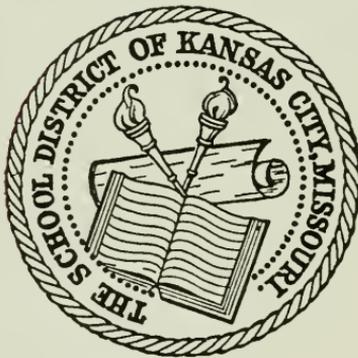


Bound
Periodical **1164153**

**Kansas City
Public Library**



This Volume is for
REFERENCE USE ONLY

From the collection of the

o P^{z n}re^minger^a
v L^{t p}ibrary

San Francisco, California
2008

THE BELL SYSTEM TECHNICAL JOURNAL

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

EDITORS

R. W. KING

J. O. PERRINE

EDITORIAL BOARD

M. R. SULLIVAN

O. E. BUCKLEY

O. B. BLACKWELL

M. J. KELLY

H. S. OSBORNE

A. B. CLARK

J. J. PILLIOD

S. BRACKEN

TABLE OF CONTENTS

AND

INDEX

VOLUME XXIII

1944

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

PRINTED IN U. S. A.

THE BELL SYSTEM TECHNICAL JOURNAL

VOLUME XXIII, 1944

Table of Contents

JANUARY, 1944

The Discernibility of Changes in Program Band Width— <i>D. K. Gannett and Iden Kerney</i>	1
Use of the Etch Technique for Determining Orientation and Twinning in Quartz Crystals— <i>G. W. Willard</i>	11
Modes of Motion in Quartz Crystals, the Effects of Coupling and Methods of Design— <i>R. A. Sykes</i>	52
Response of a Linear Rectifier to Signal and Noise— <i>W. R. Bennett</i>	97
Dielectric Constants and Power Factors at Centimeter Wave-Lengths— <i>Carl R. Englund</i>	114

APRIL, 1944

Indicial Response of Telephone Receivers— <i>E. E. Mott</i>	135
Theoretical Analysis of Modes of Vibration for Isotropic Rectangular Plates Having All Surfaces Free— <i>H. J. McSkimin</i>	151
Principles of Mounting Quartz Plates— <i>R. A. Sykes</i>	178
The Magnetically Focused Radial Beam Vacuum Tube— <i>A. M. Skellett</i>	190

JULY, 1944

Effect of Telegraph Distortion on the Margins of Operation of Start-Stop Receivers— <i>W. T. Rea</i>	207
The Mounting and Fabrication of Plated Quartz Crystal Units— <i>R. M. C. Greenidge</i>	234
Effects of Manufacturing Deviations on Crystal Units for Filters— <i>A. R. D'heedene</i>	260
Mathematical Analysis of Random Noise— <i>S. O. Rice</i>	282

iii

1161153

FEB 28 1945

OCTOBER, 1944

The Conquest of Distance by Wire Telephony— <i>Thomas Shaw</i>	337
Some Aspects of Powder Metallurgy— <i>Earle E. Schumacher and Alexander G. Souden</i>	422

Index to Volume XXIII

A

Alloys: Some Aspects of Powder Metallurgy, *Earle E. Schumacher and Alexander G. Souden*, page 422.

B

Band Width, Program, The Discernibility of Changes in, *D. K. Gannett and Iden Kerney*, page 1.
Bennett, W. R., Response of a Linear Rectifier to Signal and Noise, page 97.

C

Crystal Units for Filters, Effects of Manufacturing Deviations on, *A. R. D'heedene*, page 260.
Crystal Units, Plated Quartz, The Mounting and Fabrication of, *R. M. C. Greenidge*, page 234.
Crystals, Quartz, Use of the Etch Technique for Determining Orientation and Twinning in, *G. W. Willard*, page 11.
Crystals, Quartz, Modes of Motion in, the Effects of Coupling and Methods of Design, *R. A. Sykes*, page 52.
Crystals: Theoretical Analysis of Modes of Vibration for Isotropic Rectangular Plates having All Surfaces Free, *H. J. McSkimin*, page 151.
Crystals: Principles of Mounting Quartz Plates, *R. A. Sykes*, page 178.

D

D'heedene, A. R., Effects of Manufacturing Deviations on Crystal Units for Filters, page 260.
Dielectric Constants and Power Factors at Centimeter Wave-Lengths, *Carl R. Englund*, page 114.

E

Electronics: The Magnetically Focused Radial Beam Vacuum Tube, *A. M. Skellett*, page 190.
Englund, Carl R., Dielectric Constants and Power Factors at Centimeter Wave-Lengths, page 114.

F

Filters, Effects of Manufacturing Deviations on Crystal Units for, *A. R. D'heedene*, page 260.

G

Gannett, D. K. and Iden Kerney, The Discernibility of Changes in Program Band Width, page 1.
Greenidge, R. M. C., The Mounting and Fabrication of Plated Quartz Crystal Units, page 234.

K

Kerney, Iden and D. K. Gannett, The Discernibility of Changes in Program Band Width, page 1.

L

Loading: The Conquest of Distance by Wire Telephony (A Story of Transmission Development From the Early Days of Loading To the Wide Use of Thermionic Repeaters, *Thomas Shaw*, page 337.

M

McSkimin, H. J., Theoretical Analysis of Modes of Vibration for Isotropic Rectangular Plates having All Surfaces Free, page 151.

Metallurgy, Powder, Some Aspects of, *Earle E. Schumacher and Alexander G. Souden*, page 422.

Mott, E. E., Indicial Response of Telephone Receivers, page 135.

N

Noise, Response of a Linear Rectifier to Signal and, *W. R. Bennett*, page 97.

P

Powder Metallurgy, Some Aspects of, *Earle E. Schumacher and Alexander G. Souden*, page 422.

Q

Quartz Crystal Units, Plated, The Mounting and Fabrication of, *R. M. C. Greenidge*, page 234.

Quartz Crystals, Use of the Etch Technique for Determining Orientation and Twinning in, *G. W. Willard*, page 11.

Quartz Crystals, Modes of Motion in, the Effects of Coupling and Methods of Design, *R. A. Sykes*, page 52.

Quartz Plates, Principles of Mounting, *R. A. Sykes*, page 178.

Quartz: Theoretical Analysis of Modes of Vibration for Isotropic Rectangular Plates having All Surfaces Free, *H. J. McSkimin*, page 151.

R

Radio: The Discernibility of Changes in Program Band Width, *D. K. Gannett and Iden Kerney*, page 1.

Radio: Dielectric Constants and Power Factors at Centimeter Wave-Lengths, *Carl R. Englund*, page 114.

Rea, W. T., Effect of Telegraph Distortion on the Margins of Operation of Start-Stop Receivers, page 207.

Rectifier, a Linear, Response of to Signal and Noise, *W. R. Bennett*, page 97.

Rice, S. O., Mathematical Analysis of Random Noise, page 282.

S

Schumacher, Earle E. and Alexander G. Souden, Some Aspects of Powder Metallurgy, page 422.

Shaw, Thomas, The Conquest of Distance by Wire Telephony (A Story of Transmission Development From the Early Days of Loading To the Wide Use of Thermionic Repeaters, page 337.

Skellott, A. M., The Magnetically Focused Radial Beam Vacuum Tube, page 190.

Souden, Alexander G. and Earle E. Schumacher, Some Aspects of Powder Metallurgy, page 422.

Sykes, R. A., Modes of Motion in Quartz Crystals, the Effects of Coupling and Methods of Design, page 52.

Principles of Mounting Quartz Plates, page 178.

T

Telegraph Distortion on the Margins of Operation of Start-Stop Receivers, Effect of, *W. T. Rea*, page 207.

Telephone Receivers, Indicial Response of, *E. E. Mott*, page 135.

