham, "Theory of the Internal Action of Thermionic Systems at Moderately High Frequencies." *Phil. Mag.*, p. 641, March 1928. Part II, *Phil. Mag. Supp.*, Vol. 11, p. 457, February 1931.
2. F. B. Llewellyn, "Vacuum Tube Electronics at Ultra-High Frequencies." *Proc. I.R.E.*, Vol. 21, November 1933; also *B.S.T.J.*, Vol. XIII, January 1934.

3. Johannes Müller, "Elektronenschwingungen im Hochvakuum." *Hochfrequenztechnik und Elektroakustik*, Vol. 41, May 1933.
4. F. B. Llewellyn, "Note on Ultra-High-Frequency Electronics." *Proc. I.R.E.*, Vol. 23, February 1935.
5. W. E. Benham, "Electronic Theory and the Magnetron Oscillator." *Proc. Phys. Soc.*, Vol. 47, January 1935.
6. Compton and Langmuir, "Electrical Discharges in Gases," Part II. *Rev. of Mod. Physics*, Vol. 3, April 1931.
7. B. J. Thompson and W. R. Ferris, Oral paper at Rochester I.R.E., 1934.
8. J. G. Chaffee—Forthcoming discussion on (7).
9. J. G. Chaffee, "The Determination of Dielectric Properties at Very High Frequencies." *Proc. I.R.E.*, Vol. 22, August 1934.
10. W. R. Ferris, "Input Resistance of Vacuum Tubes as Ultra-High Frequency Amplifiers." *
11. D. O. North, "Analysis of the Effects of Space Charge on Grid Impedance." *

* To be published shortly. Much of the material was presented orally by B. J. Thompson and W. R. Ferris at various technical meetings including:
U. R. S. I. and *I. R. E.*, Washington, April 27, 1934.
Philadelphia section of *I. R. E.*, January 3, 1935.
Reference (7) above.

Further Extensions of the Theory of Multi-Electrode Vacuum Tube Circuits

By S. A. LEVIN and LISS C. PETERSON

The response of circuits containing vacuum tubes with any number of electrodes due to impressed electromotive forces, and under such circumstances that the time of transit of the electrons is negligible, is discussed when arbitrary feedback is present between the circuits connected to the electrodes, each of which may carry conductive current. The use of the theory is illustrated by obtaining first and second order effects in typical three-electrode tube circuits.

In a previous paper in the *Bell System Technical Journal*, October, 1934, the treatment was restricted to three-electrode tube circuits in which it was assumed that the amplification factor of the tube was constant and that no conductive grid current was present. In the present paper these restrictions are removed.

Introduction

THE response in multi-electrode vacuum tube circuits due to impressed electromotive forces has been the subject of several papers. For the three-electrode vacuum tube circuit J. R. Carson [1] has used a method of successive approximations, assuming constant amplification factor and no conductive grid current. E. Peterson and H. P. Evans [2] removed the restriction on the amplification factor but maintained the assumption regarding the grid current, while F. B. Llewellyn [3] considered the general case with both plate and grid currents. Finally, J. G. Brainerd [4] has treated the general case of the four-electrode tube circuit. The theories given by these authors did not take into account any feedback between the circuits of the electrodes except in the first approximation.

In a previous paper [5] the theory given by Carson was extended to include the effects of feedback between plate and grid circuits not only in the first but also in the second and higher approximations. The aim of the present paper is to extend similarly the other theoretical work mentioned above[2, 3, 4] to circuits containing tubes with three, four, or any number of electrodes.

Theory of Three-Electrode Tube Circuits

We shall consider the three-electrode tube circuit shown in Fig. 1 where Z_1, Z_2, and Z_3 are impedances which may include inter-electrode

[1] J. R. Carson: *I. R. E. Proc.*, April, 1919, p. 187.
[2] E. Peterson and H. P. Evans: *B. S. T. J.*, July, 1927, p. 442.
[3] F. B. Llewellyn: *B. S. T. J.*, July, 1926, p. 433.
[4] J. G. Brainerd: *I. R. E. Proc.*, June, 1929, p. 1006.
[5] S. A. Levin and Liss C. Peterson: *B. S. T. J.*, October, 1934, p. 523.

admittances. The impressed variable electromotive forces are ϵ_p and ϵ_g in series with the impedances Z_p and Z_g, respectively. We will designate by E_p and I_p the total plate voltage and current, respectively, while the corresponding quantities for the grid are E_g and I_g. In the

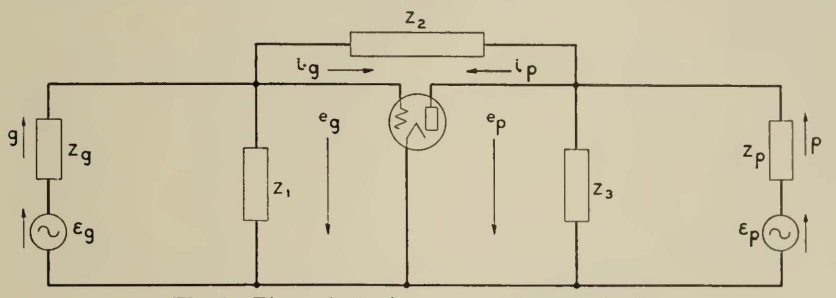


Fig. 1—Three-electrode vacuum tube and circuit.

absence of the variable electromotive forces the d-c. values of these voltages and currents are E_{p0}, I_{p0}, E_{g0}, and I_{g0}, respectively, while the increments due to the impressed forces are e_p, i_p, e_g, and i_g, respectively. Similarly, g and p denote the incremental voltages across Z_g and Z_p. All quantities referring to currents and voltages are instantaneous values.

We will now assume that I_p and I_g are functions of E_p and E_g, and that we can derive from these functions the expansions

$$i_p = \sum b_{mn} e_p{}^m e_g{}^n, \qquad i_g = \sum \beta_{mn} e_p{}^m e_g{}^n, \tag{1}$$

where

$$b_{mn} = \frac{1}{m!n!} \frac{\partial^{(m+n)} I_p}{\partial E_p{}^m \partial E_g{}^n}, \qquad \beta_{mn} = \frac{1}{m!n!} \frac{\partial^{(m+n)} I_g}{\partial E_p{}^m \partial E_g{}^n}, \tag{2}$$

evaluated at the operating point (E_{p0}, E_{g0}).

The important tube parameters are by definition

$$\left.\begin{aligned} \frac{1}{r_p} &= \frac{\partial I_p}{\partial E_p}, \qquad \mu_p = \frac{\dfrac{\partial I_p}{\partial E_g}}{\dfrac{\partial I_p}{\partial E_p}} = -\left(\frac{dE_p}{dE_g}\right)_{I_p=\text{const.}} \qquad S_{pg} = \frac{\partial I_p}{\partial E_g} = \frac{\mu_p}{r_p} \\ \frac{1}{r_g} &= \frac{\partial I_g}{\partial E_g}, \qquad \mu_g = \frac{\dfrac{\partial I_g}{\partial E_p}}{\dfrac{\partial I_g}{\partial E_g}} = -\left(\frac{dE_g}{dE_p}\right)_{I_g=\text{const.}} \qquad S_{gp} = \frac{\partial I_g}{\partial E_p} = \frac{\mu_g}{r_g} \end{aligned}\right\}, \tag{3}$$

where r denotes electrode resistances, μ the mu-factors, and S the transconductances.

It follows readily from (2) and (3) that

$$\left.\begin{aligned} b_{10} &= \frac{1}{r_p} & \beta_{10} &= \frac{\mu_g}{r_g} \\ b_{01} &= \frac{\mu_p}{r_p} & \beta_{01} &= \frac{1}{r_g} \\ b_{20} &= -\frac{1}{2r_p^2}\frac{\partial r_p}{\partial E_p} = P_2 & \beta_{02} &= -\frac{1}{2r_g^2}\frac{\partial r_g}{\partial E_g} = T_2 \\ b_{11} &= \frac{1}{r_p}\frac{\partial \mu_p}{\partial E_p} + 2\mu_p P_2 & \beta_{11} &= \frac{1}{r_g}\frac{\partial \mu_g}{\partial E_g} + 2\mu_g T_2 \\ b_{02} &= \frac{1}{2r_p}\frac{\partial \mu_p}{\partial E_g} + \frac{\mu_p}{2r_p}\frac{\partial \mu_p}{\partial E_p} + \mu_p^2 P_2 & \beta_{20} &= \frac{1}{2r_g}\frac{\partial \mu_g}{\partial E_p} + \frac{\mu_g}{2r_g}\frac{\partial \mu_g}{\partial E_g} + \mu_g^2 T_2 \end{aligned}\right\}, \quad (4)$$

where P_2 and T_2 are new notations for b_{20} and β_{02}, respectively. Similar expressions may be derived for the coefficients b_{30}, β_{30}, etc.

If we now apply the circuital laws to the network external to the tube, we get a number of equations, two of which are

$$\epsilon_g = g + e_g, \qquad \epsilon_p = p + e_p. \quad (5)$$

To obtain a solution of (1) and (5) we utilize a method of successive approximations. Let

$$i_p = \sum i_{pk},\ i_g = \sum i_{gk},\ e_p = \sum e_{pk},\ e_g = \sum e_{gk},\ g = \sum g_k,\ p = \sum p_k, \quad (6)$$

where the summations extend from $k = 1$ to $k = \infty$. Let us further define the terms in the series (6) by the following equations:

$$\left.\begin{aligned} r_p i_{p1} - e_{p1} &= \mu_p e_{g1} \\ r_g i_{g1} - e_{g1} &= \mu_g e_{p1} \\ \epsilon_g = g_1 + e_{g1}, \qquad \epsilon_p &= p_1 + e_{p1} \end{aligned}\right\}, \quad (7)$$

$$\left.\begin{aligned} r_p i_{p2} - e_{p2} &= \mu_p e_{g2} + r_p(b_{20}e_{p1}^2 + b_{11}e_{p1}e_{g1} + b_{02}e_{g1}^2) \\ r_g i_{g2} - e_{g2} &= \mu_g e_{p2} + r_g(\beta_{20}e_{p1}^2 + \beta_{11}e_{p1}e_{g1} + \beta_{02}e_{g1}^2) \\ 0 = g_2 + e_{g2}, \qquad 0 &= p_2 + e_{p2} \end{aligned}\right\}, \quad (8)$$

$$\left.\begin{aligned} r_p i_{p3} - e_{p3} &= \mu_p e_{g3} + r_p[2b_{20}e_{p1}e_{p2} + b_{11}(e_{p1}e_{g2} + e_{p2}e_{g1}) + 2b_{02}e_{g1}e_{g2} \\ &\quad + b_{30}e_{p1}^3 + b_{21}e_{p1}^2e_{g1} + b_{12}e_{p1}e_{g1}^2 + b_{03}e_{g1}^3] \\ r_g i_{g3} - e_{g3} &= \mu_g e_{p3} + r_g[2\beta_{20}e_{p1}e_{p2} + \beta_{11}(e_{p1}e_{g2} + e_{p2}e_{g1}) + 2\beta_{02}e_{g1}e_{g2} \\ &\quad + \beta_{30}e_{p1}^3 + \beta_{21}e_{p1}^2e_{g1} + \beta_{12}e_{p1}e_{g1}^2 + \beta_{03}e_{g1}^3] \\ 0 = g_3 + e_{g3}, \qquad 0 &= p_3 + e_{p3} \end{aligned}\right\} \quad (9)$$

and so forth for subsequent terms.[5]

The physical interpretation of equations (7) to (9) is readily obtained. It follows from (7) that the equivalent circuit of Fig. 1 for first order quantities is given by Fig. 2. The equivalent circuit of

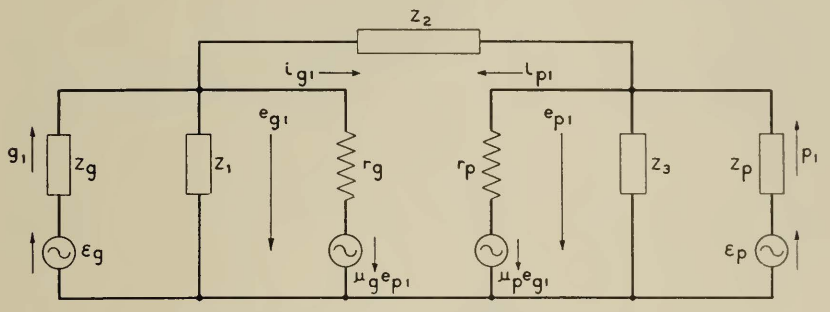


Fig. 2—Equivalent circuit—first-order effects.

Fig. 1 for second and third order effects are those shown in Fig. 3, and Fig. 4, respectively.

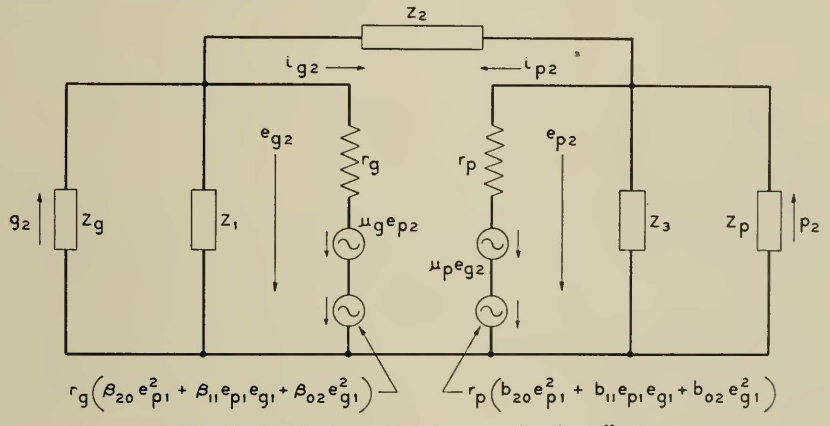


Fig. 3—Equivalent circuit—second-order effects.

It follows from (4) and (8) that

$$\begin{cases} r_p(b_{20}e_{p1}^2 + b_{11}e_{p1}e_{g1} + b_{02}e_{g1}^2) \\ \quad = r_pP_2(e_{p1} + \mu_pe_{g1})^2 + \frac{1}{2}\left(\frac{\partial\mu_p}{\partial E_g} + \mu_p\frac{\partial\mu_p}{\partial E_p}\right)e_{g1}^2 + \frac{\partial\mu_p}{\partial E_p}e_{p1}e_{g1}, \\ r_g(\beta_{20}e_{p1}^2 + \beta_{11}e_{p1}e_{g1} + \beta_{02}e_{g1}^2) \\ \quad = r_gT_2(e_{g1} + \mu_ge_{p1})^2 + \frac{1}{2}\left(\frac{\partial\mu_g}{\partial E_p} + \mu_g\frac{\partial\mu_g}{\partial E_g}\right)e_{p1}^2 + \frac{\partial\mu_g}{\partial E_g}e_{p1}e_{g1}. \end{cases} \tag{10}$$

The corresponding terms in (9) can be expressed similarly.

Equations (7), (8) and (9) contain the general theory of the three-electrode vacuum tube circuit. In the special case when conductive grid current is absent it is only necessary to omit the second equation

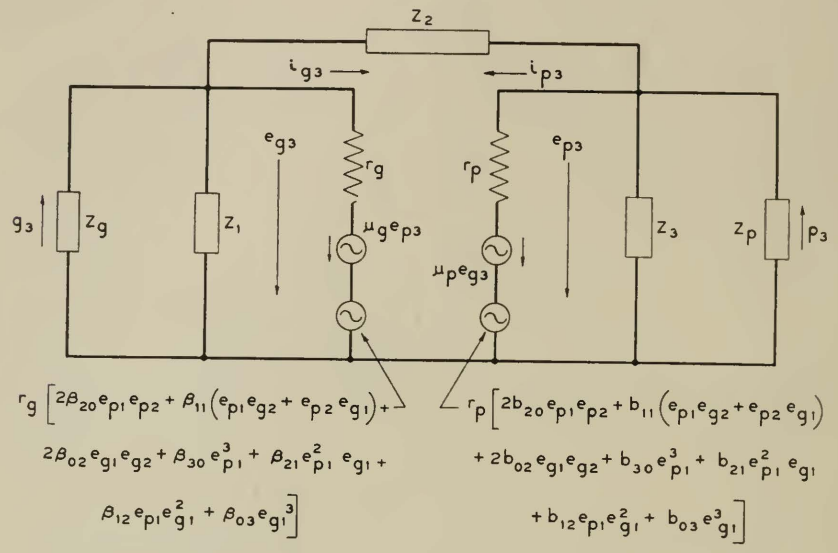


Fig. 4—Equivalent circuit—third-order effects.

in each of the equations (7) to (10), inclusive, and to omit in each of the Figs. 2, 3, and 4, the branch containing r_g. If it is assumed that not only conductive grid current is absent, but also that μ_p is constant, the second plate e.m.f. in (8) reduces to $r_p P_2(e_{p1} + \mu_p e_{g1})^2$ as is seen from (10), and (8) thus becomes identical with the corresponding equation already obtained previously.[5] A similar reduction and correspondence occurs for the second plate e.m.f. in (9), as well as in subsequent equations.