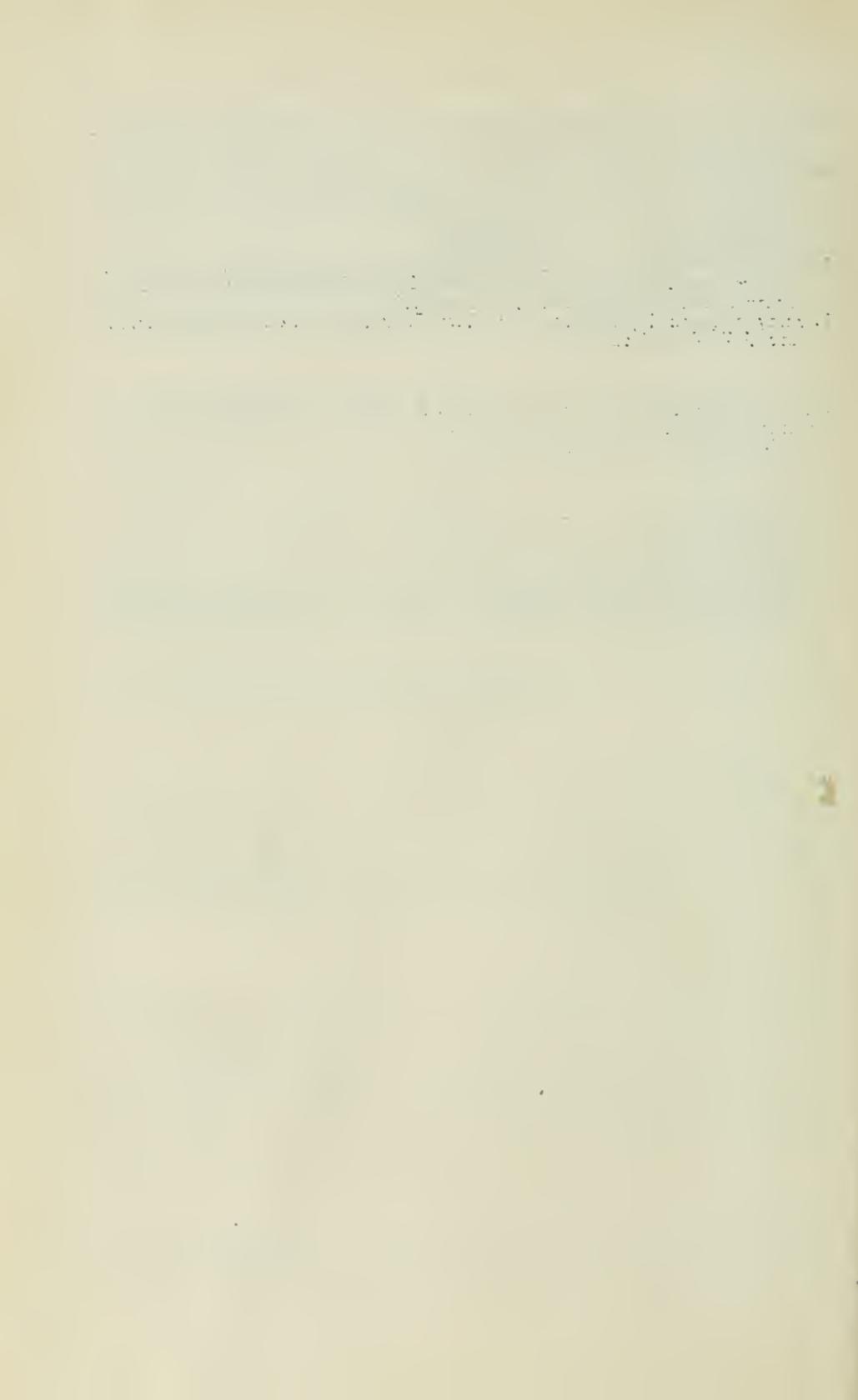
- Telephony, Wire, The Conquest of Distance by (A Story of Transmission Development From the Early Days of Loading To the Wide Use of Thermionic Repeaters), *Thomas Shaw*, page 337.
- Thermionic Repeaters: The Conquest of Distance by Wire Telephony (A Story of Transmission Development From the Early Days of Loading To the Wide Use of Thermionic Repeaters), *Thomas Shaw*, page 337.
- Transmission, Telegraph: Effect of Telegraph Distortion on the Margins of Operation of Start-Stop Receivers, *W. T. Rea*, page 207.
- Transmission Development: The Conquest of Distance by Wire Telephony (A Story of Transmission Development From the Early Days of Loading To the Wide Use of Thermionic Repeaters), *Thomas Shaw*, page 337.
- Twinning in Quartz Crystals, Use of the Etch Technique for Determining Orientation and, *G. W. Willard*, page 11.

V

- Vacuum Tube, Radial Beam, The Magnetically Focused, *A. M. Skellett*, page 190.
- Vibration for Isotropic Rectangular Plates having All Surfaces Free, Theoretical Analysis of Modes of, *H. J. McSkimin*, page 151.

W

- Wave Filters: Effects of Manufacturing Deviations on Crystal Units for Filters, *A. R. D'heedene*, page 260.
- Willard, G. W.*, Use of the Etch Technique for Determining Orientation and Twinning in Quartz Crystals, page 11.
- Wire Telephony, The Conquest of Distance by (A Story of Transmission Development From the Early Days of Loading To the Wide Use of Thermionic Repeaters), *Thomas Shaw*, page 337.



THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS
OF ELECTRICAL COMMUNICATION

The Discernibility of Changes in Program Band Width
—*D. K. Gannett and Iden Kerney* 1

Use of the Etch Technique for Determining Orientation
and Twinning in Quartz Crystals
—*G. W. Willard* 11

Modes of Motion in Quartz Crystals, the Effects of Coup-
pling and Methods of Design . . . *R. A. Sykes* 52

Response of a Linear Rectifier to Signal and Noise
—*W. R. Bennett* 97

Dielectric Constants and Power Factors at Centimeter
Wave-Lengths *Carl R. Englund* 114

Abstracts of Technical Articles by Bell System Authors 130

Contributors to this Issue 133

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

THE BELL SYSTEM TECHNICAL JOURNAL

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



EDITORS

R. W. King

J. O. Perrine

EDITORIAL BOARD

F. B. Jewett

M. R. Sullivan

O. B. Blackwell

O. E. Buckley

A. B. Clark

H. S. Osborne

S. Bracken

M. J. Kelly

F. A. Cowan



SUBSCRIPTIONS

Subscriptions are accepted at \$1.50 per year. Single copies are 50 cents each.
The foreign postage is 35 cents per year or 9 cents per copy.



Copyright, 1944
American Telephone and Telegraph Company

The Bell System Technical Journal

Vol. XXIII

January, 1944

No. 1

The Discernibility of Changes in Program Band Width*

By D. K. GANNETT and IDEN KERNEY

One of the factors that should be considered in determining how wide a transmission band is required for high fidelity broadcasting is the ability of people to perceive the effects of restricting the band to various limits, when listening to typical radio programs. Tests are described in which this was directly measured. The tests were concerned only with the physical ability to hear the differences in band width and disregarded the question of the enjoyment or aesthetic appreciation of wider bands. It is concluded that changes in band width are detectable about twice as readily with music as with speech; that one must go from 8 to 15 kc. to obtain a change as readily detected as a change from 5 to 8 kc.; and that both these changes, for speech, are just sufficient to have an even chance of being detected by listeners having experience in such tests.

THE question of how wide a frequency band it is necessary to transmit to provide high fidelity broadcasting involves consideration of a number of factors. Among these are the limits of hearing of the human ear, the spectra of program material, the aesthetic sensibilities of listeners, the effect of room noise in studios and homes, and the acoustic properties of rooms. A true engineering solution of the problem would attempt to assign numerical values to each of these factors, and then to combine them in some way to obtain a figure of merit versus band width. Sufficient information to do this in a complete and satisfactory manner is not available, however, and in practice the final answer is usually obtained by the exercise of judgment, bolstered by such technical data as can be found on the component factors.

The first two of the above factors, the limits of hearing and the spectra of program material, have been separately investigated and the results published in the technical literature by a number of experimenters. Because of the intangibles involved, however, even these two sets of data cannot readily be combined, forgetting the other factors, with complete assurance that their contribution to the answer is established. The authors, therefore, undertook a series of tests to measure directly their combined effect.

* This paper is a publication, substantially without change, of a report prepared some time ago before work non-productive to the war effort was suspended.

These experiments tested the ability of critical listeners to hear changes in band width on direct comparison when listening to representative program material. The purpose of this paper is to present the data from these tests. Similar experiments have of course been done before. The excuse for this paper is that the experiments represent a complete set of data and the analysis of the data is believed to be in such form as to be useful in further consideration of the requirements of program fidelity.

The circuit arrangements used for the tests are shown schematically in Fig. 1. The essential features are a source of program, a switch for connecting into the circuit either of two low-pass filters, and a high-quality loud-speaker. Controls for adjusting levels, volume indicators, etc., are omitted from the diagram. The arrangements included a signal visible to the

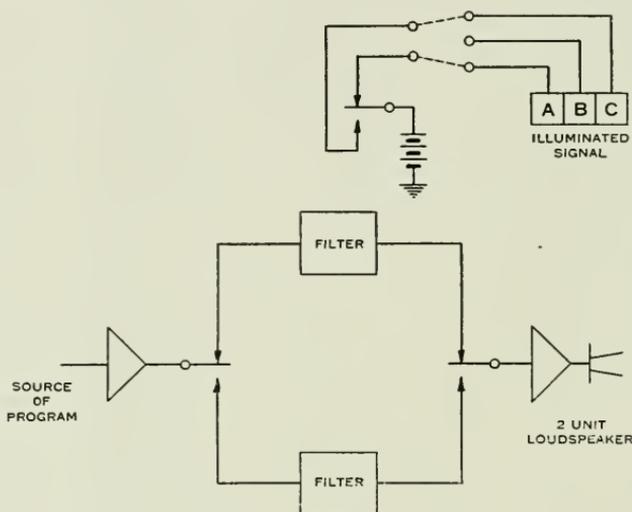


Fig. 1—Arrangement for testing program band widths.

listeners in which one of the letters, A, B, or C, could be illuminated. On a given test two of these letters were associated with the switch so that one letter was illuminated for one position and the other letter for the other position. The choice of letters among the three was varied more or less at random for different tests. Low pass filters were available to provide cut-offs of 3, 5, 8, 11 and 13 kc. When no filter was inserted the band was considered to extend to 15 kc. as this was about the upper limit of transmission of the testing circuits and loud speaker. The lower limit of the transmitted band for all conditions was approximately 40 cycles.

In conducting a test, a group of observers listened to comparisons between two of the available band widths, the conditions being switched every few seconds until a sufficient number of comparisons had been made. The

conditions were unknown to the observers, being designated to them only by the letters in the signal. At the conclusion of the test the observers were asked to mark on a ballot which letter appeared to coincide with the wider band (not which they preferred). A series of tests consisted of comparisons between substantially all of the possible band widths among those available. There were also included in some of the series as a check, one or two tests in which the band width was the same for both positions of the switch. Ten complete series of tests were carried out, two on each of five different programs.