Application to Steady State Solutions

In this section the use of the theory is illustrated by obtaining first and second-order effects assuming the circuit configuration to be that shown in Fig. 1. To avoid unnecessary complications the discussion is limited to steady-state solutions, and it is also assumed that no plate e.m.f. is impressed. We shall first obtain the solutions in the general case and then indicate how these are simplified in such special cases which have been treated by some previous investigators.[2, 3, 5]

General Case

Let the impressed grid e.m.f. be

$$\epsilon_g = \sum k_h \cos(\omega_h t + \kappa_h) = R\sum k_h e^{i(\omega_h t+\kappa_h)}, \qquad i = \sqrt{-1} \tag{11}$$

where the summation extends from $h = 1$ to $h = n$, and the letter R before an expression means its real part. Referring to Fig. 2 it may be shown that

$$\left.\begin{aligned} e_{g1} &= R\sum \frac{\alpha_1(\omega_h)}{Z(\omega_h)} k_h e^{i(\omega_h t+\kappa_h)} \\ e_{p1} &= R\sum -\frac{\alpha_2(\omega_h)}{Z(\omega_h)} k_h e^{i(\omega_h t+\kappa_h)} \\ (e_{p1} + \mu_p e_{g1}) &= R\sum r_p \frac{\alpha_3(\omega_h)}{Z(\omega_h)} k_h e^{i(\omega_h t+\kappa_h)} \\ (e_{g1} + \mu_g e_{p1}) &= R\sum \frac{\alpha_4(\omega_h)}{Z(\omega_h)} k_h e^{i(\omega_h t+\kappa_h)} \end{aligned}\right\}, \tag{12}$$

where

$$\begin{aligned} Z(\omega) = \frac{1}{(Z_g' + Z_2)r_g Z_2} \{&[Z_2(r_g + Z_g') + r_g Z_g'] \\ &\times [Z_2(r_p + Z_p') + r_p Z_p'] \\ &- Z_g' Z_p'(\mu_g Z_2 - r_g)(\mu_p Z_2 - r_p)\}, \end{aligned} \tag{13}$$

$$\alpha_1(\omega) = \frac{Z_1[Z_2(r_p + Z_p') + r_p Z_p']}{(Z_1 + Z_g)(Z_2 + Z_g')}, \tag{14}$$

$$\alpha_2(\omega) = \frac{Z_1 Z_p'(\mu_p Z_2 - r_p)}{(Z_1 + Z_g)(Z_2 + Z_g')}, \tag{15}$$

$$\alpha_3(\omega) = \frac{Z_1(\mu_p Z_2 + \mu_p Z_p' + Z_p')}{(Z_1 + Z_g)(Z_2 + Z_g')}, \tag{16}$$

$$\alpha_4(\omega) = \frac{Z_1 Z_2 r_p + Z_1 Z_2 Z_p'(1 - \mu_g\mu_p) + Z_1 r_p Z_p'(1 + \mu_g)}{(Z_1 + Z_g)(Z_2 + Z_g')}, \tag{17}$$

$$Z_p'(\omega) = \frac{Z_3 Z_p}{Z_3 + Z_p}, \qquad Z_g'(\omega) = \frac{Z_1 Z_g}{Z_1 + Z_g}. \tag{18}$$

The right-hand expressions in (13)–(18) are to be evaluated at ω. If we write

$$\frac{\alpha_k(\omega)}{Z(\omega)} = \left|\frac{\alpha_k(\omega)}{Z(\omega)}\right| e^{-i\varphi_k(\omega)}, \qquad k = 1, 2, 3, 4, \tag{19}$$

the equations (12) can be written

$$\left.\begin{aligned} e_{g1} &= \sum \left|\frac{\alpha_1(\omega_h)}{Z(\omega_h)}\right| k_h \cos\left[\omega_h t + \kappa_h - \varphi_1(\omega_h)\right] \\ e_{p1} &= -\sum \left|\frac{\alpha_2(\omega_h)}{Z(\omega_h)}\right| k_h \cos\left[\omega_h t + \kappa_h - \varphi_2(\omega_h)\right] \\ e_{p1} + \mu_p e_{g1} &= \sum r_p \left|\frac{\alpha_3(\omega_h)}{Z(\omega_h)}\right| k_h \cos\left[\omega_h t + \kappa_h - \varphi_3(\omega_h)\right] \\ e_{g1} + \mu_g e_{p1} &= \sum \left|\frac{\alpha_4(\omega_h)}{Z(\omega_h)}\right| k_h \cos\left[\omega_h t + \kappa_h - \varphi_4(\omega_h)\right] \end{aligned}\right\}. \tag{20}$$

This concludes our consideration of effects of the first order and we now turn to those of the second order. For this purpose we substitute the values given by (20) in the right-hand side of the expressions (10) for the grid and plate e.m.f.'s, and we then obtain two expressions, each of which is equal to a sum of sinusoidal terms. If we limit our attention to the terms of frequency $(\omega_1-\omega_2)$, it is readily shown that the grid e.m.f. of this frequency is equal to the real part of

$$\left[r_g T_2 \frac{\alpha_4(\omega_1)}{Z(\omega_1)} \left(\overline{\frac{\alpha_4(\omega_2)}{Z(\omega_2)}}\right) + \frac{1}{2}\left(\frac{\partial \mu_g}{\partial E_p} + \mu_g \frac{\partial \mu_g}{\partial E_g}\right) \frac{\alpha_2(\omega_1)}{Z(\omega_1)} \left(\overline{\frac{\alpha_2(\omega_2)}{Z(\omega_2)}}\right) - \frac{1}{2}\frac{\partial \mu_g}{\partial E_g}\left\{ \frac{\alpha_1(\omega_1)}{Z(\omega_1)} \left(\overline{\frac{\alpha_2(\omega_2)}{Z(\omega_2)}}\right) + \left(\overline{\frac{\alpha_1(\omega_2)}{Z(\omega_2)}}\right) \frac{\alpha_2(\omega_1)}{Z(\omega_1)} \right\} \right] k_1 k_2 e^{i((\omega_1-\omega_2)t+\kappa_1-\kappa_2)} \tag{21}$$

and the plate e.m.f. is equal to the real part of

$$\left[r_p{}^3 P_2 \frac{\alpha_3(\omega_1)}{Z(\omega_1)} \left(\overline{\frac{\alpha_3(\omega_2)}{Z(\omega_2)}}\right) + \frac{1}{2}\left(\frac{\partial \mu_p}{\partial E_g} + \mu_p \frac{\partial \mu_p}{\partial E_p}\right) \frac{\alpha_1(\omega_1)}{Z(\omega_1)} \left(\overline{\frac{\alpha_1(\omega_2)}{Z(\omega_2)}}\right) - \frac{1}{2}\frac{\partial \mu_p}{\partial E_p}\left\{ \frac{\alpha_1(\omega_1)}{Z(\omega_1)} \left(\overline{\frac{\alpha_2(\omega_2)}{Z(\omega_2)}}\right) + \left(\overline{\frac{\alpha_1(\omega_2)}{Z(\omega_2)}}\right) \frac{\alpha_2(\omega_1)}{Z(\omega_1)} \right\} \right] k_1 k_2 e^{i((\omega_1-\omega_2)t+\kappa_1-\kappa_2)}, \tag{22}$$

where a bar over a quotient indicates its conjugate complex.

It follows from Fig. 3 by straightforward calculations that the currents i_{p2} and i_{g2} produced by the e.m.f.'s (21) and (22), are

$$\left.\begin{aligned} i_{p2}(\omega_1-\omega_2) &= R\left[-\frac{[\epsilon_g]}{Z_b(\omega_1-\omega_2)} + \frac{[\epsilon_p]}{Z_d(\omega_1-\omega_2)} \right] \\ i_{g2}(\omega_1-\omega_2) &= R\left[\frac{[\epsilon_g]}{Z_a(\omega_1-\omega_2)} - \frac{[\epsilon_p]}{Z_c(\omega_1-\omega_2)} \right] \end{aligned}\right\}, \tag{23}$$

where $[\epsilon_g]$ and $[\epsilon_p]$ are abbreviations for the complex quantities (21)

and (22), respectively, and

$$\left.\begin{aligned} Z_a(\omega) &= Z(\omega)\frac{r_g(Z_g' + Z_2)}{Z_2(r_p + Z_p') + r_pZ_p'} = |Z_a(\omega)|e^{i\psi_a(\omega)} \\ Z_b(\omega) &= Z(\omega)\frac{r_g(Z_g' + Z_2)}{Z_g'(\mu_pZ_2 - r_p)} = |Z_b(\omega)|e^{i\psi_b(\omega)} \\ Z_c(\omega) &= Z(\omega)\frac{r_g(Z_2 + Z_g')}{Z_p'(\mu_gZ_2 - r_g)} = |Z_c(\omega)|e^{i\psi_c(\omega)} \\ Z_d(\omega) &= Z(\omega)\frac{r_g(Z_2 + Z_g')}{Z_2(r_g + Z_g') + r_gZ_g'} = |Z_d(\omega)|e^{i\psi_d(\omega)} \end{aligned}\right\}. \qquad (24)$$

In (24) the introduction of the angles ψ is convenient when it is desired to evaluate the real parts of the expressions (23).

The expressions (21) to (24) can be used to obtain any second-order current of frequency $(\omega_a - \omega_b)$ by replacing ω_1 with ω_a and ω_2 with ω_b. The remaining second-order currents are found by a process similar to that above. For instance, $i_{p2}(2\omega_1)$ and $i_{g2}(2\omega_1)$ are given by the right-hand expressions in (23) provided the e.m.f.'s $[\epsilon]$ are those of frequency $(2\omega_1)$ and the impedances Z are evaluated at $(2\omega_1)$. In passing it may be remarked that equations similar to those in (23) and (24) also occur when third and higher-order effects are calculated.

Special Cases

If the impedances Z_1, Z_2, and Z_3 are infinite the case treated above reduces to that considered by Llewellyn,[3] and after proper simplifications the previous equations give results identical with those obtained by him. For instance, if we take the limiting values of e_{p1} in (12) as Z_1, Z_2, and Z_3 tend to infinity, and if we then divide the quantity inside the summation sign by $-Z_p(\omega_h)$, we get an expression for i_{p1} which may be shown to be identical with equation (33) in Llewellyn's paper, except for differences in notations. Similarly, the plate current $i_{p2}(\omega_1 - \omega_2)$ in (23) reduces to a value which may be shown to be equal to the sum of his equations (35) and (36), evaluated for this type of second-order current.

Another special case is that when the impedances Z_1, Z_2, and Z_3 are all finite but conductive grid current is absent. We then have μ_g equal to zero, and R_g equal to infinity, and the previous general equations are simplified correspondingly.

We arrive at the case treated by Peterson-Evans [2] by maintaining the assumption of no conductive grid current but by assuming Z_1, Z_2, and Z_3 to be infinite. For instance, if then i_{p1} and $i_{p2}(\omega_1 - \omega_2)$ are

evaluated on this basis for a plate impedance Z_p equal to a pure resistance at all frequencies, it can be shown that the currents so obtained are identical with the corresponding currents given by equations (4) and (6) in the paper referred to.

Finally, if we assume finite values for Z_1, Z_2, and Z_3, no conductive grid current, and constant μ_p, we have the case treated in the previous paper.[5]

Theory of Four-Electrode Tube Circuits

Circuits with tubes having more than three electrodes can be treated by a process similar to that adopted above, as will be made clear by outlining the theory for the four-electrode tube circuit.

The circuit to be considered is shown in Fig. 5 where Z_1 to Z_6 are

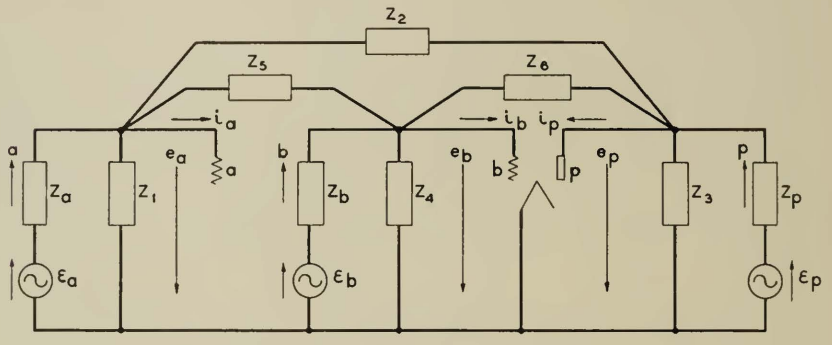


Fig. 5—Four-electrode vacuum tube and circuit.

impedances which may include inter-electrode admittances. On the electrodes denoted by a, b, and p are impressed the variable electromotive forces ϵ_a, ϵ_b, and ϵ_p in series with the impedances Z_a, Z_b, and Z_p, respectively. The significance of the quantities E_p, E_{p0}, e_p, I_p, I_{p0}, i_p and the corresponding quantities with indices a and b, is obvious from the preceding discussion of the three-electrode tube circuit. Let a, b, and p be the incremental voltages across the impedances Z_a, Z_b, and Z_p. As before, instantaneous values are implied.

For the currents we get the expansions

$$\left.\begin{aligned} i_p &= P_1e_a + P_2e_b + P_3e_p + P_4e_a^2 + P_5e_b^2 + P_6e_p^2 \\ &\quad + P_7e_ae_b + P_8e_ae_p + P_9e_be_p + \cdots \\ i_a &= A_1e_a + A_2e_b + A_3e_p + A_4e_a^2 + A_5e_b^2 + A_6e_p^2 \\ &\quad + A_7e_ae_b + A_8e_ae_p + A_9e_be_p + \cdots \\ i_b &= B_1e_a + B_2e_b + B_3e_p + B_4e_a^2 + B_5e_b^2 + B_6e_p^2 \\ &\quad + B_7e_ae_b + B_8e_ae_p + B_9e_be_p + \cdots \end{aligned}\right\} \quad (25)$$

where

$$\left.\begin{aligned} P_1 &= \frac{\partial I_p}{\partial E_a},\quad P_2 = \frac{\partial I_p}{\partial E_b},\quad P_3 = \frac{\partial I_p}{\partial E_p},\quad P_4 = \frac{1}{2}\frac{\partial^2 I_p}{\partial E_a{}^2},\quad P_5 = \frac{1}{2}\frac{\partial^2 I_p}{\partial E_b{}^2} \\ P_6 &= \frac{1}{2}\frac{\partial^2 I_p}{\partial E_p{}^2},\ P_7 = \frac{\partial^2 I_p}{\partial E_a \partial E_b},\ P_8 = \frac{\partial^2 I_p}{\partial E_a \partial E_p},\ P_9 = \frac{\partial^2 I_p}{\partial E_b \partial E_p},\ \cdots \end{aligned}\right\} \tag{26}$$

and similar expressions hold for the A- and B-values.

The electrode resistances are by definition

$$\frac{1}{r_p} = \frac{\partial I_p}{\partial E_p},\qquad \frac{1}{r_a} = \frac{\partial I_a}{\partial E_a},\qquad \frac{1}{r_b} = \frac{\partial I_b}{\partial E_b}. \tag{27}$$

The mu-factors are

$$\left.\begin{aligned} \mu_{pa} &= \frac{\dfrac{\partial I_p}{\partial E_a}}{\dfrac{\partial I_p}{\partial E_p}} = -\left(\frac{dE_p}{dE_a}\right)_{E_b,\ I_p=\text{const.}} \\ \mu_{ap} &= \frac{\dfrac{\partial I_a}{\partial E_p}}{\dfrac{\partial I_a}{\partial E_a}} = -\left(\frac{dE_a}{dE_p}\right)_{E_b,\ I_a=\text{const.}} \\ \mu_{bp} &= \frac{\dfrac{\partial I_b}{\partial E_p}}{\dfrac{\partial I_b}{\partial E_b}} = -\left(\frac{dE_b}{dE_p}\right)_{E_a,\ I_b=\text{const.}} \\ \mu_{pb} &= \frac{\dfrac{\partial I_p}{\partial E_b}}{\dfrac{\partial I_p}{\partial E_p}} = -\left(\frac{dE_p}{dE_b}\right)_{E_a,\ I_p=\text{const.}} \\ \mu_{ab} &= \frac{\dfrac{\partial I_a}{\partial E_b}}{\dfrac{\partial I_a}{\partial E_a}} = -\left(\frac{dE_a}{dE_b}\right)_{E_p,\ I_a=\text{const.}} \\ \mu_{ba} &= \frac{\dfrac{\partial I_b}{\partial E_a}}{\dfrac{\partial I_b}{\partial E_b}} = -\left(\frac{dE_b}{dE_a}\right)_{E_p,\ I_b=\text{const.}} \end{aligned}\right\} \tag{28}$$

and the transconductances

$$\left.\begin{aligned} S_{pa} &= \frac{\partial I_p}{\partial E_a} = \frac{\mu_{pa}}{r_p} \qquad S_{pb} = \frac{\partial I_p}{\partial E_b} = \frac{\mu_{pb}}{r_p} \\ S_{ap} &= \frac{\partial I_a}{\partial E_p} = \frac{\mu_{ap}}{r_a} \qquad S_{ab} = \frac{\partial I_a}{\partial E_b} = \frac{\mu_{ab}}{r_a} \\ S_{bp} &= \frac{\partial I_b}{\partial E_p} = \frac{\mu_{bp}}{r_b} \qquad S_{ba} = \frac{\partial I_b}{\partial E_a} = \frac{\mu_{ba}}{r_b} \end{aligned}\right\}. \tag{29}$$

It can now be shown that

$$\left.\begin{aligned} &P_1 = \frac{\mu_{pa}}{r_p}, \qquad P_2 = \frac{\mu_{pb}}{r_p}, \qquad P_3 = \frac{1}{r_p} \\ &P_6 = -\frac{1}{2{r_p}^2}\frac{\partial r_p}{\partial E_p}, \; P_8 = \frac{1}{r_p}\frac{\partial \mu_{pa}}{\partial E_p} + 2\mu_{pa}P_6, \; P_9 = \frac{1}{r_p}\frac{\partial \mu_{pb}}{\partial E_p} + 2\mu_{pb}P_6 \\ &P_4 = \frac{1}{2r_p}\frac{\partial \mu_{pa}}{\partial E_a} + \frac{\mu_{pa}}{2r_p}\frac{\partial \mu_{pa}}{\partial E_p} + {\mu_{pa}}^2 P_6 \\ &P_5 = \frac{1}{2r_p}\frac{\partial \mu_{pb}}{\partial E_b} + \frac{\mu_{pb}}{2r_p}\frac{\partial \mu_{pb}}{\partial E_p} + {\mu_{pb}}^2 P_6 \\ &P_7 = \frac{1}{r_p}\frac{\partial \mu_{pb}}{\partial E_a} + \frac{\mu_{pb}}{r_p}\frac{\partial \mu_{pa}}{\partial E_p} + 2\mu_{pb}\mu_{pa}P_6 \\ &\quad = \frac{1}{r_p}\frac{\partial \mu_{pa}}{\partial E_b} + \frac{\mu_{pa}}{r_p}\frac{\partial \mu_{pb}}{\partial E_p} + 2\mu_{pa}\mu_{pb}P_6 \end{aligned}\right\}, \tag{30}$$

$$\left.\begin{aligned} &A_1 = \frac{1}{r_a}, \qquad A_2 = \frac{\mu_{ab}}{r_a}, \qquad A_3 = \frac{\mu_{ap}}{r_a} \\ &A_4 = -\frac{1}{2{r_a}^2}\frac{\partial r_a}{\partial E_a}, \; A_7 = \frac{1}{r_a}\frac{\partial \mu_{ab}}{\partial E_a} + 2\mu_{ab}A_4, \; A_8 = \frac{1}{r_a}\frac{\partial \mu_{ap}}{\partial E_a} + 2\mu_{ap}A_4 \\ &A_5 = \frac{1}{2r_a}\frac{\partial \mu_{ab}}{\partial E_b} + \frac{\mu_{ab}}{2r_a}\frac{\partial \mu_{ab}}{\partial E_a} + {\mu_{ab}}^2 A_4 \\ &A_6 = \frac{1}{2r_a}\frac{\partial \mu_{ap}}{\partial E_p} + \frac{\mu_{ap}}{2r_a}\frac{\partial \mu_{ap}}{\partial E_a} + {\mu_{ap}}^2 A_4 \\ &A_9 = \frac{1}{r_a}\frac{\partial \mu_{ap}}{\partial E_b} + \frac{\mu_{ap}}{r_a}\frac{\partial \mu_{ab}}{\partial E_a} + 2\mu_{ap}\mu_{ab}A_4 \\ &\quad = \frac{1}{r_a}\frac{\partial \mu_{ab}}{\partial E_p} + \frac{\mu_{ab}}{r_a}\frac{\partial \mu_{ap}}{\partial E_a} + 2\mu_{ab}\mu_{ap}A_4 \end{aligned}\right\}, \tag{31}$$

$$\left.\begin{aligned}
&B_1 = \frac{\mu_{ba}}{r_b}, \qquad B_2 = \frac{1}{r_b}, \qquad B_3 = \frac{\mu_{bp}}{r_b} \\
&B_5 = -\frac{1}{2r_b{}^2}\frac{\partial r_b}{\partial E_b}, \quad B_7 = \frac{1}{r_b}\frac{\partial \mu_{ba}}{\partial E_b} + 2\mu_{ba}B_5, \quad B_9 = \frac{1}{r_b}\frac{\partial \mu_{bp}}{\partial E_b} + 2\mu_{bp}B_5 \\
&B_4 = \frac{1}{2r_b}\frac{\partial \mu_{ba}}{\partial E_a} + \frac{\mu_{ba}}{2r_b}\frac{\partial \mu_{ba}}{\partial E_b} + \mu_{ba}{}^2B_5 \\
&B_6 = \frac{1}{2r_b}\frac{\partial \mu_{bp}}{\partial E_p} + \frac{\mu_{bp}}{2r_b}\frac{\partial \mu_{bp}}{\partial E_b} + \mu_{bp}{}^2B_5 \\
&B_8 = \frac{1}{r_b}\frac{\partial \mu_{bp}}{\partial E_a} + \frac{\mu_{bp}}{r_b}\frac{\partial \mu_{ba}}{\partial E_b} + 2\mu_{bp}\mu_{ba}B_5 \\
&\quad = \frac{1}{r_b}\frac{\partial \mu_{ba}}{\partial E_p} + \frac{\mu_{ba}}{r_b}\frac{\partial \mu_{bp}}{\partial E_p} + 2\mu_{ba}\mu_{bp}B_5
\end{aligned}\right\}. \quad (32)$$

The circuital laws applied to the external network furnish a number of equations, three of which are

$$\epsilon_a = a + e_a, \qquad \epsilon_b = b + e_b, \qquad \epsilon_p = p + e_p. \quad (33)$$

Let now

$$\left.\begin{aligned}
i_p &= \sum i_{pk}, & i_a &= \sum i_{ak}, & i_b &= \sum i_{bk} \\
e_p &= \sum e_{pk}, & e_a &= \sum e_{ak}, & e_b &= \sum e_{bk} \\
p &= \sum p_k, & a &= \sum a_k, & b &= \sum b_k
\end{aligned}\right\}. \quad (34)$$

We then obtain the equations

$$\left.\begin{aligned}
r_p i_{p1} - e_{p1} &= \mu_{pa}e_{a1} + \mu_{pb}e_{b1} \\
r_a i_{a1} - e_{a1} &= \mu_{ab}e_{b1} + \mu_{ap}e_{p1} \\
r_b i_{b1} - e_{b1} &= \mu_{ba}e_{a1} + \mu_{bp}e_{p1} \\
\epsilon_a = a_1 + e_{a1}, &\qquad \epsilon_b = b_1 + e_{b1}, \qquad \epsilon_p = p_1 + e_{p1}
\end{aligned}\right\}, \quad (35)$$

which show that the equivalent circuit for first order effects is that given in Fig. 6.