The programs consisted of a dance orchestra, two large symphony orchestras, speech from a male speaker repeating a test sentence, and a radio dramatic sketch. The programs, except for the spoken test sentences, were obtained by special arrangement over direct wire lines from the studio or theater in which the performance took place. The entire system from microphones to and including the loud-speaker had a substantially flat transmission characteristic from 40 to 15,000 cycles, with no filters in the circuit. The loud-speaker was of the two-unit type and was one of a number built for the demonstration of auditory perspective in 1933. The tests were conducted in the program laboratory of the Bell Telephone Laboratories where the acoustic noise level was about +30 decibels. The noise contributed by the electrical parts of the system was considerably below the acoustic noise. The loudness of the programs was adjusted to about unity reproduction, that is, to the volume that would be heard by listeners in a favorable position at the original performance.

The observers were engineers having a considerable experience in tests of program quality. They were doubtless therefore considerably more critical than the average radio listener. The number of observers varied somewhat during the tests but averaged about sixteen. The ages of the observers were in the 30's and 40's so that neither very young nor very old ears were represented.

The immediate outcome of the tests was some 2,000 ballots which were meaningless until analyzed. Before the analysis could be made, however, it was necessary to decide how to express the results.

There are no familiar units to express fidelity or program quality. It was decided therefore to employ the very useful concept of the limen and the liminal unit. These terms have occasionally been applied to other subjective data and may be roughly defined as the least change in a quantity which is detectable. In the present case, if the band widths being compared differ greatly, there will be a nearly unanimous agreement among the observers as to which is the wider. If they differ only slightly, however, many of the observers will vote wrongly for the narrower band and on successive repetitions of the test many will reverse themselves. An average of

a large number of votes will show a plurality for the wider band, the margin of choice increasing as the difference in band width is made greater. A significant measure of the detectable difference in band width will be taken to be that difference such that 75% of the observers correctly select the wider band and 25% wrongly select the narrower band. This difference in band widths will be designated one "difference limen." The sensory effect of a change of one difference limen will be called one "liminal unit".

The significance of the vote of 75 to 25% is assumed to be as follows: On a particular test some of the observers can detect the difference between the conditions while the remainder will guess. Of the latter, half are likely to guess right and half wrong. When 25% vote wrongly they are assumed to be guessing and must be paired with another 25% who also guessed but happened to guess right. Therefore a vote of 75 to 25% is taken to indicate that 50% of the observers were guessing and the remainder could actually detect the difference. The difference limen may now be more specifically defined as that difference in band widths which is detectable to half the observers.

It may be commented that this attempt to explain the definition of "liminal unit" is perhaps over-simple. The observers themselves are frequently uncertain whether they are guessing or are influenced in their choice by some minute difference. The test could be done with a single observer, repeated many times to obtain the same number of observations as with a group. When the conditions are nearly equal he will vote about as often one way as the other, but as the difference between the conditions is increased he will vote a larger per cent of the time correctly for the wider band, just as did the group. When the two conditions are separated by one difference limen he will vote correctly 75% of the time and wrongly 25% of the time, which may be said, in line with the argument given earlier, to indicate that he is guessing half the time and can discern the difference half the time. The difference limen could therefore be defined as that threshold difference for which there is an even chance of its discernment by a listener.

Having chosen a method of expressing the results, the analysis can now be attacked. The first step is to group together all tests on similar types of program material, and to determine for each band width comparison the per cent of votes for the wider and narrower band, respectively. The data thus obtained for music and speech are shown by the solid curves of Figs. 2 and 3. A curve labeled 8 kc., for example, shows the per cent of the total votes which selected as the wider each of the other band widths to which 8 kc. was compared. The points, although somewhat irregular, fell systematically enough to permit drawing the smooth curves with the application of some judgment and having due regard to the necessary symmetry between them. (For example, the 8 kc. curve at an abscissa of 5 kc. must

agree with the 5 kc. curve at an abscissa of 8 kc.) A much larger volume of data would be needed to obtain points falling accurately on a smooth curve. To facilitate obtaining the best approximations, the curves were plotted on several kinds of coordinates, including rectangular, semi-logarithmic (shown in the illustrations), probability and logarithmic probability.

The dotted curves were interpolated between the solid curves and progress in steps of 1 kc. The interpolation was readily accomplished with consider-

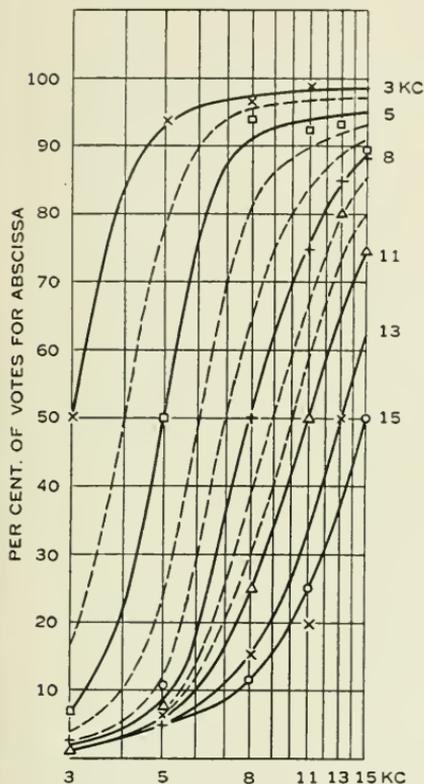


Fig. 2—Music

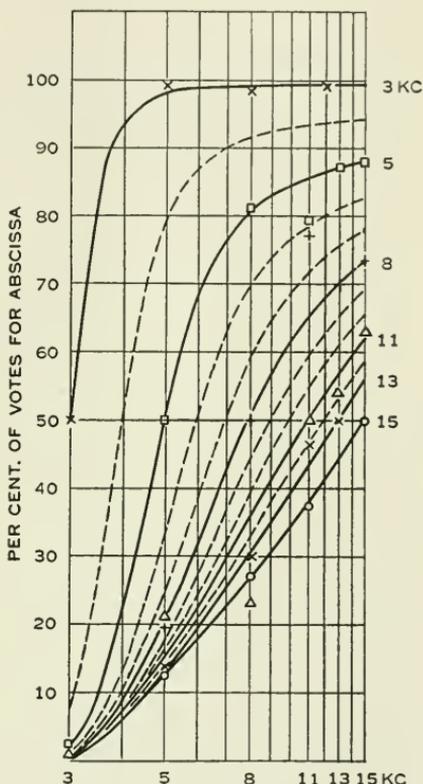


Fig. 3—Speech

Figs. 2 & 3—Detectability of changes in band width.

able accuracy. For example, points for the 10 kc. curve are obtained from the values of each of the solid curves corresponding to an abscissa of 10 kc.

From these curves, the difference limens for each band width were determined by reading directly the bands corresponding to votes of 25% and 75%. The bands at which these votes occur therefore by definition differ from the reference band by one limen. The following table gives the intervals of one limen as thus derived from the curves.

DIFFERENCES IN UPPER LIMIT OF PROGRAM BAND IN KC, CORRESPONDING TO ONE LIMEN

Music	Speech
3—3.6	3—3.3
3.3—4—4.8	3.4—4—4.8
4.1—5—6	4.1—5—6.9
5—6—7.4	4.6—6—9.4
5.8—7—9.3	5.1—7—12.8
6.4—8—11	5.5—8
6.9—9—12.2	5.8—9
7.4—10—13.4	6.2—10
8—11—15	6.4—11
9.8—13	7—13
11—15	7.6—15

The difference limens are seen to vary with the frequency of cut-off, increasing as the frequency increases. Since each difference limen corresponds to a sensory effect of one liminal unit, it is obvious that the reciprocal

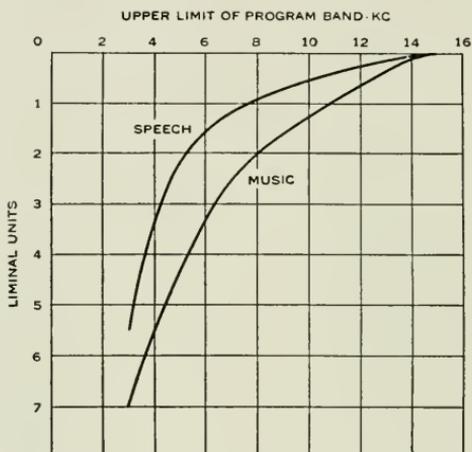


Fig. 4—Ability to detect changes in program band width.

of the difference limen gives the rate of change of liminal units with changes of program band width in terms of liminal units per kilocycle. Therefore, curves of liminal units versus the upper limit of the program band may be constructed from the figures in the table. Such curves are plotted in Fig. 4. The actual mechanics of the process used to plot the curves was as follows, taking the data for "music" for illustration. The lowest frequency occurring in the table is 3 kc., and it is seen that raising the band width to 3.6 kc. will bring about a subjective increase of one liminal unit. Therefore, on an arbitrary scale, 3 kc. was plotted at 0 and 3.6 kc. at one liminal unit. Next a smooth curve was drawn through these points and the location of 3.3 kc. (next line of table) was determined by interpolation. Since 4 kc. is one liminal unit above 3.3 kc., and 4.8 is one liminal unit above 4 kc., these points were plotted and the curve extended through them. By a similar process

the curve was extended step by step up to 15 kc. Finally, the origin was shifted so as to express the liminal curve with respect to 15 kc. instead of 3 kc.

It was mentioned above that a number of tests were introduced without the knowledge of the observers in which the conditions were not changed, the band width remaining constant while the illuminated letters were switched. This produced the most interesting psychological result that observers voted nearly two to one for the letter appearing in the right-hand position in the signal, on each of the six tests of this kind. This raises the question as to whether this effect impaired the results on the other tests.

In the course of the tests, comparisons between each pair of band widths were presented 10 times, 6 times with music and 4 times with speech. The letters corresponding to the two conditions were assigned more or less at random from the three letters A, B, and C. Taking 11 of these groups of tests in which the narrower band was represented about as often by the right hand as by the left hand of the pair of letters chosen, the average vote for the right-hand letter was 51.1% and for the left-hand letter was 48.9%. The difference between these two figures is too small to be significant. It is therefore concluded that when there was a real difference, the observers were not measurably influenced by their slight subconscious predilection for the right-hand letter. It would be interesting to correlate this phenomenon with the right or left-handedness of the observers. This point illustrates the extreme care that must be taken in conducting judgment tests of this sort to insure that no irrelevant factors affect the statistical result.

The curves of Fig. 4 permit drawing the following conclusions:

1. Increases in band width can be detected up to 15 kc. for both music and speech. The fact that this is true for speech is rather surprising. However, above about 5 kc., changes in band width are twice as readily detectable on music as on speech.
2. It requires an increase in band width from 8 to 15 kc. to be as readily detected as an increase from 5 to 8 kc., for both speech and music.
3. The following intervals correspond to one liminal unit and are therefore just discernible half of the time to the observers:
Speech: 5 to 8 kc.; 8 to 15 kc.
Music: 5 to $6\frac{1}{2}$ kc.; $6\frac{1}{2}$ to 8 kc.; 8 to 11 kc.; 11 to 15 kc.

In considering these conclusions, the fundamental assumption and limitations of the data should be borne in mind. First, the data were obtained from tests with a certain group of observers and on certain program material. Curves of somewhat different slope would doubtless be obtained with observers of different average age, experience, musical appreciation, etc. It is likely, however, that this would affect the absolute importance of the different intervals in liminal units rather than the relative values. As noted

earlier, the observers in these tests were considerably more experienced and critical than average radio audiences. The program material tested was representative of most of the programs on the air, but different results would be obtained with material markedly different in nature. This would probably be particularly true of selected sound effects. Secondly, it should not be forgotten that the results are based only on the ability of the ear to detect the changes, with no weighting for factors such as aesthetic values or per-

TABLE I

	Upper Frequency Limit Versus Unrestricted Band, Corresponding to One Liminal Unit
Musical Instruments	
1. Flute.....	13,500 cycles
2. Snare Drum.....	13,000
3. Violin.....	13,000
4. Soprano Saxophone.....	12,700
5. Oboe.....	12,700
6. 14 in. Cymbals.....	12,000
7. Bass Clarinet.....	10,500
9. Piccolo.....	10,200
9. Bassoon.....	10,000
10. Cello.....	9,800
11. Bass Saxophone.....	8,600
12. Clarinet.....	8,500
13. Trumpet.....	8,300
14. Bass Viol.....	7,800
15. Trombone.....	7,200
16. Bass Tuba.....	6,300
17. French Horn.....	6,100
18. Piano.....	5,600
19. Bass Drum.....	4,300
20. Timpani.....	3,500
Speech	
Male.....	7,300
Female.....	9,200
Sound Effects	
Footsteps.....	12,000
Handclapping.....	15,000
Key Jangling.....	15,000

sonal preferences, or for the effects of room noise and other factors present in the practical case. Thirdly, it should be appreciated that comparison tests such as these are very sensitive tests, showing up differences that could not be detected under usual home listening conditions.

It is of interest to compare the above results with previously published data. In a paper "Audible Frequency Ranges of Music, Speech and Noise,"¹ W. B. Snow gave data for 20 musical instruments, certain noises, and

¹ Jour. Acous. Soc. Amer., July 1931; *Bell Sys. Tech. Jour.*, Oct. 1931.

speech. The data showed the frequency limitations as compared with unlimited bands (about 15 kc.) which yielded a vote of 60 to 40%, and 80 to 20% among a considerable number of observations. In Table I these data have been interpolated to determine the limits that would correspond to a vote of 75 to 25%, in line with the criterion assumed in this paper. In making the interpolation, it was assumed that the curve of per cent of observers voting correctly for the wider band versus logarithm of the frequency is a straight line in the range of interest.

TABLE II

	Lower Frequency Limit Versus Unrestricted Band, Corresponding to One Liminal Unit
Musical Instruments	
1. Bass Viol.....	53 cycles
2. Bass Tuba.....	55
3. Timpani.....	60
4. Bass Drum.....	72
5. Bass Saxophone.....	72
6. Bassoon.....	74
7. Bass Clarinet.....	80
8. Cello.....	83
9. Snare Drum.....	87
10. Piano.....	95
11. Trombone.....	110
12. French Horn.....	125
13. Clarinet.....	140
14. Trumpet.....	160
15. Soprano Saxophone.....	210
16. Violin.....	230
17. Oboe.....	240
18. Flute.....	250
19. 14 in. Cymbals.....	370
20. Piccolo.....	510
Speech	
Male.....	115
Female.....	190
Sound Effects	
Footsteps.....	95
Handclapping.....	135
Key Jingling.....	915

It is difficult to interpret these data from individual instruments in terms of results to be expected from whole orchestras and other music as usually heard. However, comparing Table I with Fig. 4, it will be seen that the frequency limit determined from the present tests as corresponding to one liminal unit for music falls about one third the way down the list of instruments in the table, and the limit corresponding to two liminal units falls about two thirds down the table, which seems reasonable. Also the frequency limit found in the present tests to correspond to one liminal unit for

speech lies between the figures given in the table for male and female speech, which is a good check.

The present tests did not include measurements on the lower end of the frequency band. However, some clue to the results that would be expected may be obtained from Mr. Snow's paper. Table II, derived from Mr. Snow's data in a manner similar to that just described, gives the lower limit of the frequency band corresponding to a degradation of one liminal unit compared with transmitting a much lower frequency.

The frequency corresponding to one liminal unit for speech may be taken as the mean of the figures for male and female speech, or about 150 cycles. In the case of music, it may be expected that at the lower as well as the upper end of the frequency range one liminal unit for an orchestra should fall about one third the way down the list of individual instruments, and two liminal units about two thirds the way down the list. This would make one liminal unit for music correspond to about 80 cycles and two liminal units to about 150 cycles. This speculation leads to the interesting hypothesis that the relations are probably the same at the lower as at the upper end of the frequency scale, that is, changes in band widths are twice as readily detected for music as for speech, and that the frequency limit corresponding to one liminal unit for speech corresponds to two liminal units for music.

CHAPTER V

Use of the Etch Technique for Determining Orientation and Twinning in Quartz Crystals

By G. W. WILLARD

This paper is one of a series of papers dealing with piezoelectric circuit elements and their manufacture.¹ Certain parts of the paper are not new or original, but have been added for the sake of completeness and for the convenience of the reader.

5.1 INTRODUCTION

THE manufacture of piezoelectric plates from crystalline material involves orientation problems not encountered in the fabrication of objects from non-crystalline materials. The reason for this is that crystalline materials have physical properties which vary with the orientation, or direction, in which they are measured. Since the operating characteristics (activity, frequency, and temperature-coefficient) of the finished piezoelectric plate depend, not only upon the shape and dimensions of the plate, but upon the physical properties (electrical, elastic and thermal) of the crystalline material, the finished piezoelectric plate must have a *specific orientation* with respect to the material as well as a specific shape and dimensions. In the case of quartz piezoelectric plates the orientation problem is complicated by two factors. First, a large portion of the available natural quartz crystals lack such natural faces as are required to determine accurately the structure-orientation from the shape of the original stone. Thus the raw stones must be examined for structure orientation by physical instruments before even the first cuts may be made. Secondly, a large portion of natural quartz crystals are twinned, i.e. not of the same structure orientation throughout the stone. The boundaries of the respective, homogeneous regions are not predictable, and cannot be completely located in the uncut stone. Thus the processing of quartz involves a step by step examination for twinning boundaries and orientation as the raw stone is cut into sections, the sections cut into bars or slabs, and the bars or slabs cut into blanks. Even when using untwinned stones the orientation must be redetermined and corrected at each cutting step when making such plate types as require very exact orientation.

The most widely used methods of determining the structure orientation

¹ See B.S.T.J., Vol. XXII: No. 2, July 1943 for Chaps. I and II; No. 3, Oct. 1943 for Chaps. III and IV.

of quartz are: (1) by optical effects (birefringence and rotatory power), (2) by X-ray reflections from atomic planes, and (3) by the use of etch pits which are developed when the quartz surface is etched in fluorine compounds. Other methods are or may be used in rather special cases. For example, in finished plates of known orientation types, the electrical axis direction is distinguished from other directions by electrical polarity tests (on tension or compression), or a plate known to be one of several types may be tested in an electric circuit for activity, frequency and temperature-coefficient, to determine which type it is. The selective fracture characteristics of quartz offer another method of determining orientation. Microscopic fractures resulting from grinding a quartz surface may be used for determining orientation. Thus unetched, ground, Z-cut surfaces of quartz give a hexagonal figure, when examined by pinhole illumination, which may be used to determine the approximate orientation (but not sense) of the electric axes.²

By optical methods (see Chapter II) it is possible to determine the orientation of a quartz body relative to only one direction of the structure, the optic or Z axis. Thus optical methods are limited to determining the angle between the optic axis and a line or surface of the body (but not the rotation of that line or surface about the optic axis). Twinning of the "optical" variety may be detected optically, even when located internally, but the determination of its location in depth is approximate.