We get further for second-order quantities

$$\left.\begin{aligned}
r_p i_{p2} - e_{p2} &= \mu_{pa}e_{a2} + \mu_{pb}e_{b2} + r_p L \\
r_a i_{a2} - e_{a2} &= \mu_{ab}e_{b2} + \mu_{ap}e_{p2} + r_a M \\
r_b i_{b2} - e_{b2} &= \mu_{ba}e_{a2} + \mu_{bp}e_{p2} + r_b N \\
0 = a_2 + e_{a2}, &\qquad 0 = b_2 + e_{b2}, \qquad 0 = p_2 + e_{p2}
\end{aligned}\right\}, \quad (36)$$

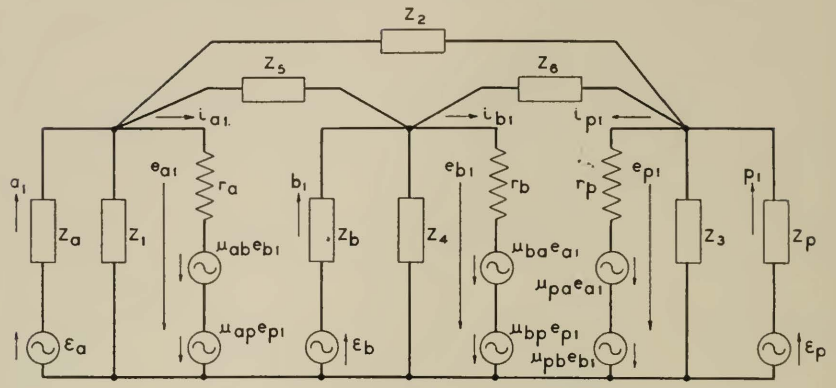


Fig. 6—Equivalent circuit—first-order effects.

where

$$\left.\begin{aligned} r_pL &= (P_4e_{a1}^2+P_5e_{b1}^2+P_6e_{p1}^2+P_7e_{a1}e_{b1}+P_8e_{a1}e_{p1}+P_9e_{b1}e_{p1})r_p \\ r_aM &= (A_4e_{a1}^2+A_5e_{b1}^2+A_6e_{p1}^2+A_7e_{a1}e_{b1}+A_8e_{a1}e_{p1}+A_9e_{b1}e_{p1})r_a \\ r_bN &= (B_4e_{a1}+B_5e_{b1}+B_6e_{p1}+B_7e_{a1}e_{b1}+B_8e_{a1}e_{p1}+B_9e_{b1}e_{p1})r_b \end{aligned}\right\}, \quad (37)$$

which in view of (30), (31) and (32) may be written

$$\left.\begin{aligned} r_pL = {} & r_pP_6(\mu_{pa}e_{a1}+\mu_{pb}e_{b1}+e_{p1})^2+\frac{1}{2}\left(\frac{\partial\mu_{pa}}{\partial E_a}+\mu_{pa}\frac{\partial\mu_{pa}}{\partial E_p}\right)e_{a1}^2 \\ & +\frac{1}{2}\left(\frac{\partial\mu_{pb}}{\partial E_b}+\mu_{pb}\frac{\partial\mu_{pb}}{\partial E_p}\right)e_{b1}^2+\left(\frac{\partial\mu_{pb}}{\partial E_a}+\mu_{pb}\frac{\partial\mu_{pa}}{\partial E_p}\right)e_{a1}e_{b1} \\ & +\frac{\partial\mu_{pa}}{\partial E_p}e_{a1}e_{p1}+\frac{\partial\mu_{pb}}{\partial E_p}e_{b1}e_{p1} \end{aligned}\right\}, \quad (38)$$

$$\left.\begin{aligned} r_aM = {} & r_aA_4(e_{a1}+\mu_{ab}e_{b1}+\mu_{ap}e_{p1})^2+\frac{1}{2}\left(\frac{\partial\mu_{ab}}{\partial E_b}+\mu_{ab}\frac{\partial\mu_{ab}}{\partial E_a}\right)e_{b1}^2 \\ & +\frac{1}{2}\left(\frac{\partial\mu_{ap}}{\partial E_p}+\mu_{ap}\frac{\partial\mu_{ap}}{\partial E_a}\right)e_{p1}^2+\frac{\partial\mu_{ab}}{\partial E_a}e_{a1}e_{b1} \\ & +\frac{\partial\mu_{ap}}{\partial E_a}e_{a1}e_{p1}+\left(\frac{\partial\mu_{ap}}{\partial E_b}+\mu_{ap}\frac{\partial\mu_{ab}}{\partial E_a}\right)e_{b1}e_{p1} \end{aligned}\right\}, \quad (39)$$

$$\left.\begin{aligned} r_bN = {} & r_bB_5(\mu_{ba}e_{a1}+e_{b1}+\mu_{bp}e_{p1})^2+\frac{1}{2}\left(\frac{\partial\mu_{ba}}{\partial E_a}+\mu_{ba}\frac{\partial\mu_{ba}}{\partial E_b}\right)e_{a1}^2 \\ & +\frac{1}{2}\left(\frac{\partial\mu_{bp}}{\partial E_p}+\mu_{bp}\frac{\partial\mu_{bp}}{\partial E_b}\right)e_{p1}^2+\frac{\partial\mu_{ba}}{\partial E_b}e_{a1}e_{b1} \\ & +\left(\frac{\partial\mu_{bp}}{\partial E_a}+\mu_{bp}\frac{\partial\mu_{ba}}{\partial E_b}\right)e_{a1}e_{p1}+\frac{\partial\mu_{bp}}{\partial E_b}e_{b1}e_{p1} \end{aligned}\right\}. \quad (40)$$

From (36) it now follows that the equivalent circuit for second-order effects for the four-electrode tube circuit is as indicated in Fig. 7.

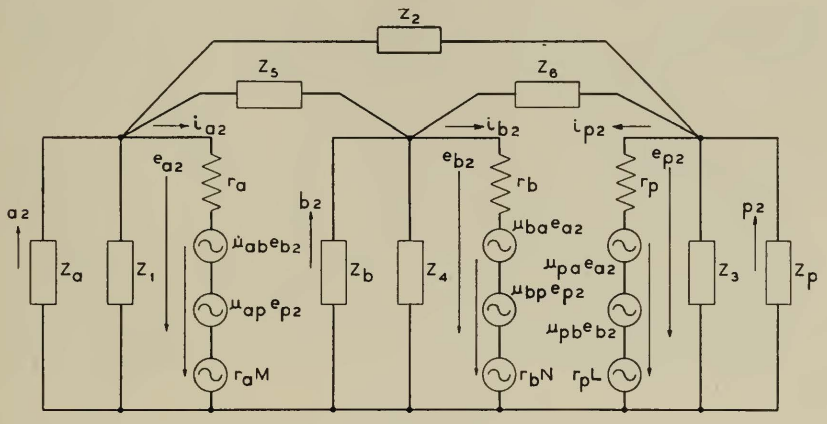


Fig. 7—Equivalent circuit—second-order effects.

Transatlantic Long-Wave Radio Telephone Transmission and Related Phenomena from 1923 to 1933

By AUSTIN BAILEY and HOWARD M. THOMSON

It is shown that transatlantic long-wave radio field strength is related to the 11-year cycles of terrestrial magnetic activity, sunspots, solar limb-prominences, and ultra-violet radiation. The directness of correlation between long-wave radio and these other phenomena is apparently only approximate. Very good seasonal and monthly correlation is obtained between magnetic activity and daylight radio transmission. Magnetic storms are shown to have prolonged and delayed effects on day and night radio transmission, obscuring tendencies for 27-day recurrences on long waves. No reference is made to the probable mechanism of long-wave radio transmission because a paper, now in preparation, will be concerned primarily with this subject.

Introduction

THIS paper summarizes a study of the transatlantic long-wave radio transmission data (near 60 kilocycles) collected by the American Telephone and Telegraph Company and the General Post Office of Great Britain during the period from 1923 to 1933, including the development tests from 1923 to 1926 and the period of operation of the commercial radio telephone circuits from January 1927 to December 1933.

A correlation is presented between terrestrial magnetic activity, sunspot-numbers, solar limb-prominences, ultra-violet radiation, and transatlantic long-wave radio telephone field strength observations during one 11-year sunspot cycle. By expressing the variables with different scales better correlations have been obtained. These are given on averages of years, months, seasons, and days. Examples of delayed night and day field strength changes accompanying long magnetic storms are included in this presentation.

It is generally believed that solar radiation influences the transmission [1*] of radio signals but the detailed mechanism by which the influence is exerted [2] is not entirely clear. The number and relative effects of other factors that may influence radio transmission are also uncertain. Solar phenomena, magnetic activity, and radio transmission appear to have some time-phase relations but these relations have not been clearly established. Some of the results of the studies reported herein contribute to the evaluation of these effects while other results give new or corroborative information on the qualitative relations between the solar, magnetic, and radio phenomena.

* Numbers refer to the references in the bibliography.

In any particular study, the field strengths of the radio transmitters were corrected to a constant antenna current and expressed in decibels above one microvolt per meter. Where a study involved a comparison between data observed at different locations, or on different frequencies, corrections were applied to reduce the data to a common basis by the application of the Bell System Long-Wave Radio Transmission Formula.[3]

The transmitters and transmission paths involved are indicated in Table I.

TABLE I

Years	Transmitter	Rocky Point, N. Y., to
	West–East	
1923–1924	57 Kc. 2XS	New Southgate, England
1925	57 Kc. 2XS	Chedzoy, England
1926 (to Sept.)	57 Kc. 2XS	Wroughton, England
1926 (Oct.) to 1928	60 Kc. WNL	Wroughton, England
1927 (Mar.) to 1933	60 Kc. WNL	Cupar, Scotland
	East–West	
		Rugby, England, to
1927–1930, part of 1931	60 Kc. GBT	Houlton, Maine
Remainder of 1931	68 Kc. GKA	Houlton, Maine
1932–1933	68 Kc. GBY	Houlton, Maine

Statistical Frequency Distributions

In making a correlation study the best results are obtained when observations of the variables involved have the same type of statistical "frequency distribution." As illustrated in Fig. 1, the cumulative frequency distribution curves of long-wave field strength observations, when plotted (in decibels) on "arithmetic probability" paper, tend to be straight lines and thus of the Normal Law type. The best correlation will therefore be obtained when the other variables involved have approximately straight-line distributions on the same type (arithmetic probability) of coordinates.

Accordingly, cumulative distribution curves were so made for all the variables used in this study. It was found that good approximations to the Normal Law distribution were obtained with logarithmic scales for radio field strength (db above 1 $\mu v/m$), monthly averages of solar limb-prominences, sunspot-numbers, and u_1 measure * of magnetic activity. However, for the "C" measure of magnetic character of days and monthly averages of ultra-violet radiation, linear scales were found to be better than logarithmic scales.

Correlation by Years

In spite of the irregularities in the long-wave field strength curves in Fig. 2, it is evident that during the sunspot maximum period the radio

* See Appendix I.

field strength tended to be higher than during the years of sunspot minimum. The same applies to solar limb-prominences, magnetic activity, and ultra-violet radiation.† The irregularities in the radio telephone field strength curves are more nearly duplicated in the curves

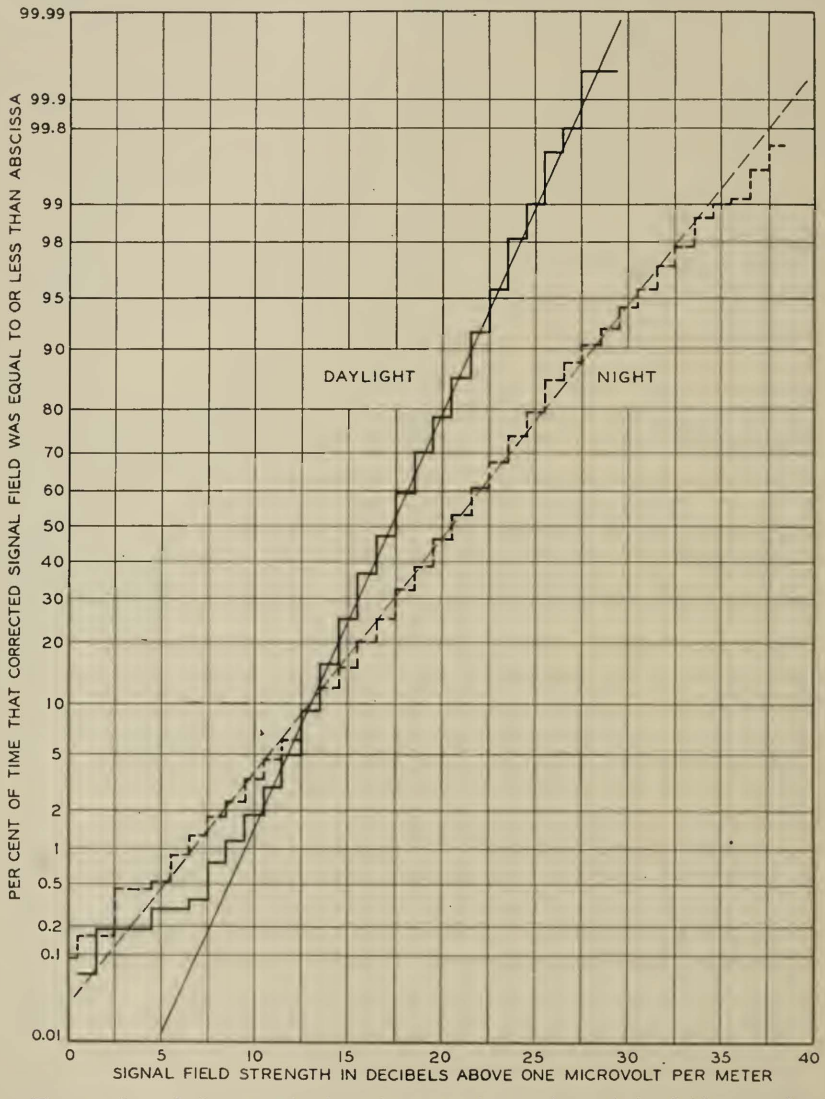


Fig. 1—Cumulative distribution of hourly observations of the field strength of WNL—60 kc. at Cupar, Scotland, in 1933, corrected to 375 amperes. (1699 observations for all-daylight path, 1146 observations for all-night-time path.)

† The measure of ultra-violet radiation is the ratio of intensity of ultra-violet ($\lambda = 0.32\mu$) to green ($\lambda = 0.50\mu$). Ratio for June 1924 = 1.

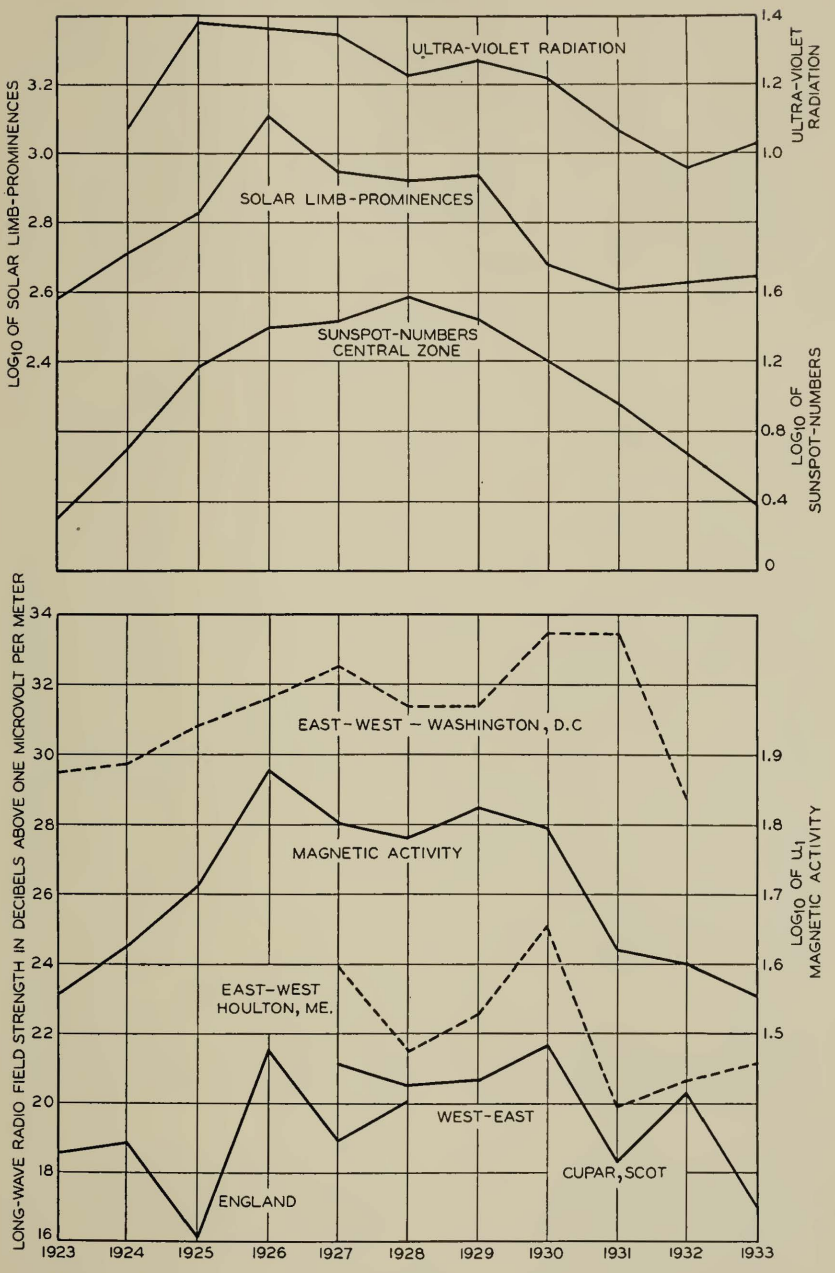


Fig. 2—Annual averages of solar phenomena, terrestrial magnetic activity, and long-wave radio transmission during one 11-year sunspot cycle.