By X-ray methods (see Chapter III) it is possible to determine the structure orientation of a quartz body exactly and completely. However, this method is limited in application by the complexity of analysis, except when the approximate orientation is already known. Though twinning can be detected on the surface of the body, it is not generally feasible to explore the surface to locate twinning boundaries. Further, though positive or negative sense of angular orientation is obtainable by X-rays, this part of the complete determination is not reliable unless the specimen examined is known to be free of twinning, or unless the twinning boundary locations are known. Thus X-ray determinations of orientation are generally limited to determining exact orientations in quartz bodies of approximately known orientation (which includes the case in which only one axis is approximately known).

The etch method of determining orientation is commonly used in conjunction with the optical and X-ray methods to give the information that those methods do not give. The etch method, as most commonly and practically applied, does not give exact orientation angles, nor is it applied to specimens of entirely unknown orientation. However, when a surface of approximately known orientation is etched, it is possible to determine approximately the complete orientation (including sense) of the specimen, and further to detect at this surface both electrical and optical twinning and to

² See Fig. 5.20, and further explanation at the end of Sec. 5.53.

determine exactly the twinning boundary locations. The detection of twinning and twinning boundaries by this method has been practiced for years. The determination of orientation and sense of orientation has been exploited only more recently. At present the etch methods play an important and extensive role in the processing of quartz plates, not only in the routine determination of orientation, but also in the detection of twinning so that the most economical cutting methods may be practiced.³

5.2 TWINNING (GENERAL)

Although the problems related to twinning are largely those of determining orientation of the crystal structure, the nature and prevalence of twinning in crystal quartz presents a special group of problems that would be absent were the twinning absent, and hence are separately grouped as twinning problems. As pointed out in Chapter IV, there are only two common types of twinning in the commercial quartz used for piezoelectric plates, namely, electrical and optical twinning. A simplifying feature of both these types is that the structure axes (optic axis and electric axes) of all portions of a single crystal are parallel each to each. However, they are not of the same sense, or handedness. The difference between the two types is as follows:

In a crystal which is only **ELECTRICALLY TWINNED**, the crystal is entirely of one handedness (either right or left), but one portion is of **OPPOSITE ELECTRICAL SENSE** to another portion, i.e., the electric axes are of opposite sense.

In a crystal which is only **OPTICALLY TWINNED**, one portion of the crystal is of **OPPOSITE HANDEDNESS**, and electrical sense, to another portion. This twinning (but not electrical) is detectable by optical means (polarized light) and is named optical twinning for this reason.

The extent of twinning that may be present in commercial crystals is seen in Fig. 5.1, which shows both electrical and optical twinning boundaries at the top surface of some Z-cut (basal) sections of quartz (which were cut up for the manufacture of quartz oscillators). Though the crystals are seldom entirely free of twinning, they do not on the average run as badly twinned as here shown. These views, taken by means to be described, correspond to what one sees when examining an etched quartz surface by reflection from a strong light.

Since untwinned finished plates must be cut entirely from one twin or another (not across a boundary), and since the proper sense of angular orientation of the plate is opposite for two adjacent electrical twins, the economic utilization of twinned quartz is a difficult problem.⁴ It involves cutting the

³ Etching is also used on finished plates for removing grinding debris, and for frequency adjustment.

⁴ As herein used, a *twin* is one of the homogeneous, untwinned portions of a twinned crystal.

stone into separate parts when the twins are large enough to be utilized separately. Further, at some stage before reaching the finished plate all twin portions but one must be cut away.⁵

In this connection it is important to note a size and form difference between electrical and optical twins. Fig. 5.2 shows the appearance of twinning boundaries when only ELECTRICAL TWINNING is present. Note that electrical twins are commonly large, hence may often be separated ap-

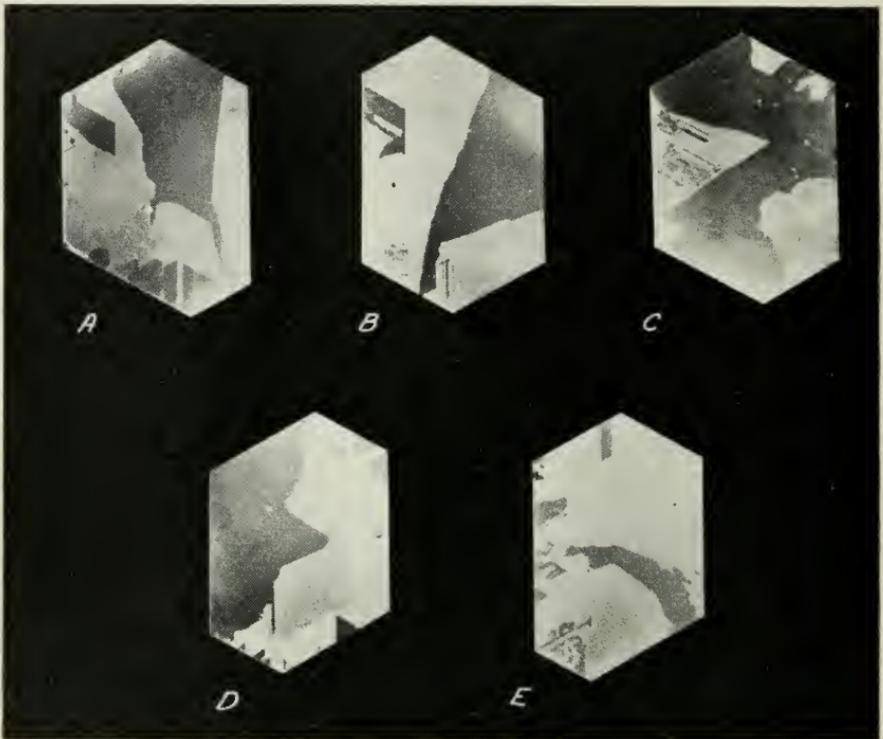


Fig. 5.1—Examples of ELECTRICAL and OPTICAL twinning, as exhibited at the etched surface of Z-cut sections. These examples are typical of an appreciable portion of the quartz that is cut up for quartz plates.

proximately along a boundary and both portions utilized. Fig. 5.3 shows the appearance of twinning boundaries when only OPTICAL TWINNING is present. Since optical twins are commonly small and in the form of thin laminations, it is seldom possible to cut optical twins apart and use both parts separately.

The conventions here used, regarding handedness and axial sense, are

⁵ See Section 5.7 for the possibility of utilizing partially twinned finished plates.

according to those of the proposed "I. R. E. Standard."⁶ Figure 5.4 shows the relation of these conventions to the natural faces of right and left quartz, to the electric charges developed on compression and tension, and to the more common cuts of oscillator plates. Also given are the relations of handedness to the conoscope and the polariscope means of detecting handedness (Section 2.7, Chap. II describes these instruments). It is important to

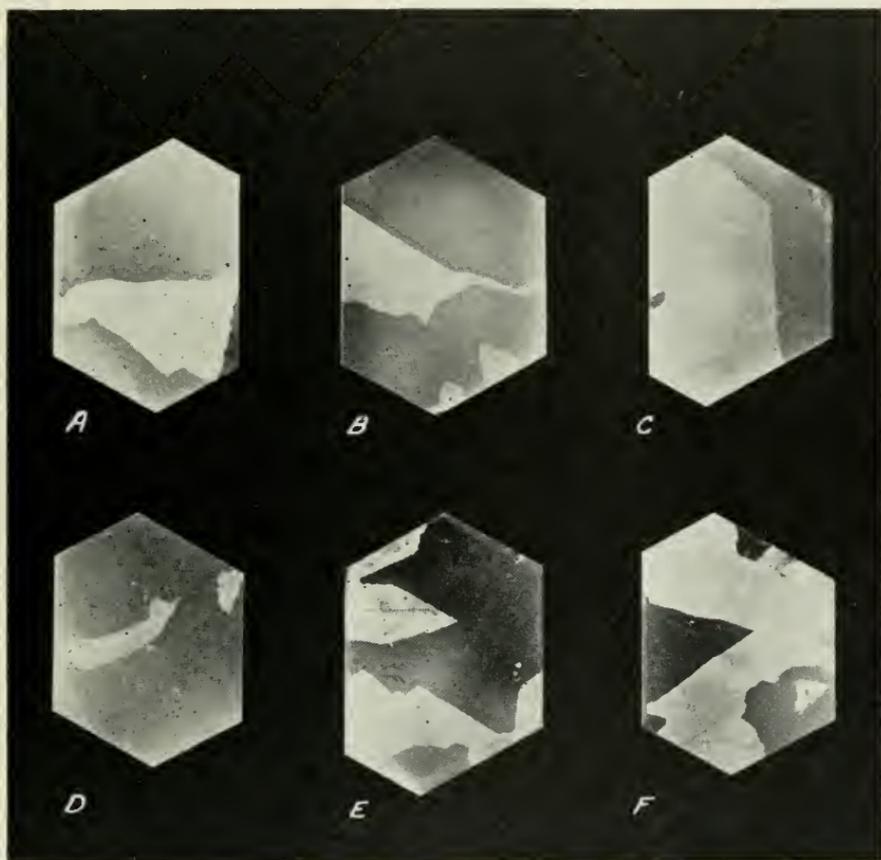


Fig. 5.2—Examples of ELECTRICAL twinning alone. Electrical twins are commonly large, and hence may be cut apart and used individually.

note that AT and CT plates are always cut at such an angular sense, relative to the Z and X axes, as to be roughly parallel to a *minor pyramidal face*, whereas the BT and DT plates are roughly parallel to a *major pyramidal face*. Thus a stone exhibiting these faces may be cut into any of these plates

⁶ "Proposed Standard Conventions for Expressing the Elastic and Piezoelectric Properties of Right and Left Quartz", *Proc. I. R. E.*, Nov. 1942, p. 495.

without determining the handedness and electrical sense of the stone (if twinning is negligible). As will be seen later, a similar situation prevails when analyzing etched X-cut sections for cutting into plates.