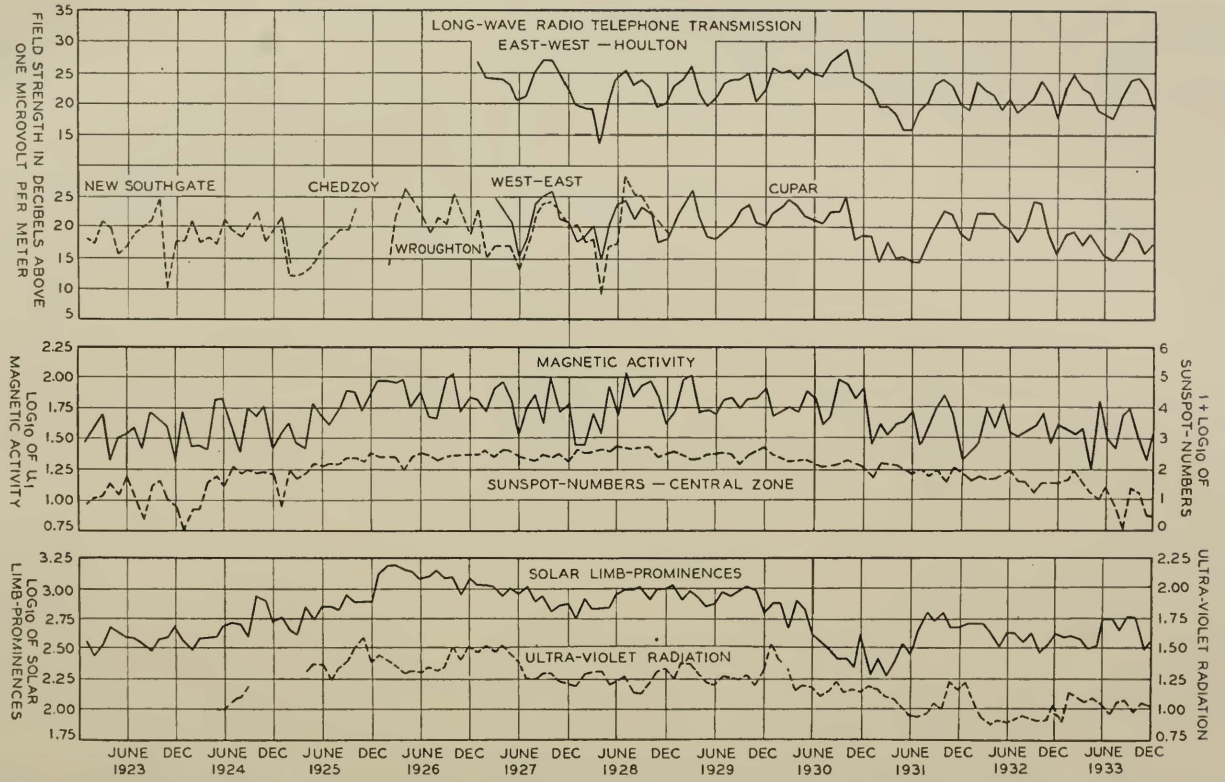


Fig. 3—Monthly averages of solar phenomena, terrestrial magnetic activity, and long-wave radio transmission during one 11-year sunspot cycle.

of magnetic activity and solar limb-prominences than in the sunspot and ultra-violet curves. The curve of reception at Washington, D. C., of signals from several European telegraph stations on longer wavelengths (15 to 30 kc.) followed the Houlton and Cupar curves fairly well except in 1931.

The decrease in long-wave field strength apparently lagged a year or two behind the decrease in sunspot-numbers. This may perhaps be explained by the lower heliographic latitude of the sunspots near the minimum of the solar cycle, offsetting the decrease in number of spots.

Correlation by Months

Monthly averages of the same phenomena as shown in Fig. 2 are plotted in Fig. 3, omitting the curve of radio telegraph reception at Washington. There seems to be little or no obvious correlation between the field strength curves and the other curves except for the one of magnetic activity. Sunspots in the central portion of the solar disc are believed to have the major influence on terrestrial phenomena,[4, 5, 6, 7] and therefore the sunspot-numbers employed in this analysis are those for the central zone * of the sun. Solar limb-prominences would hardly be expected to have considerable influence † on the terrestrial phenomena, although some relation is suggested by Fig. 2. Monthly averages of ultra-violet radiation also offer little explanation of the variations of long-wave fields.

Magnetic activity easily gives the best correlation with long-wave fields on a month-to-month basis. A scatter diagram was constructed in which the West–East and East–West monthly averages of daylight long-wave radio transmission were correlated with terrestrial magnetic activity for the years 1927–1932. For WNL the coefficient of correlation [8] is 0.526 ± 0.059; for GBT and GBY, 0.747 ± 0.067, indicating a high degree of correlation. Such correlation means either that magnetic activity affects radio transmission or that both are affected similarly by a common cause of disturbance.[9]

The coefficient of correlation of magnetic activity with the monthly averages of WNL and 2XS at New Southgate in 1923 and 1924, at Chedzoy in 1925, and at Wroughton in 1926, 1927, and 1928 is 0.38 ± 0.10.

* From 1917 to 1928, inclusive, the "central zone" of the sun was defined as that part of the sun's surface included between two meridians situated 30° on either side of the central meridian. (The central meridian is that meridian of the sun which bisects the sun's disc.) Beginning in the year 1929 the central zone is defined as the area on the sun's disc enclosed by a central circle having a diameter half that of the disc.

† Solar limb-prominences may have a more direct relation to sunspots without having any significant relation to terrestrial phenomena.

Very striking and significant results were obtained when seasonal averages of magnetic activity and radio field strength were plotted in Fig. 4 to indicate seasonal trends. Except for the magnetic submaximum in May the curves of field strength and magnetic activity

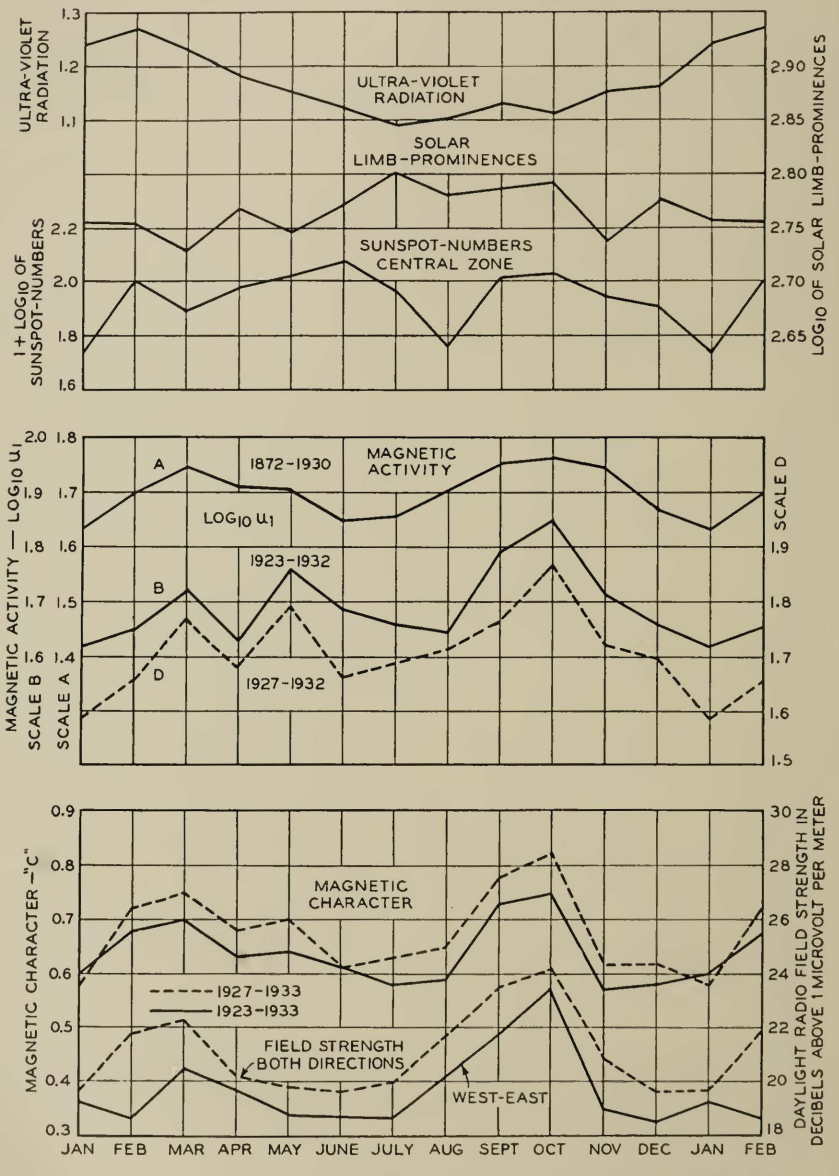


Fig. 4—Seasonal averages of long-wave radio field strength compared with solar phenomena and terrestrial magnetic activity.

are closely in phase. While the two curves of average magnetic characterization ("*C*") are considered not so reliable * for this analysis as the u_1 curves, the same equinoctial maxima appear in all magnetic curves. However, seasonal averages of solar limb-prominences, sunspot-numbers, and ultra-violet radiation apparently have no significance in relation to radio field strength. Bartels [10] found that there is no real seasonal cycle of sunspot-numbers. Cupar and Houlton observations were combined † for the years 1927 to 1933 to make the dashed curve.

West–East average transmission for eleven years is shown by the solid curve at the bottom of Fig. 4. The February dip was recorded at Cupar, Scotland, in 1931 only but was previously observed at the receiving locations in England (also during four of the seven years at Houlton). The $\log_{10} u_1$ curve for 1872–1930 was derived from data published by J. Bartels [10] in a comprehensive analysis of magnetic data, which he critically examined for the reality of seasonal variability. The other $\log_{10} u_1$ magnetic activity curves ‡ cover the first ten of the eleven years for which the radio data are available in one direction, and the first six of the seven years for which radio data are available in both directions, East–West and West–East.

Correlation with Magnetic Storms

The average daylight radio field strength of long-wave signals does not reach a maximum until five to eight days § after the peak of a long magnetic storm, nor does the average night-time field reach a minimum until about the same time. This is shown (Fig. 5) by a study of 25 storms of four or five days duration in the four years from 1930 to 1933, inclusive. The incidence of succeeding storms did not permit the study to include a longer time after the storms had subsided, so that less than one complete cycle of field variation is all that can be shown with any confidence. Data for WNL, GBT, and GBY were combined to obtain the trend curves for the radio field strength. A study of 16 periods (in the years 1927–1933) of one and two months' duration, including one to three significant magnetic storms, showed that the difficulty involved in extending this kind of study to individual storms lies primarily in the lack of highly accurate radio data.

* See Appendix I.

† See Appendix II.

‡ These curves were drawn before the data for magnetic activity (u_1) for 1933 were available. The 1933 data became available shortly before the paper was released for publication but do not materially change the overall picture.

§ C. N. Anderson [11] showed a similar delay in the maximum increase of daylight field following severe magnetic storms in 1927 and 1928. I. J. Wymore [12] found a delay of about two days for signals of somewhat longer wave-length.

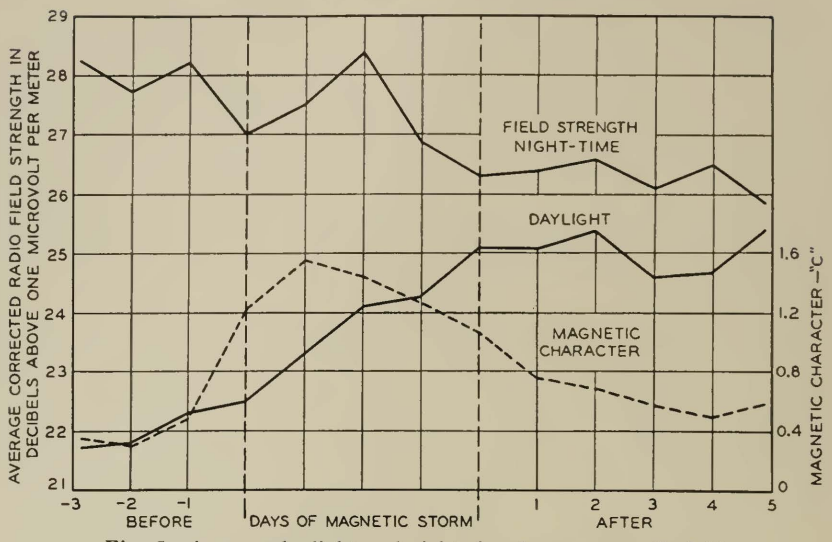


Fig. 5—Average daylight and night-time long-wave radio field strength before and after 25 long magnetic storms.

27-Day Intervals

Sunspot-numbers and terrestrial magnetic disturbances tend to be repeated at 27-day intervals.[10] A recurrence correlation for "radio character" and magnetic character has been published by A. M. Skellett,[13] who based the radio data upon the relative intensity from day to day of the disturbances of the short-wave telephone circuits between New York and London. A recurrence chart for long-wave radio is shown in Fig. 6, based upon the deviation from the 50-day moving average of the average daylight radio field strength of WNL (60 kc.) at Cupar for seven years beginning with March 1927. A measure of the international magnetic character figure ("*C*") for each day in the same period is likewise shown in this figure.

It is evident that the 27-day recurrence phenomenon is obscured on long waves, probably because magnetic storms are followed by prolonged effects on the long-wave fields, as was pointed out in connection with Fig. 5.

Correlation by Days

The average daylight field strength on long waves is improved on days of magnetic disturbance,[14] as compared with calm days.[15] Cumulative distribution curves of daily averages of daylight field strength of WNL (Rocky Point to Cupar, Scotland) for 1930 and 1933 are shown in Fig. 7, for calm and disturbed days.* Similar data for east-to-west

* A magnetically calm day is defined for the purposes of this paper as a day for which the international magnetic character figure "*C*" is from 0.0 to 0.3. A magnetically disturbed day is here characterized by a figure from 1.2 to 2.0.

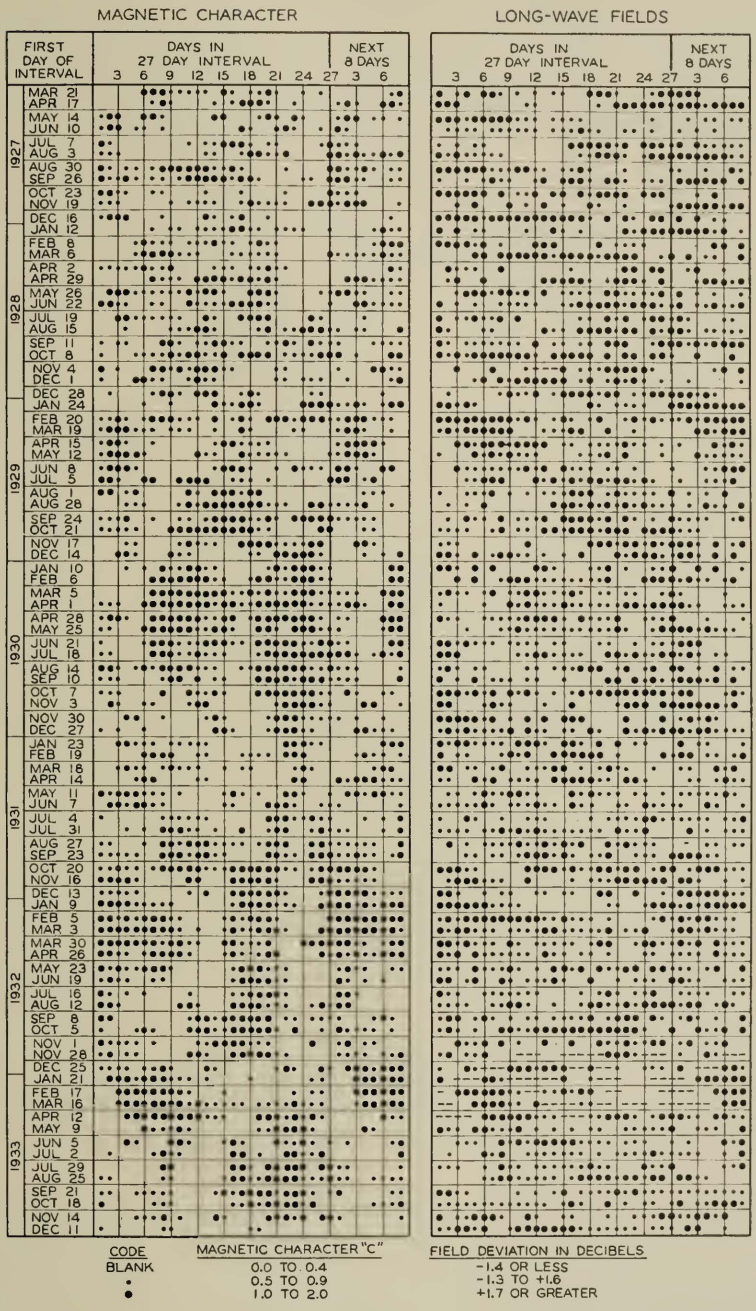


Fig. 6—Twenty-seven-day recurrence diagram of magnetic character and deviation of average daylight field of WNL from 50-day moving average.

transmission from Rugby to Houlton were obtained but are not reproduced. (It will be observed that the antenna current to which the data are corrected is different for different years. This basis results in a convenient spacing of the distribution curves to prevent

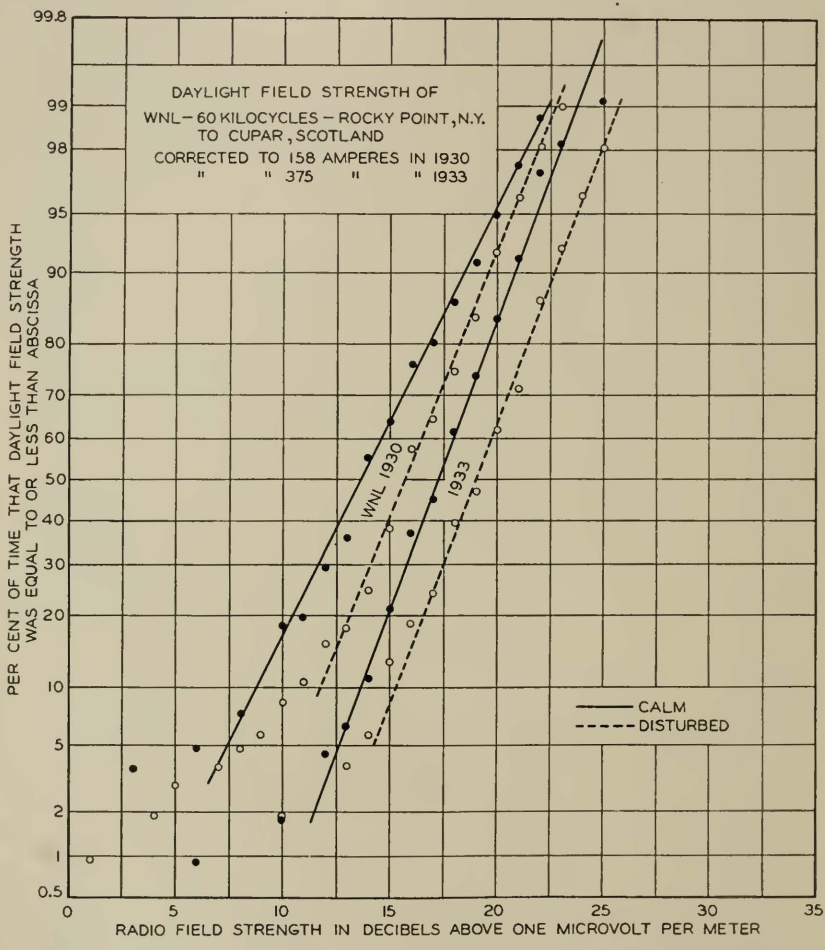


Fig. 7—Cumulative distribution of daily averages of long-wave radio field strength on days of magnetic calm and on days of magnetic disturbance. Daylight transmission path, WNL.

overlapping.) The year 1930 was magnetically more disturbed than 1933. Because of the few days' lag (Fig. 5) of the increase in daylight field strength behind magnetic storms, the full effect that accompanies magnetic disturbances is not revealed by these curves.

No consistent relationship between night-time long-wave fields and

magnetic disturbance is obtained from similar analyses. Lowered night-time fields on East–West (GBT, GBY) transmission were found to accompany magnetic disturbances in both quiet and disturbed years. However, from Fig. 8, it is seen that on West–East (WNL) trans-

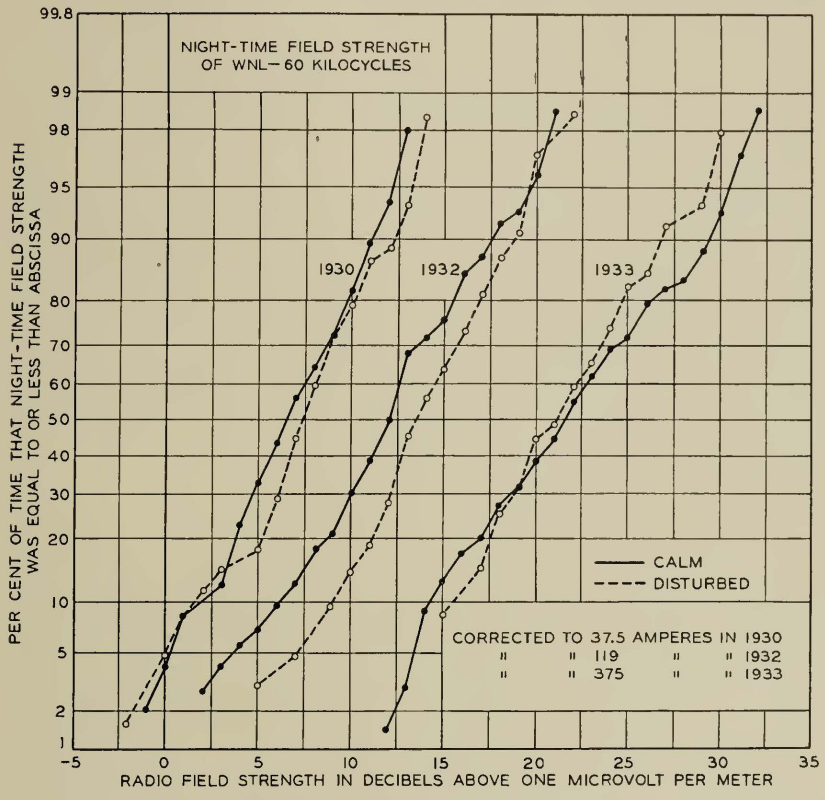


Fig. 8—Cumulative distribution of daily averages of long-wave radio field strength on days of magnetic calm and on days of magnetic disturbance. Night-time transmission path, WNL.

mission, the same result appears only in 1933, whereas in 1930 and 1932 night-time fields averaged higher on disturbed days than on calm days. No reason for this inconsistency has been found.

Correlation with Weather

Other investigators have shown some relation between long-wave field strength and temperature [16] and barometric pressure at short distances.[17] Because of the greater range of weather conditions, the relation is expected to be more obscure on long distance transmission. However, on European signals (15 to 20 kc.) received at Washington

from 1924 to 1929, during the periods from September 15 to May 15, there appears [18] "to be a decided drop in signal intensity (measured at 10:00 A.M. and 3:00 P.M.) coincident with heavy rainfall, rising temperture, and falling barometer, at the receiving station at the time of general storms, and an increased signal intensity as the storm center passes and is followed by clear and colder weather." A conclusion was presented [18] that "either radio waves are subject to large absorption in the lower atmosphere, due to a degree of ionization not known to exist, or surface storms are influenced by conditions existing at great heights."

Disturbed weather conditions may be part of the reason for disturbed transatlantic long-wave (60 kc.) radio reception, but the authors favor the hypothesis that local weather storms are more a symptom of disturbed conditions in higher regions. The same could be said of earth currents.[19] However, since reflection probably occurs at the surface of the earth one or more times on the transatlantic paths, as will be discussed more fully in a paper now in preparation, signals using these paths probably encounter surface weather conditions and earth current influences at least three times.

Correlation with Meteor Showers

An attempt was made to find some correlation between meteor showers and long-wave radio transmission. The sketchy nature of the available meteor data does not provide a satisfactory basis for such a study. Comparison of monthly average meteor hour rates [20] with monthly average radio transmission showed little or no promise of significant relations. Although meteors probably do exert an influence on radio transmission [13, 21, 22] by causing random variations of the received signal, due to changes in ionization of the transmission path, the evidence for major changes in field strength at times of meteoric activity apparently is lacking in our data.

During the four-month period centering on the Perseid shower,* Pickard [23] observed a considerable depression in the day reception of long-wave signals from Nauen (23 kc.) and a considerable increase in medium-wave signals at night from WBBM (1330 kc.), with maximum effects close to the day of meteor shower maximum. No such effect was detected, however, in the data on the long-wave transatlantic radio telephone circuits for the period from 1923 to 1933. In fact, as shown in Fig. 4, if the Perseid shower could be credited with having con-

* According to Olivier in Bulletin No. 8, the American Meteor Society (quoted also in reference 23), the Perseid shower is by far the longest and largest annual meteor shower and extends about 17 days on each side of August 11.

siderable effect on long-wave radio transmission, the effect would apparently be an increase in field strength in daytime, opposite to the result observed by Pickard on signals from Nauen. For the period covered by the curves in Fig. 4, the increased daylight field strength in August is not explained by correlation with the curves of magnetic activity.

Nature and Sources of Data

The long-wave radio telephone data used in this study are measurements of field strength at the receiving stations, and each individual observation is a simple physical measurement. The data taken before 1927, although relatively meager, were observations made under controlled conditions, a single-frequency signal being employed. Since 1927 the data, while much more abundant, had to be taken under operating conditions with speech transmission. However, since such measurements can be made on a commercial circuit only when in the judgment of the operating forces it is desirable and feasible to establish that circuit, it is inevitable that the overall picture will be distorted by the requirements of practical circuit operation. This could readily obscure such minor effects as correlation with meteoric showers, and would be likely to introduce some discrepancies between day and night behavior.

Sources of the data used in Figs. 2 and 3 are as follows:

Ultra-violet and sunspot data from *Terrestrial Magnetism*, **39**, 234, September 1934.
Solar limb-prominences data from *Astronomische Mitteilungen*, **130**, 217, 1933, and from **131**, 23, 1934.
Data for the Washington measurements—for 1923, figure by correspondence with the Bureau of Standards; for 1924 to 1930, reference 24; for 1931–1932, reference 25; observations were discontinued in 1933.
Magnetic character data, "*C*" measure, from volumes 29 to 39, inclusive, of *Terrestrial Magnetism*.
Magnetic activity, u_1 measure, from references 10, 26, and 28.

Acknowledgments

The authors wish to thank the personnel of the American Telephone and Telegraph Company and British General Post Office who made the observations and supplied the radio transmission data used in this paper. Acknowledgments are also due the Carnegie Institution of Washington, the United States Bureau of Standards, and Dr. Charles P. Olivier for other information used by the authors in correlating and interpreting the results.

Bibliography

1. H. Plendl, "Concerning the Influence of the Eleven-Year Solar Activity Period Upon the Propagation of Waves in Wireless Telegraphy," *Proc. I. R. E.*, **20**, 520, March 1932, and references.

2. G. W. Kenrick and G. W. Pickard, "Summary of Progress in the Study of Radio Wave Propagation Phenomena," *Proc. I. R. E.*, **18**, 649, April 1930, and references.
3. Lloyd Espenschied, C. N. Anderson, and Austin Bailey, "Transatlantic Radio Telephone Transmission," *Proc. I. R. E.*, **14**, 7, February 1926, *B. S. T. J.*, July 1925.
4. C. N. Anderson, "Correlation of Long-Wave Transatlantic Radio Transmission with Other Factors Affected by Solar Activity," *Proc. I. R. E.*, **16**, 297, March 1928.
5. L. W. Austin, "Solar Activity and Radio Telegraphy," *Proc. I. R. E.*, **20**, 280, February 1932, and references.
6. G. W. Pickard, "The Correlation of Radio Reception with Solar Activity and Terrestrial Magnetism," *Proc. I. R. E.*, **15**, 83, February 1927; also, **15**, 749, September 1927.
7. G. W. Pickard, "The Relation of Radio Reception to Sunspot Position and Area," *Proc. I. R. E.*, **15**, 1004, December 1927.
8. H. L. Reitz, "Mathematical Statistics" (book), Open Court Publishing Company, Chicago, 1927, Chapter IV.
9. H. B. Maris and E. O. Hulburt, "A Theory of Auroras and Magnetic Storms," *Phys. Rev.*, March 1929.
10. J. Bartels, "Terrestrial-Magnetic Activity and Its Relations to Solar Phenomena," *Terr. Mag.*, **37**, 1, March 1932.
11. C. N. Anderson, "Notes on the Effect of Solar Disturbances on Transatlantic Radio Transmission," *Proc. I. R. E.*, **17**, 1528, September 1929.
12. I. J. Wymore, "The Relation of Radio Propagation to Disturbances in Terrestrial Magnetism," *Proc. I. R. E.*, **17**, 1206, July 1929.
13. A. M. Skellett, "Disturbances in Radio Transmission," *Bell Laboratories Record*, **13**, 164, February 1935.
14. H. B. Maris and E. O. Hulburt, "Wireless Telegraphy and Magnetic Storms," *Proc. I. R. E.*, **17**, 494, March 1929.
15. C. N. Anderson, "Notes on Radio Transmission," *Proc. I. R. E.*, **19**, 1150, July 1931.
16. L. W. Austin and I. J. Wymore, "Radio Signal Strength and Temperature," *Proc. I. R. E.*, **14**, 781, December 1926.
17. K. Sreenivasan, "On the Relation Between Long-Wave Reception and Certain Terrestrial and Solar Phenomena," *Proc. I. R. E.*, **17**, 1793, October 1929.
18. I. J. Wymore-Shiel, "A Correlation of Long-Wave Radio Field Intensity with the Passage of Storms," *Proc. I. R. E.*, **19**, 1675, September 1931.
19. Isabel S. Bemis, "Some Observations of the Behavior of Earth Currents and Their Correlation with Magnetic Disturbances and Radio Transmission," *Proc. I. R. E.*, **19**, 1931, November 1931.
20. Charles P. Olivier, "Meteors" (book), Williams and Wilkins, Baltimore, 1925, Chapter 16, p. 182.
21. A. M. Skellett, "The Ionizing Effect of Meteors," *Proc. I. R. E.*, **23**, 132, February 1935.
22. A. M. Skellett, "The Ionizing Effect of Meteors in Relation to Radio Propagation," *Proc. I. R. E.*, **20**, 1933, December 1932.
23. G. W. Pickard, "A Note on the Relation of Meteor Showers and Radio Reception," *Proc. I. R. E.*, **19**, 1166, July 1931.
24. L. W. Austin, "Tables of North Atlantic Radio Transmission Conditions for Long-Wave Daylight Signals for the Years 1922–1930," *Proc. I. R. E.*, **20**, 689, April 1932.
25. L. J. Briggs, "North Atlantic Radio Transmission Conditions for Long-Wave Daylight Signals for the Years 1931 and 1932" (Letter to the Editor), *Terr. Mag.*, **38**, 67, March 1933.
26. J. Bartels, "Terrestrial-Magnetic Activity in the Years 1931 and 1932," *Terr. Mag.*, **39**, 1, March 1934.
27. Charles Chree, "The Criterion of Magnetic Disturbance," *Terr. Mag.*, **28**, 33, March 1923.
28. J. Bartels, "Terrestrial-Magnetic Activity in the Year 1933 and at Huancayo," *Terr. Mag.*, **40**, September 1935.

APPENDIX I

Measures of Terrestrial Magnetic Activity "C" and "u_1"

Characterization (C)

"Terrestrial-magnetic activity at a given station, and in a certain interval, may be defined as an expression for the frequency and intensity of magnetic disturbances in that interval. There are many ways in which this general definition may be expressed as a numerical measure. *Characterization*, the simplest, is now widely used. In this measure every observatory assigns, from the character of its photographic records, to each interval of 24 hours, between successive Greenwich midnights, a character-figure, '0' for quiet, '1' for moderately disturbed, and '2' for greatly disturbed days. The average for all collaborating observatories (the number of which increased from 30 to about 45, since this measure was begun in 1906) is the international magnetic character-figure C." [10]

"A primary desideratum is to arrive at a clear idea of exactly what it is we want to measure. If our object is simply to discriminate between the days of a single month, with a view to selecting for special purposes the five quietest or five most disturbed days of the month, it is very doubtful whether the existing scheme of international 'character' figures can be improved on. Its simplicity and the small amount of labor it entails are great recommendations. The disadvantages it seems to me to possess are: 1. The significance of any particular 'character' figure, *e.g.*, 1.5, is variable; it connotes decidedly less disturbance in a quiet than in a disturbed year. Also, while the mean 'character' figure for the year does to a certain extent wax and wane with disturbance, the variation seems to me inadequate. 2. The assigning of 'character figures' at an individual station is largely a psychological process, depending on the temperament and knowledge of the judge. The standards in use at different stations at the same time are widely different, and the standard in use at any particular station may vary largely from time to time. The 'character' figures supplied by any two stations do not suffice for a satisfactory intercomparison of the stations, and if we wish to compare one year or season with another the international 'character' figures leave a great deal to be desired." [27]

The "C" measure of magnetic activity may be inferior as a basis for comparing years, but is relatively safe for comparing short intervals such as days within a month. For comparing longer intervals a

different measure is desirable, having a homogeneous standard of evaluating magnetic activity from year to year and for all time. This standard measure was called "u," from which was derived the "u_1" measure used in this paper.

u_1 Measure

"A short *definition of* u may be repeated as the average change, taken without regard to sign, measured in the unit 0.0001 c.g.s. = 10γ, of the daily means of horizontal intensity on the magnetic equator of the earth. The monthly means of u_1 are derived from those of u:

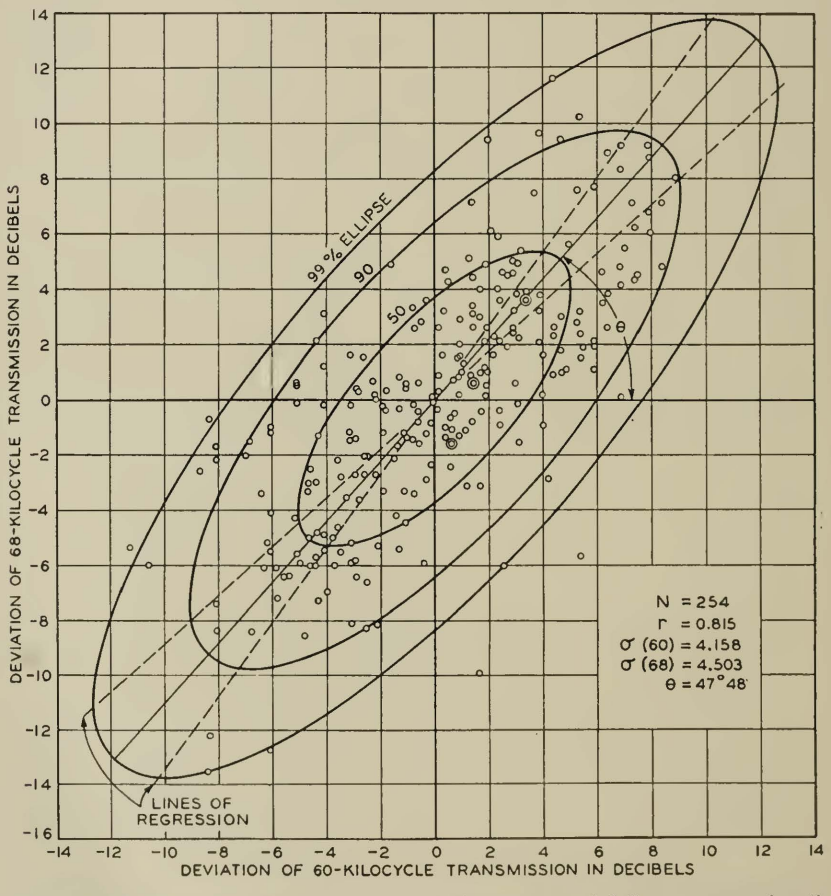


Fig. 9—Correlation diagram of scattering from mean of daily averages of radio transmission over all-daylight paths for the 254 days from May 5, 1931 to March 31, 1932, when GKA–GBY (Rugby, England) was transmitting on 68 kc. and WNL (Rocky Point, N. Y.) was transmitting on 60 kc.

For values of u from 0.0 to 0.6, $u_1 = 100u - 30$.

For higher values,[10]

$$0.6 \text{ to } 1.6,\ u_1 = 30 + 100(u - 0.6) - 30(u - 0.6)^2,$$
$$1.6 \text{ to } 3.6,\ u_1 = 100 + 40(u - 1.6) - 10(u - 1.6)^2,$$
$$3.6 \text{ to } \infty,\ u_1 = 140,$$

u_1 increases less rapidly, approaching asymptotically the limiting value 140. The quantity u_1 has been introduced in order to obtain a measure of activity which has a frequency-distribution similar to that of the relative sunspot-numbers, and therefore is more suitable for research on correlations between terrestrial-magnetic activity and solar activity."[26]

APPENDIX II

Data for East–West transmission were combined with data for West–East transmission for two illustrations, Figs. 4 and 5, given in this paper. Two unpublished studies by the authors of this paper showed that the correlation between transmission from Rugby to Houlton on 68 kilocycles and transmission from Rocky Point to Cupar on 60 kilocycles was very high, 0.80 ± 0.01. The scatter diagrams for these studies were so similar that only one is shown here, Fig. 9. For transmission on 60 kilocycles in both directions the correlation was 0.76 ± 0.02. These results are considered sufficient justification for combining East–West and West–East radio transmission data to obtain average curves.

TECHNICAL DIGESTS

With this issue the first of a series of briefer articles, known as "Technical Digests," is being introduced. These will review papers by Bell System authors which appear in the current numbers of other scientific or engineering periodicals.

The purpose of these digests, which in length will run to several times that of the usual abstract, may be given as twofold. By eliminating the need hitherto felt that numerous articles be reprinted in full, a larger variety of subjects will find compass in a given number of Technical Journal pages. Secondly and more important, it is hoped that the digests will succeed in so outlining the important features of their respective articles as to make them available to a much wider circle of readers than would have opportunity for a detailed reading or study of the original.

Superiorities of Lead-Calcium Alloys for Storage Battery Construction *

RECENT investigations, conducted at the Bell Telephone Laboratories and elsewhere, have demonstrated that the lead-antimony alloys almost universally employed in storage cell construction are far from ideal for the purpose from the electrochemical standpoint. It has been shown that *in the course of normal operation* of the present type cell, antimony is leached out of the positive electrode, passes through the solution and deposits on the negative, where it promotes "local action" and self-discharge. Also, it has been demonstrated that stibine is generated in perceptible amounts by the present type battery on over-charge.

The continued use of lead-antimony alloys for over fifty years has been due primarily to their desirable metallurgical and physical characteristics and the fact that other equally satisfactory alloys of lead have not been available. Electrochemical theory indicates that for use in storage cell construction, lead should be alloyed only with metals *electronegative* to it. The alloying constituents should have little tendency to diffuse or segregate at the normal operating temperatures reached in a cell. The resulting alloy should be considerably stronger than lead, easily cast, and resistant to electrolytic corrosion. It should also have high electrical conductivity and small solidification and thermal contraction.

As a result of a comprehensive investigation of lead-calcium alloys in connection with cable sheathing materials, data were accumulated at the Bell Telephone Laboratories which suggested the use of certain of these alloys for storage battery grids and plates. Tests have been conducted to determine the value of this suggestion, and with very promising results.

In the course of the cable sheath studies, the thermal equilibrium diagram of lead-calcium alloys containing a very small percentage of the calcium component was determined and is illustrated in Fig. 1. This diagram shows that the amount of calcium soluble in solid lead

* Digest of Two Papers: "The Electrochemical Behavior of Lead, Lead-antimony and Lead-calcium Alloys in Storage Cells" by H. E. Haring and U. B. Thomas, and "Some Physical and Metallurgical Properties of Lead-calcium Alloys for Storage Cell Grids and Plates" by E. E. Schumacher and G. S. Phipps. These papers are, being presented before the Convention of the Electrochemical Society in Washington, October, 1935, and published in the *Transactions* of the Society.