5.3 NATURE OF ETCH-PITS

When crystal quartz is etched by contact with hydrofluoric acid (or other etching agents) the surface of the quartz is eaten away in such a manner as

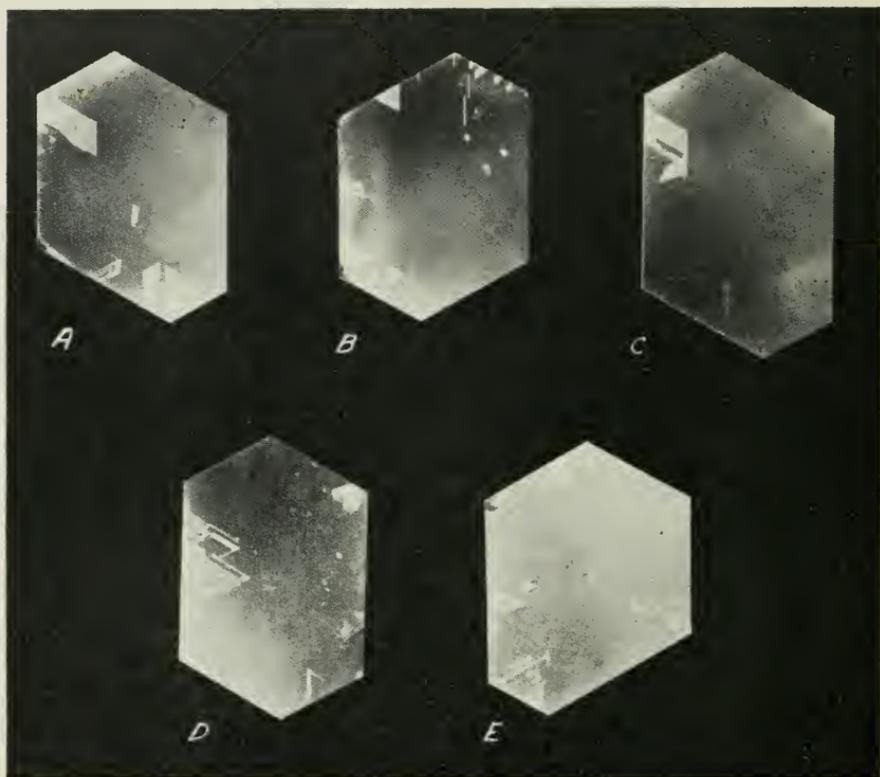


Fig. 5.3—Examples of OPTICAL twinning alone. Optical twins are commonly small and interlayered, and hence may not be separated and used individually.

to leave microscopic *etch-pits* (or hills). These etch-pits are formed of minute facets which are definitely related to the crystal structure. The form of these pits and the orientation of the facets may be used to determine the orientation of the crystal structure at the etched surface being examined.

The general appearance of four types of etch-pits is shown in the photomicrographs of Fig. 5.5. These are the pits that are developed on ground surfaces which are approximately parallel to the well known X-, Y-, and Z-cut surfaces of right hand quartz, by the action of hydrofluoric acid. It is

seen that the positive and negative X-surfaces produce different etch-pits, and are thus usable in determining electrical sense. Further, the pits on all surfaces have directional properties which allow them to be used for determining the approximate directions of the axis which lie in the etched surface. However, to be able to determine orientations from etched surfaces of other

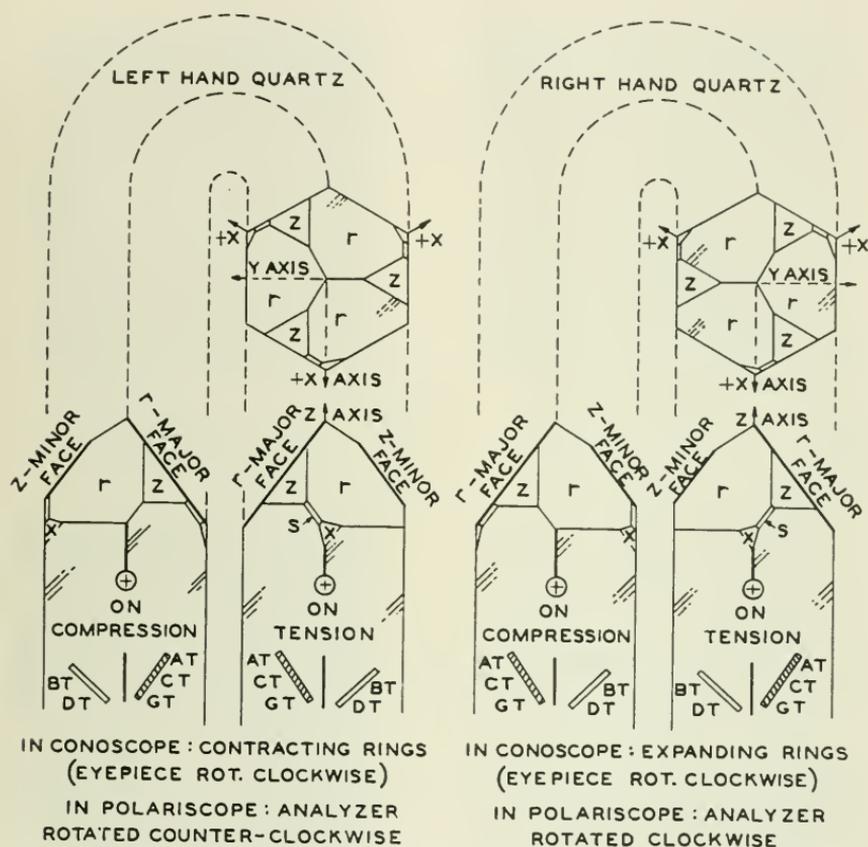


Fig. 5.4—The conventions of handedness, axes, natural faces, and angular sense-of-cut of common oscillator plates, together with the electrical and optical rules for determining these characteristics in unfaced stones.

orientations than those shown above, requires a knowledge of the appearance of the etch-pits developed on such surfaces.

A rather complete catalog of etch-pits on all possible surfaces of quartz was prepared by W. L. Bond,⁷ using an etched sphere of quartz (Figs. 5.5, 5.6 are from Bond). Thirty-six different types of etch-pits were obtained and their angular range of coverage was found (the X-, Y-, and Z- surface

⁷ "Etch Figures of Quartz," *Z. Kristallogr.* (a) 99, 1938, pp. 488-498.

pits are obtained only on surfaces within 6° to 8° , from the X-, Y-, and Z-surfaces, respectively). Since the development of good etch pits and their exact appearance is considerably affected by the preparation of the surface for etching (fineness of grind), and by the strength of the acid and the length of etching time, and by the manner of illumination when viewing, the

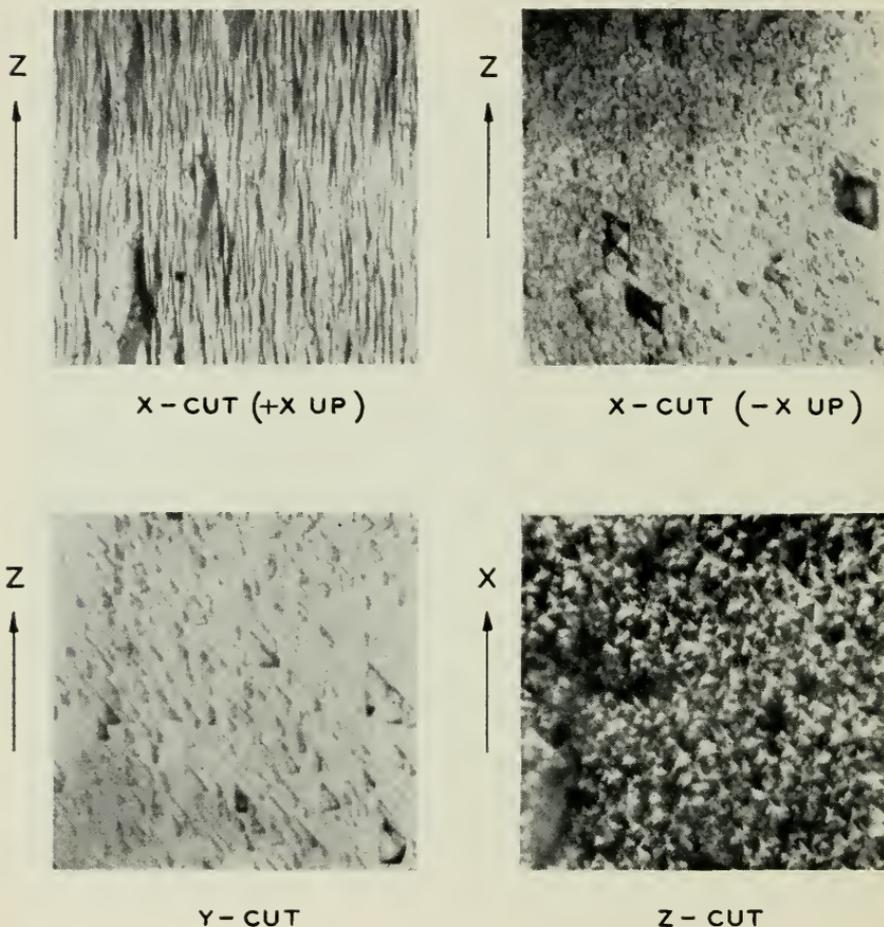


Fig. 5.5—Photomicrographs of etch-pits on the etched surfaces of common orientations. As seen the etch-pits are definitely related to the structure axes of the quartz.

figures shown here do not represent the exact appearance of pits obtained by other manners of development. However, such figures are reproducible.