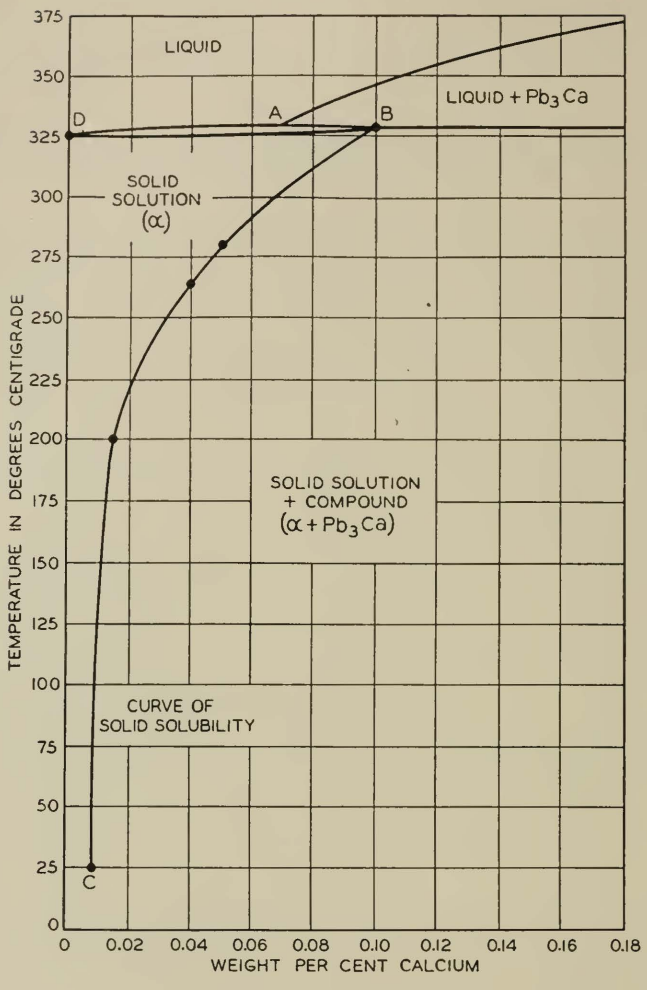


Fig. 1—A partial equilibrium diagram of the system lead-calcium, according to Schumacher and Bouton.

varies continuously from 0.10 per cent at 328° C. to about 0.01 per cent at room temperature. Moreover, when the 0.10 per cent alloy is cooled below 328° C., the calcium precipitates out of the solid solution in the form of Pb_3Ca. This precipitation, while a gradual process, can be hastened by suitable heat treatment. Physically, the effect of the Pb_3Ca molecules highly dispersed throughout the body of the alloy is to strengthen and harden it very materially. Thus, the alloy with a calcium content of 0.10 per cent has a tensile strength of around 8,000 pounds per square inch, a value comparable to that reached by

the lead-antimony alloys most generally used in storage battery construction. It has been found, furthermore, that at ordinary temperatures no decrease of tensile strength occurs with age. This maintenance of strength is probably related to the low rate of diffusion of the hardening constituent Pb_3Ca which, being a large and stable molecule, does not diffuse readily in the matrix. The electrical conductivity of this lead-calcium alloy is approximately 20 per cent greater than that of lead-9 per cent antimony, a factor of importance in securing uniform distribution of current when large currents are drawn from a battery.

Experimental cells of both Planté and Fauré (pasted) types have been constructed and are being subjected to a variety of tests. To date, forty-two cells of the starting and lighting type have been pasted, assembled, formed and cycled. This investigation has progressed sufficiently to make it quite evident that the behavior of a complete cell can be predicted with a high degree of accuracy if the electrochemical properties of the structural alloy are known. It has been definitely established that when the negative electrodes of lead-calcium cells of the starting and lighting type (previously subjected to 100 charge-discharge cycles) are charged and allowed to stand one month, they lose only 4 to 5 per cent of their charge as compared to 20 to 25 per cent for lead-antimony cells similarly treated. The efficiencies of lead-calcium cells have been found to be superior to those of lead-antimony cells. No undue corrosion of lead-calcium grids has been observed.

Marine Radio Telephone Service for Boston Harbor *

By F. A. GIFFORD and R. B. MEADER

THERE has been a constantly increasing interest in an inexpensive service for small harbor and coastal craft such as tugboats, private yachts, coastal passenger ships, merchant craft and fishing vessels. This interest became particularly evident in New England in 1931 and since equipment suitable for the purpose had recently been developed by Bell Telephone Laboratories, the New England Telephone and Telegraph Company undertook the establishment of a marine radio telephone service.

A survey consisting of a comprehensive series of field strength measurements on shipboard and at various points along the coasts of Massachusetts and Cape Cod Bays resulted in selecting Green Harbor as the location for a shore station. Green Harbor is in the town of Marshfield, Massachusetts, about 28 miles southeast of Boston.

A commercial survey indicated that initially the service would be of interest chiefly to the Boston fishing industry. Consequently, boat radio telephone equipment was installed on the trawler "Flow" of the Bay State Fishing Company and the service was opened in June, 1932, on a demonstration basis. As tests with this vessel progressed, it became evident that the radio telephone service would fulfill the communications requirements of the fishing industry. It also became apparent that a complete service of this type should include some means for determining the vessel's position at any time by means of radio. Therefore, the development of such equipment was initiated by Bell Telephone Laboratories as an adjunct to the radio telephone service, and the outcome of tests of an experimental model indicates that the problem of providing suitable radio compass equipment in the price range satisfactory to the fishing fleet owners has been satisfactorily solved.

The radio transmitter is a 400-watt crystal-controlled type similar to those designed for use at aviation ground stations and adjusted to operate at a frequency of 2506 kilocycles. This frequency is maintained within limits of better than 0.025 per cent.

In order to combine the two unidirectional radio channels into a two-way circuit suitable for connection to the ordinary wire circuits in the

* Digest of a paper to be published in full in *Communication and Broadcast Engineering*, October, 1935.

land telephone network, apparatus similar to that provided at the terminals of the transatlantic and high seas ship-to-shore radio telephone circuits has been provided. This apparatus includes controls for adjusting the volume of speech into the transmitter from the receiver to the wire lines and the usual voice operated devices (termed "vodas") provided for the suppression of echoes and singing.

The apparatus mentioned, together with a volume indicator, means for talking, monitoring and signaling, and testing apparatus, is as-



Fig. 1—Installation of control unit on fishing trawler.

sembled on one floor-mounted apparatus bay. This, mounted adjacent to the bay containing the receiver and a noise suppression device, constitutes the operating position. This position is continuously attended by a technical operator who adjusts the controls during the progress of each call, guided by indications of the meters provided, to insure the best possible connection under the conditions obtained at the time. Power for the terminal apparatus is supplied by a motor generator set operating from a 110-volt 60-cycle source.

The noise suppression device, termed a "codan" (carrier operated device anti-noise) is employed in connection with the receiver which introduces a predetermined high loss into the audio-frequency portion of the radio receiver during intervals when there is no incoming carrier, and insures relatively quiet conditions on the receiving line. When a carrier is received, the codan action removes this loss, allowing the speech with which the carrier is modulated to pass from the receiver

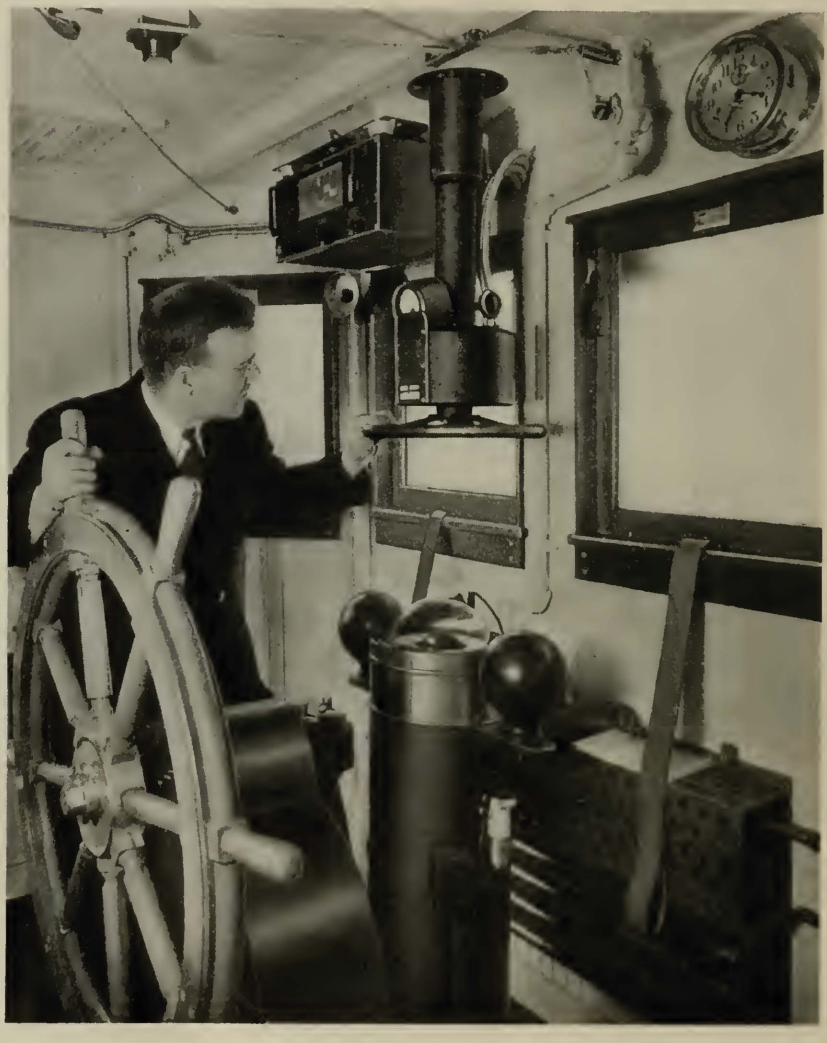


Fig. 2—Marine radio compass installed on a trawler.

output to the receiving line. This device makes it possible to deliver higher speech volumes to the telephones on shore since it prevents the operation of the receiving vodas relays on radio noise during the idle intervals. It also prevents the high radio noise which would otherwise result due to the increase in gain inserted by the automatic volume control whenever the carrier is interrupted.

It can be seen from this discussion that the use of the codan presupposes suppression of the carrier of the distant transmitter except when the user wishes to talk. This method of operation has been adopted for this type of marine radio telephone service.

A 10-kw. 220-volt three-phase 60-cycle alternator driven by a Buffalo gasoline engine has been provided to furnish the necessary emergency power supply in case the normal commercial supply fails.

The frequency designated by the Federal Communications Commission for use by ships communicating with the shore through the Green Harbor radio telephone station is 2110 kilocycles. This carrier frequency is maintained within limits of 0.025 per cent.

Two crystals are provided and the circuit arranged so that the receiver may be quickly adjusted to operate on either of two frequencies by means of a local mechanical or an electrically operated remote control. The receiver is so designed as to make possible boat-to-boat conversations on a separate frequency.

The signaling unit which is normally connected to the output of the radio receiver consists of a selector operated under the control of an arrangement of relays which in turn are controlled by incoming signal pulses of 600 and 1500-cycle tones. The bell on each vessel is operated only in response to the particular code of pulses to which the selector is adjusted. Arrangements are also included so that the vessels of any one fleet may be called simultaneously.

A motor generator set operates continuously while the vessel is standing by for the reception of signals and furnishes 12- and 200-volt power for the operation of the radio receiver and signaling unit. A second motor generator set is automatically started when the handset is lifted from the switch hook to place a call or in response to an incoming signal, and furnishes power to operate the transmitter. On several of the smaller boats having 32-volt power supply with wide voltage fluctuations, power supply equipment consisting of two dynamotors operated from a 12-volt battery charged from the vessel's storage battery has been employed successfully.

The control unit for the radio telephone consists of a small panel on which are mounted a switch for turning the set on and off, a meter for indicating antenna current, a manual volume control, pilot lamp

signals and a bell for announcing incoming calls. A special handset with push button completes the control unit assembly.

At Boston, the marine radio telephone traffic is handled at two positions on the outgoing toll board especially modified for this purpose. The wire lines from the Green Harbor station terminate at this point, and calls from vessels can be switched by the operator to any point connected to Bell System facilities. The normal wire lines are three loaded cable pairs. One pair is used for transmission from the shore telephone to a boat, the second conducts speech received from a boat through the toll position to the land line telephone, and the third is employed as an order wire for communication between the operator at the marine position and the technical operator at the Green Harbor station. All of these circuits are duplicated over an alternate route for use in case of trouble on the normal facilities.

The marine operator dials the code assigned to the vessel desired. The dialing operation produces the desired grouping of 600 and 1500 cycle pulses which modulate the radio transmitter carrier frequency of 2506 kilocycles. The signaling unit on the vessel is actuated by these pulses and the bell rings. The captain raises the handset from the switch hook on the control unit, presses the push button in the handle of the handset and announces the name of his vessel. The operator then completes the connection and the conversation takes place.

In placing a call from boat to shore the captain or member of the crew raises the handset from the switch hook and, after listening to ascertain that no conversations are in progress, presses the push button and calls "marine operator." The marine operator who is normally monitoring on the channel ascertains the name of the calling vessel, the shore station desired and other necessary details, and while the calling party holds the line proceeds to call the land line telephone and establish the connection.

When a person on one boat wishes to talk with a person on another boat, the procedure in placing the call and establishing the connection is the same as in the case of a ship-to-shore call, except that when both are prepared to talk, the technical operator at Green Harbor operates a by-pass key which connects the radio receiver output and radio transmitter input without including the voice operated device and other equipment associated with the land circuits. The land line is bridged onto the circuit so that the marine operator may be advised of any difficulties which arise in carrying on the conversation.

During the more than two years that the system has been in experimental service, the transmission results up to distances of 500 miles

from the shore station have been quite satisfactory. Of course, during periods of abnormally heavy static the normal range is somewhat reduced. The service is available at all times, but practically all business is handled between the hours of 8 A.M. and 6 P.M. so that the relatively poor atmospheric conditions usually existing during summer nights do not adversely affect the radio telephone traffic. However, experience has indicated that calls originated during such periods from vessels within the normal range can be handled satisfactorily. During some periods of favorable atmospheric conditions experimental transmissions over distances greatly in excess of the normal range have been successfully conducted.

On fishing vessels the radio telephone equipment is accessible for maintenance work only at the conclusion of trips which are usually of about ten days' duration. It is obvious, therefore, that the equipment must be designed for reliable operation over long periods and experience indicates that these requirements have been well satisfied.

Fishing craft normally make use of the service for reporting the details of the catch, for making arrangements to return to port and for talking with other fishing vessels to locate points where fishing is best. The radio telephone has proved of vital importance on several occasions where engine breakdowns necessitated advice from shore in order to make repairs and having replacement parts available upon the vessel's arrival at port. In several instances of sickness and accidents to members of a crew, medical advice has been obtained or the Coast Guard summoned to remove the injured man for quick transportation to a hospital. In one case of severe damage to a trawler as a result of a collision, the Coast Guard were summoned and the owners were able to keep in constant touch with the situation.

Harbor Craft Ship-to-Shore Radio Telephone Service in Puget Sound Area *

By E. B. HANSEN

PUGET SOUND with an area of about 6,000 square miles serves a large region whose principal resources are lumber and fishing. Movement of the raw products to points where the manufacturing and processing are carried on comprises a major part of the short-haul shipping in the Puget Sound area. Large lumber and pulp mills have been permanently established at locations where deep sea shipping and rail facilities are readily available. Logs in the form of large booms are towed to the mills from the sources of supply which extend throughout the area drained by the Puget Sound. The average haul of each boom is about 75 miles and consumes several days. Because of the relatively long time a tug is isolated in each operation, some economical means of dispatching and directing the activities of the tow boats from the land headquarters is desirable. In the case of fishing, there is an even greater need than in the case of log towing for close contact between fishing boat and cannery in order that both the fishing and cannery activities may be coordinated so as to prevent waste of fish during large runs.

From the above, it is clear that some means of communication between harbor craft and land stations would be useful. A preliminary view of the problem indicated that low-powered radio telephone channels from each of these boats to a single land station and thence by wire lines to any telephone, would be the most economical and practical means from the standpoint of the ship owners. Furthermore, the government regulations covering the issuance of operators' licenses for radio telephone transmitters of 50 watts or less power are such that the average member of a ship's crew can qualify after a few hours' instruction, thus making it unnecessary for vessels employing this equipment to carry a special radio operator.

After it was decided that this service was feasible and should be established the initial order of procedure was the selection of the site for the land station.

Due to the fact that it is desirable to locate the transmitter and receiver in juxtaposition, a wide variety of factors had to receive con-

* Digest of a paper published in *Electrical Engineering*, August, 1935, pp. 828–831.

sideration. As to transmission, the distances involved in the harbor area are such as to make it advantageous, from the standpoint of quality and freedom from fading, to make use of a frequency which will permit covering the entire area with the direct or "ground" wave. The signal strengths obtained with ground wave transmission, however,

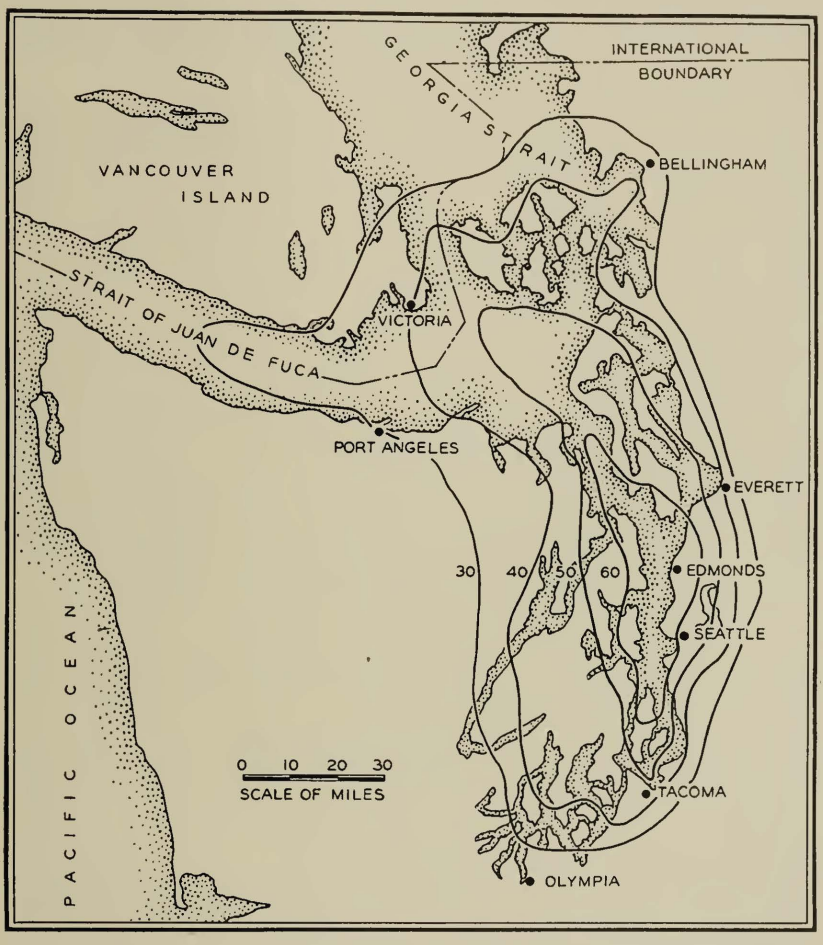


Fig. 1—Field strengths are indicated in decibels above one microvolt per meter.

are affected considerably by the terrain, and the location of the station as much as a mile inland might reduce the over-water range as much as 50 miles. The selection of a receiving site involved a consideration of interference from a variety of sources including high voltage power lines, automobiles and motor boats as well as atmospheric disturbances.

After a survey involving transmitted field strengths and noise measurements, a 13-acre plot of tide and water land having ready access to power supply and telephone connections was selected on Point Edwards about one mile south of Edmonds, Washington, and fifteen miles from Seattle.

Due to the fact that the site for the transmitting and receiving station is under water at high tide, the building housing the radio equipment is of frame construction erected on treated piling.