The use of etch-pits to determine the orientation of a perfectly general surface is complicated by the fact that some different surface orientations give pits not readily distinguished from each other. However, for the surfaces most commonly encountered in quartz plate manufacture the etch-

pits are quite distinctive, when well developed. Use may be made of a microscope or a high powered projector to view the figures. The pit outlines may be aligned with lines ruled on the eye-piece or on the screen, and a fixed marking device may be used to mark the quartz surface with orientation lines. Twinning may be detected by the appearance of different etch-pits as the specimen is moved about. For example, on an electrically twinned X-cut surface both X-cut views of Fig. 5.5 could be found. However, the location and marking of twinning boundaries involves a tedious exploration of the surface, since only a minute portion is viewed at any one time. This exploration may be eliminated if the surface is first viewed by reflection methods where the whole surface and extent of twinning is at once seen (as in Fig. 5.1) and marked.

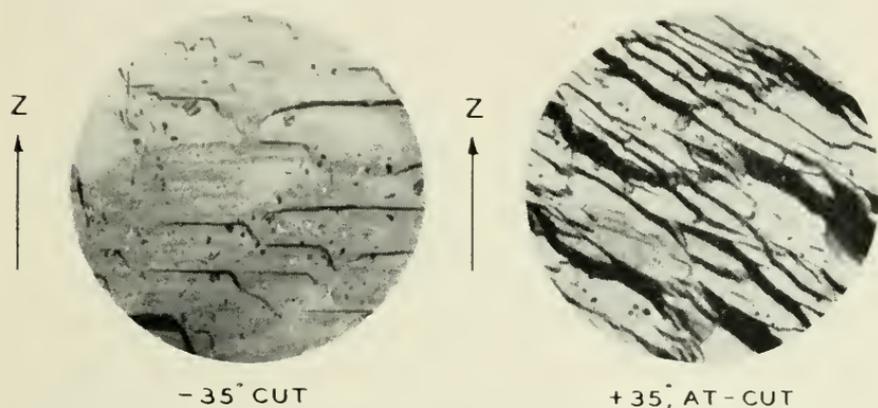


Fig. 5.6—Etch-pits on the etched surface of a $+35^\circ$ AT plate, and on an analogous but wrong sensed -35° plate. This difference in etch-pits may be used in the manufacturing process to determine the right and wrong sensed regions of twinned AT slabs.

A special case where the microscope or projector method might be employed is in the examination of thin AT, BT, CT or DT slabs for twinning and sense of cut. Here the slabs are known to be cut with a reference edge parallel to an electric axis, and with the major faces inclined at 35° to 55° (depending upon the variety of slab) from the optic axis, the sense of the inclination being positive for the AT and CT slabs, and negative for the BT and DT. The effect of electrical twinning on such etched surfaces is shown in Fig. 5.6. The etch-pits of the good $+35^\circ$ AT-portion of the slab are easily distinguished from the analogous -35° (bad) portions. This difference is similarly distinguishable in the other cuts.

Actually, orientation and twinning are seldom analyzed by the method described above, i.e. by examining their appearance in the microscope, or by projection on the screen. The method appears to be far less practical than other methods which depend upon the gross effect, of hundreds of simi-

lar etch-pits, in bending a light beam. By the latter methods the individual etch-pits are never seen, nor does their nature need to be known. Nevertheless, the resultant optical effect of hundreds of similar etch pits is as characteristic of structure orientation as the individual pits themselves.

5.4 OPTICAL EFFECT OF ETCH-PITS

The gross optical effect of hundreds of similar etch-pits results from the fact that each of the pits has minute facets which are similarly inclined to those of all the other pits. Though the pits of Figs. 5.5 and 5.6 may not appear to be formed from groups of flat facets they are generally so regarded. "Curved-facets" are theoretically considered to be made up of individual flat-facets which are parallel to possible atomic planes (and hence may be given index numbers as in Chap. III). This view is the same as that taken

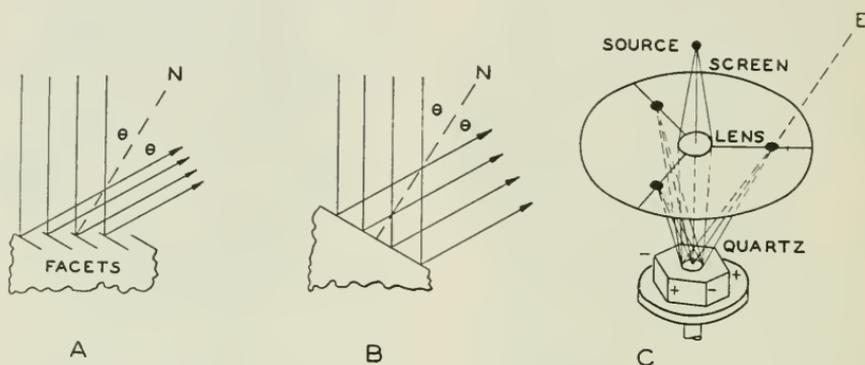


Fig. 5.7—Reflection of light from a single set of similarly oriented etch-pit facets, A, is like that from a single mirror, B. Reflection from all three sets of facets of a Z-cut section will give a three-fold *etch-figure* on a screen, as in C.

with regard to natural faces, which are of course produced by essentially opposite effects, i.e., acid corrosion in the case of etch-pits, and growth from solution in the case of natural faces. Actually, many "curved-facets" give optical effects showing no discernible evidence of individual flat facets. However, the question is academic, so far as use of the pits for orientation purposes is concerned, for such facets are still definitely related to the crystal structure.

Etch-pit facets may be used to *reflect* a light beam into specific patterns or to *refract* the beam on transmission through the material into similar (but not identical) patterns. The different basic optical means of using etch-pit facets are shown in Figs. 5.7, 5.8, 5.9. Included in each figure is a diagram of the effects obtained by illuminating an idealized Z-cut section. This idealized section is assumed to have only simple, equilateral, three-sided

pyramidal etch-pits, oriented relative to the X axes as shown in Fig. 5.5. The actual results obtained with Z sections are more complicated than this and thus indicate that the etch-pits are not exactly as idealized here.

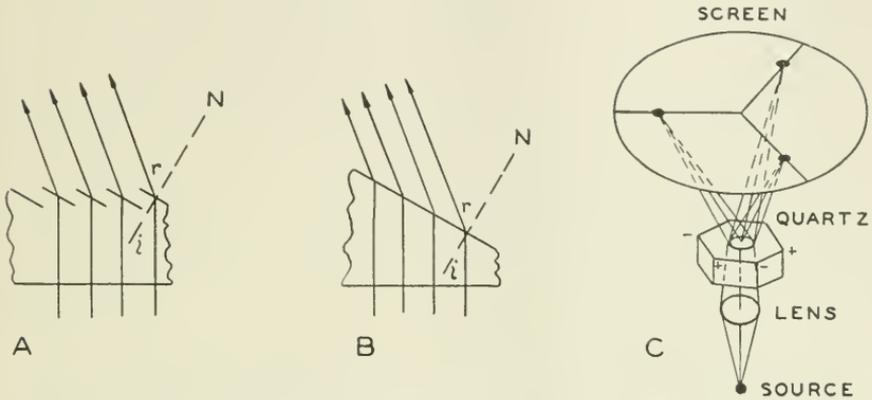


Fig. 5.8—Light transmitted thru a single set of etch-pit facets, A, is refracted as by a prism, B. The three sets of facets of a Z-cut section give a three-fold etch-figure, as in C.

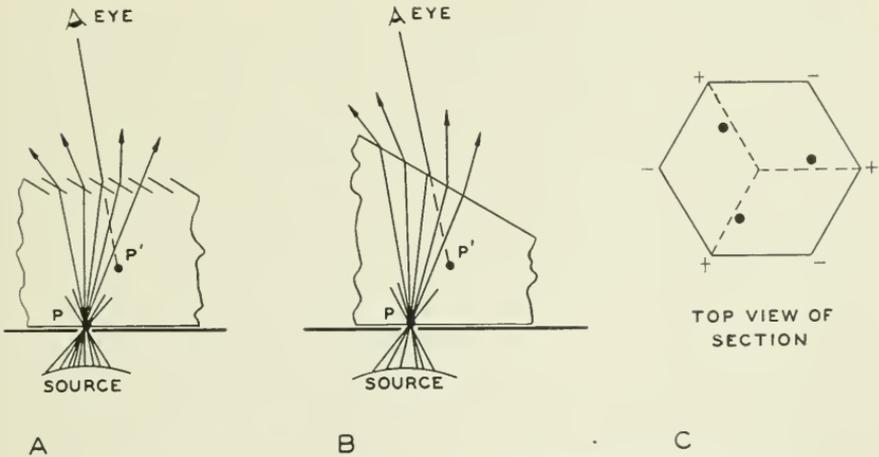


Fig. 5.9—Light transmitted thru a pin-hole is refracted by a single set of facets, A, as it would be by a prism, B. A virtual image of the pin-hole P will be observed at P'. The etch-figure seen down in a Z-cut section is three-fold, as in C.