The transmitting and receiving antennas, located 200 feet apart and near the building are also supported by piling. Both antennas are of the simple vertical type, the former 80 feet in height and supported by a single pole 110 feet high, and the latter 40 feet in height. A tuning unit is mounted in a weather-proofed box on the receiving antenna pole. The transmission line from the receiving antenna to the receiver[1] in the radio equipment building is a special armored and lead covered concentric conductor type.

The equipment at the land station consists of a 400-watt transmitter with a rectifier unit for power supply similar to the type generally used for aviation service. The frequency stability is obtained by the use of a quartz crystal frequency control and will maintain its frequency to better than 0.025 per cent. It is designed for substantially complete modulation of the carrier and under this condition little distortion occurs to the speech frequencies. A tuning unit for tuning the transmitting antenna to resonance is also housed in the building.

As shown on the block diagram Fig. 2, a four-wire circuit is used to connect the radio station and the Seattle office where the circuit becomes two-wire for switching to land subscribers. The equipment includes means for regulating speech volumes; outgoing, so as to always properly load the radio transmitter, and incoming so as to give the shore subscriber the best received volume. A voice-operated device known as the "vodas" provides means for suppressing echoes and singing. This device is essentially the same in principle and in its operation as those used in intercontinental service which have been described in detail in previous publications. In addition to the above units apparatus for monitoring and testing is provided.

The harbor station, moreover, is arranged for remote control operation, equipment being provided so that the transmitter is automatically turned on when the operator inserts a plug in the "Harbor Circuit" jack.

During operation the transmitter is continuously monitored by means of an auxiliary radio receiver in Seattle tuned to the transmitter frequency. In addition, the two-way portion of the voice-frequency

[1] For a description of the receiver see the following digest.

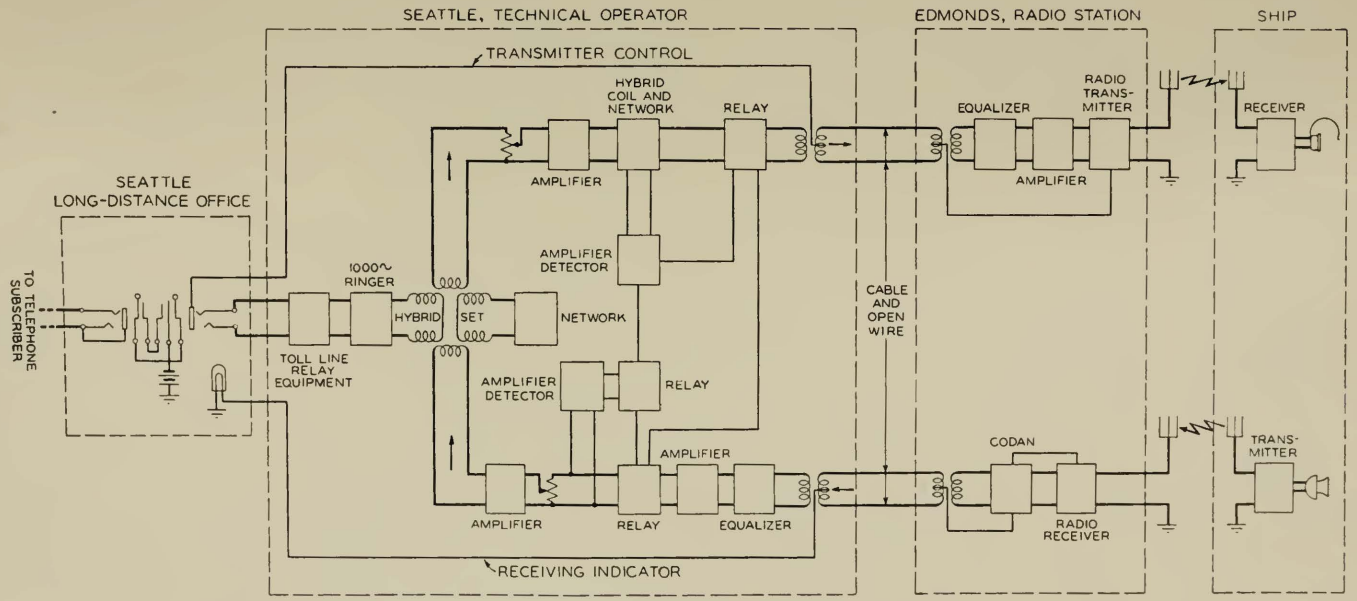


Fig. 2—Schematic of harbor radio telephone circuit.

circuit is monitored by means of an amplifier and loud speaker bridged across the circuit through a high impedance coil. This, of course, indicates to the attendant the general status of each connection as to transmission in both directions. Monitoring by means of a headset is also available and is usually resorted to in those instances where conditions require close attention to the adjustment of the apparatus to insure a satisfactory connection.

When a call originates on land, the toll operator connects the land line from the shore subscriber to the radio circuit, designated "Harbor," by means of a regular toll cord circuit. This operation, which requires the insertion of a plug in the jack of the radio circuit, automatically starts the transmitter at Point Edwards. It also indicates to the radio attendant by means of an alarm that a connection is being made and that his attention is required to ascertain whether any adjustments under his control are necessary for proper operation of the circuit. The switchboard operator then rings with a 1000-cycle signal which registers as an attention call to all ships which have their receivers operating. Selective ringing is available but requires that ships as well as station be properly equipped. The particular vessel to which a connection is to be made is called by name and station letter. This ship then starts its transmitter and reports and the two-way connection is established.

On calls originating from a ship, the boat's transmitter is energized and its carrier operates the codan circuit associated with the land station receiver. An auxiliary circuit actuated by the codan functions to signal the Seattle long distance operator by means of the regular toll line lamp signal. It also signals the radio attendant to stand by in the manner described above. The switchboard operator in responding to the signal inserts the answering plug of a cord circuit in the "Harbor" jack which energizes the land transmitter. Two-way telephone contact is then established with the ship, and from this point on the regular traffic operating procedure is followed in connecting the ship with the desired telephone station.

Ship Sets for Harbor Ship-to-Shore Service *

By H. N. WILLETS

A STUDY of conditions on board fishing trawlers and harbor craft indicates that the radio units must be small in size, rugged in construction and adaptable to remote control. Further, the units must permit operation by non-technical personnel and require a minimum of power for their proper functioning.

The striking similarity between these conditions and those already encountered in the furnishing of radio telephone equipment to aircraft permitted the use with small modification of a radio system already proved by years of use on this country's major airlines.

The ship transmitter, coded the 13-A, has an output of 50 watts. Its size is $13\frac{3}{4}'' \times 18\frac{7}{8}'' \times 10''$, its weight about 34 pounds.

The tube complement comprises a 5-watt audio-frequency amplifier tube, a 5-watt oscillator tube which is controlled by a quartz crystal oscillating at one-half the desired frequency and connected to the grid circuit of the oscillator tube, and a 50-watt screen grid tube as a first radio-frequency amplifier. The coupling from the oscillator is supplied by a radio-frequency transformer which freely passes the first harmonic of the quartz plate to drive the first radio-frequency amplifier. The output stage consists of two similar tubes in parallel to form a second radio-frequency amplifier. The two radio-frequency amplifier stages are coupled by a transformer which also acts as a band-pass filter and freely passes the output frequency. The two radio-frequency transformers are mounted in a single plug-in unit. The audio amplifier modulates the screen bias on the first and second radio-frequency amplifiers, giving substantially one hundred per cent modulation. The filament of each of the three radio-frequency amplifier tubes is in series with a ballast tube to protect them from fluctuations in the power supply.

The quartz crystal oscillator has the recently developed low temperature coefficient cut and the crystal holder is arranged with a heater which operates only should the temperature drop below zero Centigrade. The transmitter is arranged for one, two or three carrier frequencies, each requiring a separate crystal. If more than one frequency is provided, change from one to another is accomplished by a

* Digest of discussion prepared for Pacific Coast Convention, A.I.E.E., August, 1935.

single mechanical control. Remote selection of frequency may be by a flexible shaft or by a pull-wire arrangement.

The companion receiver, coded the 12-C, is a compact unit weighing about 16 pounds. It has a superheterodyne circuit employing six tubes and permits a quick shift between two fixed frequencies. High sensitivity and unusual selectivity are obtained. The antenna circuit is series tuned. The beating oscillator and modulator are combined in one tube while a quartz crystal assures the correct beating frequency. It is a plug-in unit similar to that used in the transmitter and ground to the carrier frequency plus the intermediate frequency of 385 kilocycles. Two stages of intermediate radio frequency are used. A duo diode triode tube is used as detector and first stage of audio-frequency amplification. The second audio and output tube is a pentode. Automatic volume control provides uniformity of output signal. A ballast tube protects the filaments against voltage fluctuations in the power supply.

Space is provided for two crystals and remote selection of either frequency may be provided by flexible shaft or pull wire. The controls are tied together through the receiver and transmitter mountings so that the frequency of both may be shifted simultaneously from a remote point.

While in the aircraft system the air pilot has his receiver always in use as he checks periodically on his route, it is obvious that the marine pilot would not require this constant contact with the shore station. To eliminate the necessity of loud speaker monitoring, selective ringing is available. It works from the output of the radio receiver with a conventional type selector which is stepped up in response to the impulses received from a dial actuating the shore transmitter. With this arrangement, any particular boat may be called, a bell announcing the incoming call.

For convenience of installation, the transmitter, receiver and selective ringing unit are mounted in a small metal cabinet. The cabinet provides protection to these units and incorporates the remote frequency change equipment and a junction box. All the connections between the unit mountings in the cabinet are permanent and terminate in the junction box. Provision is made in the base of the cabinet for two cables, one to the control equipment and the other to the power supply. The cabinet is provided with rubber feet which permit its installation in places subject to considerable vibration. One of the features of the installation is the ready access to any of the units for maintenance and test. The top and front of the cabinet may be removed and each unit, being "plug-in" mounted, is easily removable

without disturbing connections. Adequate ventilation is provided as well as protection against direct splashing from above.

The power supply includes 12 volts for the filaments of the transmitter, receiver and selective signaling unit; 200 volts for the plate supply of the radio receiver, and 1,050 volts for plates of the tubes in the radio transmitter. For efficient operation these voltages must be held to reasonably close limits and experience has shown that wide



Fig. 1—Cabinet housing radio equipment; front panel removed showing left top radio transmitter, right top radio receiver, ringing unit beneath.

variations in the boat's power supply are to be expected. Therefore, two forms of power supply are generally recommended:

1. Motor generators for both transmitter and receiver;
2. A 12-volt storage battery and charging unit with dynamotors for the plate supply of the transmitter and receiver.

In the case of motor generators, two machines are provided, each consisting of a driving motor and a double winding generator mounted on a common cast iron base. The machines are of splash proof con-

struction. The driving motors depend upon the source of power of each installation. Where the variation of the d-c. supply is greater than ± 3 per cent, a speed regulator is provided.

The generators of the double-voltage type have both a high and a low-voltage winding. One supplies 10 amperes at 13 volts and 0.1 ampere at 200 volts to the receiver; the other 15 amperes at 13 volts and 0.35 ampere at 1,050 volts to the transmitter. The units are equipped with filters to prevent disturbance in the radio equipment.

When voltage variations on shipboard are extreme, a 12-volt storage battery may provide the filament source as well as operate both a 200-volt and a 1,050-volt dynamotor. The battery charger may be of the automatic type or arranged for periodic charging.

The control unit is designed for installation at any convenient location. It provides facilities for remote starting and stopping the radio apparatus. It consists essentially of a single master-control switch, a telephone handset, call bell, a volume control for the receiver in the handset, and an antennæ meter for visual indication of the transmitter operation. A small lamp on the unit indicates when the receiver is in operation. To call or answer a call, one operates the single master-control switch and removes the handset from its hook switch. This controls the power to the radio transmitter. A push button in the handset handle automatically switches from receive to transmit position.

As an adjunct to the two-way radio telephone system, a radio compass has been developed which is essentially composed of a highly sensitive radio receiver and of a loop and "sense-indicating" antenna, and is extremely simple to operate. The accuracy of the arrangement is in the order of one degree when 200 to 300 miles from the marine radio beacon. Taking bearings on marine radio beacons is facilitated by a scale on the receiver calibrated directly in kilocycles and a chart on the receiver panel listing the radio beacons. After tuning, the loop is rotated by a hand wheel until the needle of a center-reading meter points to zero. A unique arrangement of the antenna permits a scale on the shaft of the loop to indicate the azimuth between the ship's heading and the direction of the received radio beacon when the meter reads zero. As the loop is turned to the right or left, the needle on the meter likewise swings to the right or left. There is no doubt as to the "sense" of the bearing as the meter gives positive visual indication when the loop is rotated. By plotting the direction of the boat from two or more radio beacons on the nautical chart, the intersection of these plotted lines determines the ship's position.

The direction-finding receiver employs a superheterodyne circuit,

covering the frequency band from 242 to 515 kilocycles. A single control tunes the entire frequency range providing coverage of all the marine beacon stations and all the ship telegraph frequencies. The power supply for the receiver is the same as that for the other radio units. A loud speaker may be provided for identifying the station and to facilitate tuning. It may also be used for taking a bearing under severe static conditions. The loop assembly consists of a statically shielded loop winding, a supporting pedestal with shaft, a mounting flange, a hand-wheel, an azimuth scale, a compensator, an indicating meter, slip rings, lamp and terminal strip.

An Electromechanical Representation of a Piezoelectric Crystal Used as a Transducer *

By W. P. MASON

THE equivalent electrical representation of a piezoelectric crystal when used as an element in an electrical circuit has been discussed by several investigators,[1] who have arrived at the circuit shown on Fig. 1. Apparently, however, no circuit has been evolved for repre-

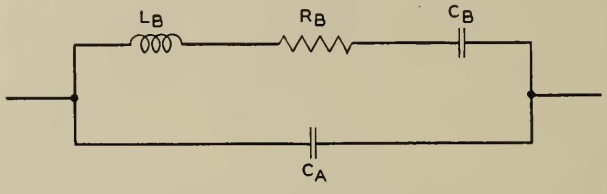


Fig. 1—Electrical representation of a piezoelectric crystal.

senting a crystal when it is used as a transducer to couple electrical circuits to mechanical systems. Since such crystals[2] are used in loud speakers, microphones, supersonic radiators, and other apparatus, it is a matter of importance to obtain such a representation. This paper discusses such an equivalent circuit and relates the elements to the mechanical, electrical, and piezoelectric constants of the material. When used as a purely electrical circuit, this representation reduces to that of Fig. 1.

When piezoelectric crystals are used to drive external mechanical systems, the modes of motion most often used are longitudinal vibrations perpendicular or parallel to the applied electric field. Accordingly, the elements of the equivalent network are derived for these cases only. The network can, however, represent any type of motion driving a load just as the network of Fig. 1 can represent the crystal for any type of motion.

Let us consider first the case of a crystal vibrating perpendicularly to

* Digest of an article appearing in the *Proc. I. R. E.* for October, 1935.

[1] W. G. Cady, *Phys. Rev.*, **XXIX**, 1 (1922); *Proc. I. R. E.*, **X**, 83 (1922). K. S. VanDyke, Ab. 52, *Phys. Rev.*, June, 1925; *Proc. I. R. E.*, June, 1928. D. W. Dye, *Proc. Phys. Soc.* (*London*), **XXXVIII**, (5), pp. 399–453. P. Vigoreaux, *Phil. Mag.*, December, 1928, pp. 1140–53.

[2] A. M. Nicolson, *Proc. A. I. E. E.*, **38**, 1315–1333, 1919. E. B. Sawyer, *Proc. I. R. E.*, **19**, No. 11, p. 2020, November, 1931. S. Ballantine, *Proc. I. R. E.*, **21**, No. 10, p. 1399, October, 1933.

the direction of the applied field. Two subdivisions of this case are usually of interest, the first when the crystal is supported at its center and drives two symmetrical loads, and the second when the crystal is supported on one end and drives a load on the other end. The symmetrical case is considered first.

By employing the well known analogies between electrical and mechanical systems, it is possible to obtain a simple network, expressed in terms of electrical symbols, which represents the properties of a piezoelectric crystal. In this representation, force is the analogue of voltage, mechanical displacement of electrical charge, and velocity of electrical current. When the electrodes are attached to the crystal faces, the equivalent network of the crystal is shown by Fig. 2, The voltage E

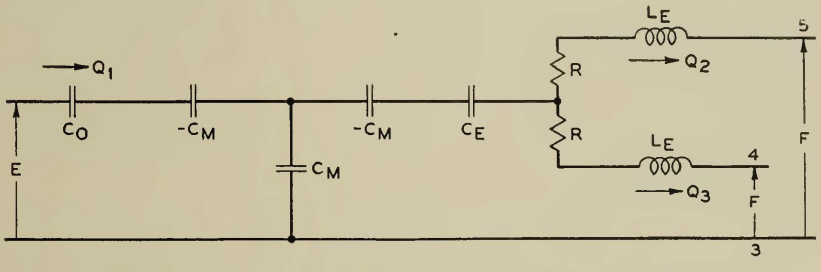


Fig. 2—Electromechanical representation of a symmetrical piezoelectric crystal.

is the voltage applied across the plates of the crystal, the force F is the force applied to each end of the symmetrical crystal, Q_1 is the electrical charge flowing in the wires connected to the crystal and Q_2 and Q_3 are the mechanical displacements of the ends of the crystal which are equal on account of the symmetry of the crystal.

The constants of the crystal can be evaluated by considering limiting cases. The capacitance C_0 is the electrostatic capacitance of the crystal clamped. The compliance C_E is the mechanical compliance of the crystal. In c.g.s. electrostatic and mechanical units, these have the values

$$C_0 = \frac{Kl_m l_0}{4\pi l_e}; \qquad C_E = \frac{sl_m}{l_e l_0}, \tag{1}$$

where K is the dielectric constant of the crystal clamped, l_e the dimension of the crystal in centimeters perpendicular to the surfaces of the electrodes, l_m the length of the crystal in the direction of vibration, l_0 the length of the third axis, and s the modulus of compliance of the crystal (the inverse of Young's modulus) along the axis of vibration. The inductance L_E represents the mass reaction of half the crystal. At low

frequencies this will be equal to half the mass of the crystal, but at higher frequencies will be less due to the fact that the crystal does not move as a whole.[3] In order to resonate with the compliance C_E at the mechanical resonance frequency of the crystal, L_E must equal

$$L_E = \frac{2l_c l_m l_0 \rho}{\pi^2}, \tag{2}$$

where ρ is the density of the crystal. The resistances R shown include the dissipation due to internal friction, supersonic radiation from the ends of the crystal, friction at the point of support and all other sources of dissipation.

If F_0 is the force required to keep the crystal from expanding when an electric charge Q_1 is applied to the crystal then C_M the mutual capacitance-compliance of the crystal is equal to

$$C_M = Q_1/F_0. \tag{3}$$

Similarly if E_0 is the open circuit voltage for a given expansion $(Q_2 + Q_3)$ of the crystal then

$$C_M = \frac{Q_2 + Q_3}{E_0}. \tag{4}$$

In order to evaluate C_M in terms of the piezo-electric coefficient d, it is necessary to find the displacement for a free crystal when an electrical potential is applied to the crystal. Short circuiting the network of Fig. 2 on the mechanical ends and setting $F = 0$, we find

$$Q_2 + Q_3 = \frac{E\left[\frac{C_0 C_E}{C_M}\right]}{1 - \frac{C_0 C_E}{C_M{}^2}} = \frac{Ek\sqrt{C_0 C_E}}{1 - k^2}, \tag{5}$$

where k is the coefficient of coupling between the electrical and mechanical system is defined by the equation

$$k = \sqrt{C_0 C_E}/C_M. \tag{6}$$

Use is now made of the piezo-electric equation

$$e = dV, \tag{7}$$

[3] Strictly speaking the value of C_E is also a function of frequency, but at the first resonance it can be shown that it differs from its static value by the factor $8/[\pi^2 - k^2(\pi^2 - 8)]$ and $L_E = l_c l_m l_0 \rho/4$. For a highly coupled crystal, this factor does not differ much from unity, and hence in the interest of simplicity, the variations of C_E have been neglected.

where e is the strain (elongation per unit length) produced in the crystal by an applied potential gradient V. Comparing (5) with (7) and noting that $Q_2 + Q_3 = el_m$, and $V = E/l_c$, we have on the insertion of the values for C_0 and C_E from (1)

$$d = \frac{k\sqrt{\frac{Ks}{4\pi}}}{1 - k^2}. \tag{8}$$

Solving for k we find

$$k = \frac{1}{2d}\sqrt{\frac{Ks}{4\pi}}\left[-1 + \sqrt{1 + \frac{16\pi d^2}{Ks}} \right] \doteq d\sqrt{\frac{4\pi}{Ks}}, \tag{9}$$

when $16\pi d^2/Ks$ is a small quantity as it is for quartz.

When the crystal is used as an element in an electrical network, and allowed to vibrate freely, the force F of Fig. 2 can be set equal to zero and the network short-circuited. Solving for the impedance on the electrical side we find

$$Z_c = \frac{-j(1 - k^2)}{2\pi f C_0}\left[\frac{1 - f^2/f_1^2 + j/q(1 - k^2)}{1 - f^2/f_2^2 + j/q}\right], \tag{10}$$

where $f_2^2 = f_A^2$; $f_1^2 = f_A^2(1 - k^2)$ (f_A being the natural mechanical resonance frequency of the crystal), and q is the ratio of the reactance of the condenser C_E to the resistance $R/2$ or

$$q = 2/2R\pi f_A C_E. \tag{11}$$

It is easily shown that the impedance Z_c is also the impedance of the network of Fig. 1 if [4]

$$\begin{aligned} C_A &= C_0; \\ C_B &= C_0k^2/(1 - k^2); \\ L_B &= 1/4\pi^2 f_A^2 k^2 C_0; \\ R_B &= 1/2\pi f_A C_0 k^2 q. \end{aligned} \tag{12}$$

Hence the representation in Fig. 2 reduces to the well known Fig. 1, when the crystal is free to vibrate.

A network representing the second case, when one end is supported with the other end used to drive a load, is shown on Fig. 3. The method of deriving the constants is the same and all of the constants

[4] If account is taken of the variation of C_E and L_E with frequency, the elements are $C_A = C_o$; $C_B = \frac{8k^2C_o}{\pi^2(1 - k^2)}$; $L_B = \frac{1}{32k^2f_A^2C_o}$; $R_B = 1/2\pi f_A C_o k^2 q$ where $q = \frac{1}{\pi R C_E f_A}$.

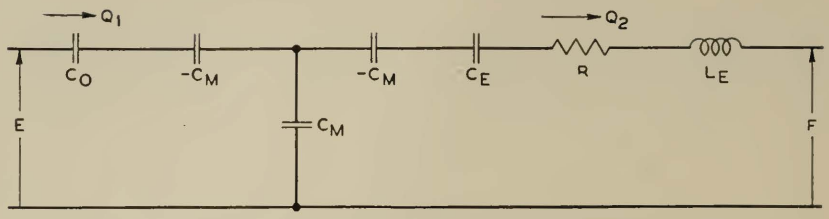


Fig. 3—Electromechanical representation of a piezoelectric crystal clamped on one end.

are the same except L_E, which is twice as large since twice the mass is moved from the clamping position.

When the direction of motion is parallel to the direction of the applied field, the same networks hold but the elements have different lengths entering into their determinations. The direction of the applied field and of the motion is designated by l_e. The other two axes are still designated by l_m and l_0. The resulting constants are

$$C_0 = \frac{Kl_m l_0}{4\pi l_e}; \qquad C_E = \frac{sl_e}{l_m l_0}; \qquad C_M \doteq \frac{Ks}{4\pi d};$$
$$L_{E_1} = \frac{2\rho l_e l_0 l_m}{\pi^2}; \qquad L_{E_2} = \frac{4\rho l_e l_0 l_m}{\pi^2}, \tag{13}$$

where L_{E_1} is the mechanical inductance for the symmetrical case (Fig. 2) and L_{E_2} for the dissymmetrical case (Fig. 3).

A simple example of the use of Fig. 2 in determining the effect of a mechanical load on the impedance of a crystal is the problem of finding out how much the energy radiated by a crystal to the surrounding air affects the decrement of a quartz crystal vibrating longitudinally. When a crystal vibrates, energy is radiated to the surrounding medium by the motion of the ends of the crystal. If the dimensions of the ends of the crystal are comparable to a wave-length or greater—which they will ordinarily be for a crystal vibrating at a high frequency—it is well known [5] that the radiating surface experiences a resistance to motion equal to

$$R_R = \rho_A b \text{ (mechanical ohms)} \tag{14}$$

per square centimeter, where ρ_A is the density of the medium and b the velocity of propagation. For air R_R is about 41 ohms per square centimeter. Hence the equivalent circuit for a crystal vibrating in air is Fig. 2 terminated at the terminals 3–4 and 3–5 by the mechanical

[5] See "Theory of Vibrating Systems and Sound," I. B. Crandall, Chap. 4, D. Van Nostrand Co.

resistances

$$R_R = 41 l_e l_0. \tag{15}$$

If all other sources of dissipation were eliminated, the radiation resistance would produce a limiting value for the decrement of a crystal which may be calculated as follows. From equation (11), the value of q for the mechanical system is

$$q = \frac{1}{\pi f_A R_R C_E} = \frac{2\sqrt{\rho/s}}{41\pi} \tag{16}$$

on inserting the values of R_R, C_E, and $f_A = 1/2l_m\sqrt{\rho s}$. Since $\rho = 2.65$ and $s = 1.27 \times 10^{-12}$ for a perpendicularly cut quartz crystal

$$q = 2.24 \times 10^4. \tag{17}$$

The decrement of a crystal in terms of the circuit of Fig. 1 is

$$\delta = \frac{R_B}{2f_A L_B}. \tag{18}$$

Inserting the values of footnote (4) we find

$$\delta = \frac{8}{\pi q} = 1.14 \times 10^{-4}. \tag{19}$$

Van Dyke [6] has measured the limiting value of the decrement of a perpendicularly cut quartz crystal vibrating in air and finds it to be 1.26×10^{-4}. Since the residual losses were about 5 per cent of the radiation losses, this agrees well with the value found in equation (19).

The equivalent circuits of Figs. 2 and 3 may also be used as the basis of design for mechanical systems, such as loud speakers, microphones and supersonic radiators.

[6] *Proc. I. R. E.*, April, 1935.

Abstracts of Technical Articles from Bell System Sources

Some Methods for Making Resonant Circuit Response and Impedance Calculations.[1] H. T. Budenbom. This article presents a series of short-cut methods for the computation of the amplitude and phase-response characteristics of one- and two-mesh circuits; the extension of the method to three- and four-mesh networks is carried out formally. It also treats the impedance-frequency characteristics of singly resonant circuits. The plan of attack in the single-circuit impedance and multi-mesh response cases is to express the desired circuit or transfer impedance as a polar numeric multiplied by a sizing constant which turns out to be a simple reactance element of the circuit.

A Rapid Method for the Determination of Sulfur in Ferromagnetic Alloys.[2] B. L. Clarke, L. A. Wooten, and C. H. Pottenger. A method is described for the determination of sulfur in ferrous alloys of high-nickel content, wherein the sample is heated in hydrogen at 1100° to 1200° C., and the liberated hydrogen sulfide absorbed in ammoniacal cadmium nitrate solution. The method is generally applicable to iron-nickel-cobalt alloys.

The precision of the method for the range of 0.005 to 0.020 per cent sulfur is shown to be ± 0.001 per cent sulfur on the basis of a 10-gram sample.

The Newly Discovered Elementary Particles.[3] Karl K. Darrow. During the three years following the discovery of heavy hydrogen in the latter part of 1931, four additional elementary particles of matter lighter than the alpha particle (nucleus of the ordinary helium atom) have been discovered where only two were known previously. This is considered extraordinary even for the present rapid pace of physics which sometimes requires physicists to revise their fundamental concepts of matter almost overnight. These five newly discovered particles formed the subject matter of a highly interesting and instructive address delivered recently by Doctor Darrow at a meeting of the American Institute of Electrical Engineers' New York Section; the essential substance of this illuminating address is presented in the published article.

[1] *Radio Engg.*, August, 1935.
[2] *Indus. & Engg. Chem.*, Analytical Edition, July 15, 1935.
[3] *Elec. Engg.*, August, 1935.

Magnetic Hysteresis at Low Flux Densities.[4] W. B. Ellwood. The energy loss per cycle in a ferromagnetic material subjected to small alternating fields is sometimes separated into three parts: the first due to eddy current loss, the second to hysteresis presumed to follow Rayleigh's laws at these flux densities; the origin of the remainder is unknown and is the subject of much controversy; it has been variously termed "magnetic viscosity," "after effect," and "square law hysteresis." In studying energy loss in a ring of compressed iron dust, hysteresis loops have been measured ballistically by a new method with a relative error in B_m as low as 0.01 per cent. The range in maximum flux density is from 2 to 100 gauss. The loops are lenticular and very slender, B_m being about 1500 times the remanence for the smallest loop. The smaller loops are at flux densities considerably below those investigated by Rayleigh. His findings as to variation of area, of remanence, and of permeability with loop amplitude are confirmed and extended through the new range of flux densities, though the shape of the loops is not as simple as that he proposed. The energy loss per cycle is proportional to B_m^3 while the remanence is proportional to B_m^2. The ballistic measurements have been compared with a-c. bridge measurements on the same specimen. The loss of unknown origin is not included in the hysteresis loss measured ballistically. Comparison is made between the third harmonic induced voltage computed by a Fourier series analysis of the ballistic loops and the harmonic actually generated by the specimen. Agreement is found between the measured and the computed values. Possible explanations of the discrepancy between the ballistic observed energy losses and a-c. findings are discussed.

A Fugue in Cycles and Bels.[5] John Mills. Engineers who are questioned by their musical acquaintances about electrical transmission and what it is likely to do to the art are likely to find an explanation, even in the simplest of terms, shooting over the heads of their audience. The reason lies not in the inherent difficulty of the concepts, but in their number, which exceeds the power of memory to retain as unrelated facts. In this volume John Mills has strung his facts together on the thread of logical relationship, but he has tied them into his readers' existing knowledge by many a deft touch of anecdote or humor. Frequency, with its relation to harmony and discord, opens the volume; there follows a physical picture of vibrations in various media and how one form is transformed into another. Pitch and intensity are next considered. An entire section is devoted to tele-

[4] *Physics*, July, 1935.
[5] Published by D. Van Nostrand Company, Inc., New York, N. Y., 1935.

phonic studies of hearing; it includes such topics as the ear, the amplifier, transmission, loudness, overloading and distortion.

A third section, which discusses the electrical future for music, will most intrigue the musician. After surveying the present state of the art of pick-up, transmission, recording and reproduction of music, the author visualizes some of the possibilities which might be most immediately realized: reproduction in auditory perspective, electrical music and electrical aids in teaching.

To permit a rapid survey by the general reader, Mr. Mills has withdrawn all tables and graphs from the main text and has grouped them, with necessary explanations, at the end. This section forms a useful compendium of numerical information on frequencies, decibels, thresholds, response curves of microphones and loud speakers, loudness and energy levels, masking and the like.

The engineer will find "A Fugue in Cycles and Bels" interesting and easy to read, and a source of data not always available, and he can recommend it to those musicians whose serious interest in their art leads them to delve into its physical basis.

Influence of Experimental Technique on the Measurement of Differential Intensity Sensitivity of the Ear.[6] H. C. MONTGOMERY. The lack of agreement among previous measurements of differential intensity sensitivity indicates that the values obtained depend to a large extent on the experimental conditions. The relative importance of various factors is indicated, and a procedure is suggested which was designed to give the smallest possible values of differential intensity sensitivity. Intensive measurements made by this method upon a single subject, using a pure tone of 1000 cycles, gave values consistently smaller than any previously reported. There is no sharp division between intensity changes which can be perceived and those which cannot. The response of the subject is essentially variable and can be described only by statistical methods.

Diurnal and Seasonal Variations in the Ionosphere During the Years 1933 and 1934.[7] J. P. SCHAFER and W. M. GOODALL. The most important results of daily ionospheric measurements made at Deal, New Jersey, latitude 40° 15′ N., longitude 74° 02′ W., over the period from March, 1933, to May, 1934, are given in this paper and may be summarized as follows:

1. There was a definite correlation between the noon value of ionic density of the F_1 region and magnetic disturbances, a decrease in ionic density being obtained on magnetically disturbed days.

[6] *Jour. Acous. Soc. Amer.*, July, 1935.
[7] *Proc. I. R. E.*, June, 1935.

2. The noon value of ionic density of the E and F_1 region attained a maximum in summer and a minimum in winter whereas the reverse condition, of minimum in summer and maximum in winter, was found for the F_2 region.

3. The time of maximum ionic density of the F_2 region varied with the seasons of the year, occurring near noon in winter and near sunset in summer.

The paper also shows a series of virtual height contour maps for the four seasons of the year.

Utilization of Electrical Resistance Measurements for Controlling the Composition of Alloys.[8] E. E. SCHUMACHER and L. FERGUSON. To stimulate interest in the more general use of the electrical resistance of solid solutions for indicating the composition of alloys, details concerning one application are set forth in this paper. While for other systems certain elements of the method as outlined here might have to be modified, it appears that the general principle could be advantageously utilized in many instances. This paper presents, therefore, a practical resistance method of supplementing ordinary chemical analyses in the determination of the antimony content of lead-antimony alloys ranging at least from 0.4 to 1.1 per cent. It has been found that, within this range, a 1 per cent change in resistance is produced by a change in antimony content of less than 0.1 per cent and that the maximum deviation in resistance data may reasonably be made so small as to correspond to not more than $\pm$ 0.01 per cent antimony.

[8] *Metals and Alloys*, June, 1935.

Contributors to this Issue

Leonard Gladstone Abraham, B.S., 1922, M.S., 1923, University of Illinois. American Telephone and Telegraph Company, Department of Development and Research, 1923–34; Bell Telephone Laboratories, 1934–. Mr. Abraham has been engaged in transmission development work on toll telephone systems.

Austin Bailey, A.B., University of Kansas, 1915; Ph.D., Cornell University, 1920; Instructor in Physics, Cornell University, 1915–18; Signal Corps, U. S. A., 1918–19; Assistant Professor of Physics, University of Kansas, 1921–22. American Telephone and Telegraph Company, Department of Development and Research, 1922–34; Bell Telephone Laboratories, 1934–. Dr. Bailey's work has been largely along the line of methods for making radio transmission measurements and of long-wave radio problems.

F. A. Gifford, B.S., Tufts College, 1920. General Electric Company, Lynn, Massachusetts, 1920–22; American Radio and Research Corporation, Medford, Massachusetts, 1922; New England Telephone and Telegraph Company, Engineering Department, Boston, 1922–. Mr. Gifford has been engaged on various types of telephone engineering work, devoting a large part of his time during the past four years to problems in connection with the Boston Marine Radio Telephone Service.

Earl B. Hansen, B.S. in Electrical Engineering, University of California, 1920. Pacific Telephone and Telegraph Company, 1920 to present date. Latterly, Mr. Hansen's duties have been in the field of toll transmission.

H. E. Haring, B.S., Franklin and Marshall College, 1916; M.A., Princeton University, 1917. Assistant Chemist, Ordnance Department, U. S. Army, 1917–19; Associate Chemist, U. S. Bureau of Standards, 1919–28; Electrochemist, Victor Talking Machine Company, 1928–29; Bell Telephone Laboratories, 1929–. Since 1919 Mr. Haring has been engaged in electrochemical research in connection with storage batteries and other electrochemical apparatus, electrodeposition, and corrosion.

H. R. Huntley, B.S., University of Wisconsin, 1921. Wisconsin Telephone Company, Engineering Department, 1917–30; American

Telephone and Telegraph Company, Department of Operation and Engineering, 1930–. Mr. Huntley's work has been concerned principally with transmission and inductive coordination matters.

Dr. Jewett, as Vice President of the American Telephone and Telegraph Company and President of the Bell Telephone Laboratories, needs no introduction to *Technical Journal* readers.

S. A. Levin, E.E., Chalmers Technical Institute, Gothenburg, 1919; Technische Hochschule, Berlin, 1920–21; Technische Hochschule, Dresden, 1921–23. Radio Department, General Electric Company, Schenectady, N. Y., 1923–26; Engineering Department, National Electric Light Association, New York, N. Y., 1926–30; Bell Telephone Laboratories, 1930–. Mr. Levin's work has to do with the development of high-frequency measuring equipment for carrier systems.

Frederick B. Llewellyn, M.E., Stevens Institute of Technology, 1922; Ph.D., Columbia University, 1928. Western Electric Company, 1923–25; Bell Telephone Laboratories, 1925–. Dr. Llewellyn has been engaged in the investigation of special problems connected with radio and vacuum tubes.

W. P. Mason, B.S. in Electrical Engineering, University of Kansas, 1921; M.A., Columbia University, 1924; Ph.D., 1928. Bell Telephone Laboratories, 1921–. Dr. Mason has been engaged in investigations on carrier transmission systems and more recently in work on wave transmission networks, both electrical and mechanical.

R. B. Meader, B.S., University of New Hampshire, 1921. New England Telephone and Telegraph Company, 1922–. Since 1930 Mr. Meader has worked chiefly on radio telephone field strength surveys and other activities leading to the establishment of the Green Harbor Radio Telephone Station and the development and operation of the Boston Marine Radio Telephone Service.

E. J. O'Connell, B.S., Northwestern University, 1924. American Telephone and Telegraph Company, Long Lines Department, 1924–25; Illinois Bell Telephone Company, Engineering Department, 1925–28; American Telephone and Telegraph Company, Department of Operation and Engineering, 1928–. Mr. O'Connell's work has been concerned principally with inductive coordination matters.

Liss C. Peterson, E.E., Chalmers Technical Institute, Gothenburg, 1920; Technische Hochschule, Charlottenburg, 1920–21; Technische Hochschule, Dresden, 1921–22; Signal Corps, Swedish Army, 1922–23.

American Telephone and Telegraph Company, 1925–30; Bell Telephone Laboratories, 1930–. Mr. Peterson is engaged in the study of modulation and other problems connected with high-frequency carrier systems.

G. S. Phipps, B.S. in Electrochemical Engineering, Pennsylvania State College, 1930. Bell Telephone Laboratories, 1930–. Mr. Phipps has been engaged principally in the metallurgical investigation of lead, aluminum and zinc alloys.

E. E. Schumacher, B.S., University of Michigan; Research Assistant in Chemistry, 1916–18. Engineering Department, Western Electric Company, 1918–25; Bell Telephone Laboratories, 1925–. As Assistant Research Metallurgist, Mr. Schumacher is in charge of a group whose work relates largely to research studies on metals and alloys.

U. B. Thomas, Jr., B.S. in Chemistry, William and Mary, 1929. Bell Telephone Laboratories, 1929–. Mr. Thomas has been engaged principally in the study of base metal contacts, storage batteries, and allied electrochemical problems.

Howard M. Thomson, B.S. in E.E., University of Washington, 1930. Bell Telephone Laboratories, 1930–32; American Telephone and Telegraph Company, Department of Development and Research. 1932–34; Bell Telephone Laboratories, 1934–. Mr. Thomson's work has been principally on studies of long-wave radio transmission and field tests on an experimental transmitting wave antenna.

H. N. Willets, Western Electric Company, Engineering Inspection Department, 1917–25; Philadelphia Instrument Shop, 1921–25; Supply Department, 1925–26. Graybar Electric Company, 1926–28; Western Electric Company, 1928–. Mr. Willets is engaged in the promotion of aircraft radio, broadcasting, and ship-to-shore radiophone and harbor craft radiophone.