5.41 THE REFLECTION METHOD

Figure 5.7 shows the reflection method, where a parallel beam of light striking the etched surface of a Z-section is reflected from one of the three sets of facets as shown in A. Each single facet reflects part of the beam by

ordinary reflection laws, and the whole groups of facets act similarly to a single mirror surface at the same angle, as in B.⁸ The individual facets being very minute and of irregular size and spacing, however, cause appreciable diffusion of the beam. The resultant effect of all *three* sets of facets is shown in C, where light passing down through a lens and a hole in the screen is reflected back to three spots on the screen. These three spots are located at equal distances from the incident beam and at 120° intervals around the incident beam. If the quartz section be rotated on its table the spots rotate around the screen correspondingly. However, lateral motion of the section across the table (without rotation) does not change the position of the spots, if the section be untwinned. If the section is twinned (or more exactly, if the etched surface is twinned) the three-fold figure will shift to a different position (angularly) on crossing a twinning boundary, for the etch pits are oriented differently in the two twins. If the twinning boundary divides the illuminating beam, then both figures appear at once, giving six spots instead of three. It is clear then that twinning, as well as orientation of the section, may be determined from the figure on the screen. The angular relation between the spots and the X-axes of the section will be considered later, where figures of actual sections are shown.

The long used method of examining etched quartz surfaces by simple reflection from a bright light, may also be explained from Fig. 5.7C. If a spot of light on the screen is viewed along the line E, and the screen then removed, the light from the associated etch-pits will fall on to the eye. The illuminated portion of the section will appear bright. If a twinning boundary crosses the illuminating beam and one of the six reflected beams falls on the eye, one of the two illuminated twins will appear bright and the other dark. As the section is rotated, first one twin and then the other will appear bright, and in each case the twinning boundary is sharply defined over the whole region covered by the illuminating beam (the appearance of twinned Z-cut surfaces examined by this means is shown in Figs. 5.1, 5.2, 5.3). Due to the greater complexity of etch-pits than here idealized, the reflected beams are not so sharply defined as to require exact location of the eye relative to the incident beam and the section. Further, when a broad unfocused light source is used, it is possible and convenient to detect twinning boundaries merely by holding the section in the hand and rocking it about in various directions until a brightness contrast is observed. Though the brightness contrast is usually not marked by this simple examination it suffices for many purposes.

⁸ That the effect of a group of facets is not identically the same as that of a single mirror, is of more concern where lenses are used for focusing. In this case the displacement of the mirror facets causes a displacement of the focus of the beam from each facet. For beams of small angular range this is of little importance.

5.42 THE TRANSMISSION METHOD

Figure 5.8 shows one form of the transmission method of examining Z-cut etched surfaces. A parallel beam of light passing normally up through the bottom polished surface and the top etched surface of a section will be bent by refraction only at the etched surface, as in A. Each facet refracts the light by ordinary laws of refraction, and the whole group acts similarly to a single refracting surface at this angle, as in B.⁹ The resultant effect of all three sets of facets is shown in C (where a lens is added for focusing the light beam). If the incident beam is not normal to the bottom surface there is an additional bending of the beam at this surface. If the incident surface is not polished (or rendered optically flat, with a cover glass and immersion fluid, for example) the diffusion at this surface will mask or completely destroy the desired effect.¹⁰

5.43 THE PINHOLE TRANSMISSION METHOD

Figure 5.9 shows the pinhole form of the transmission method, as applied to the examination of Z-cut etched surfaces. Here a section with a top, etched surface is illuminated from below through a small hole with a wide angle of illumination. The light radiates upward in all directions from the pinhole, and in passing through the upper etched surface is refracted by a single set of etch facets as in A. With the eye placed above the pinhole (and section), certain of these rays will fall on the eye. The eye then sees a virtual image of the pinhole P displaced to P', elevated from the level of P, and along the line of the ray which enters the eye. The effect of a group of facets is similar to that of a single prism, as in B.¹¹ The resultant effect of all three sets of facets of a Z-cut section is shown in C, where the section is viewed from directly above and no optical system is shown. Only the three virtual images of the pinhole are seen and they are located down in the quartz (roughly two-thirds of the way down).

Though the desired effect is due entirely to the top, etched surface, the nature of the bottom surface may cause a deleterious masking effect, which must be considered in the design of an instrument. Due to the diffusing effect of irregularities in the top surface it may act somewhat as a screen upon which the extended light source shown in Fig. 5.9A, B may be imaged by the pinhole. This extraneous image occurs if the bottom surface is polished, and to some extent if the surface is semi-polished, strongly etched, or oily.

⁹ See footnote 8.

¹⁰ Similar optics hold if the section is illuminated from the etched side instead of the polished side.

¹¹ See footnote 8.

This difficulty may be entirely obviated by the introduction of a diffusion screen directly adjacent to the pinhole.¹²

It might be noted that if it be desired to project or photograph the pinhole figure, one must focus on the virtual image which lies between the top and bottom surfaces of the etched specimen. In the simple case diagrammed in Fig. 5.10, it is assumed that the camera lens is at a distance from the section and directly over the section, so that the rays to the lens are essentially normal to the section. For a section of thickness T , and index of refraction n , the elevation E of the virtual image from the bottom surface of the section is given by: $E/T = 1 - \sqrt{1 + R^2/T^2}/n$. Here R is the radial displacement of the virtual image from the axis of the pinhole and is readily observed and measured. Also, R may be calculated from the thickness of the quartz T , the angle θ between the facets and the gross surface, and the in-

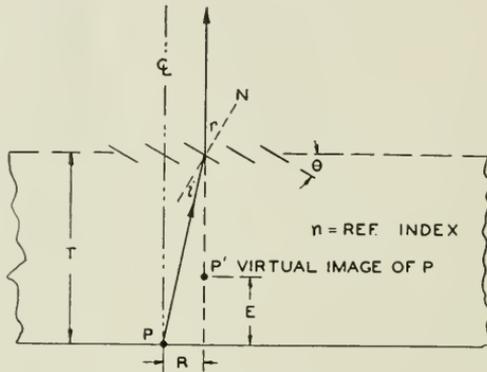


Fig. 5.10—The elevation E of the virtual image may be calculated from the thickness of the etched section T , the radial displacement of the image R , and the index of refraction n ; or from T , n , and θ , the angle between the facets and the gross surface.

dex n , (or θ may be calculated from T , R , n) by: $R/T = \tan(\theta - \sin^{-1}[(\sin^{-1} \theta)/n])$. Commonly, pinhole figures from quartz which is weakly to moderately etched (up to one hour in concentrated HF) have a maximum diameter (or double radial displacement) $2R$, nearly equal to the thickness of the section. Since the elevation of the image, E , depends upon its displacement R , an extended virtual image is not in a single plane and cannot be exactly focused (the elevation is commonly about one-fourth to one-third of the thickness of the section). The diameter of the pin-hole must always be kept small compared to the thickness of the section to give sharp figures (and the length of the pinhole must be small compared to its diameter).

¹² The diffusion screen may be a sheet of white paper placed over the pinhole, or a piece of flashed glass placed under the pinhole, with the flashed side against the pinhole. In either case it is usually necessary to increase the light intensity by focusing a concentrated light source onto the pinhole with a lens.

Choice of one of the four above methods of examining etched surfaces for twinning and orientation, depends upon many factors, as will be noted in the following section. The pinhole method is used wherever possible because of the simplicity of the optical system and the brilliance of the figures obtained.

5.5 ETCH-FIGURE INSTRUMENTS

Herein are described several instruments which have been designed for shop use in determining orientation and twinning of etched quartz sections and slabs. Their basic principles of operation are as described above. The nomenclature of handedness, sense of axes, sense of cuts, natural faces, etc. is according to Fig. 5.4, as explained at the end of Section 5.2.

The etch-figures and reflection patterns obtained on these instruments vary with the preparation of the specimen (i.e. the type of grind and the type of etch). A complete study of these factors would include a variation of the grind from a very coarse grind to polishing (and include saw-cut surface), and a variation of the etching time from short to very long, and the strength and kind of etching agent. Here chosen for illustration are the simplest practical preparations, namely, the coarsest grind usable, and the shortest etching time (in hydrofluoric acid). The etch-figures are thus markedly different than some which have appeared in the literature. Further, the photographic reproduction of etch-figures on paper, is not exact due to the limited contrast range of the paper. Thus in the accompanying illustrations detail is lost in the brilliant portions of the etch-figures in order to show details in the weaker portions, and vice-versa.¹³

5.51 THE REFLECTION ORIASCOPE

Fig. 5.11 shows diagrammatically a reflection "Oriascope", which may be used on specimens with a single flat etched surface. By the reflection principle of Section 5.41 figures are obtained on a viewing screen. Due to the relatively weak figures obtained by reflection from weakly etched surfaces, the viewing screen must be enclosed in a well blackened enclosure, and viewed through an eye chute. The screen is ruled with appropriate lines, relative to which the figure is aligned by turning the specimen on the table. The table is mounted so that when the specimen is properly oriented, the table may be slid to the right or left over a marking template, and marked through the template with appropriate lines to indicate the desired axial orientations of the specimen.

When used with Z-cut sections it is necessary to have two marking templates, one for each handedness of the quartz, since the three-fold figures

¹³ Apparent shifts in etch-figure orientation, with etching time for example, are not to be considered as resulting from an orientation shift of the individual etch-pit-facets, but as a shift in the relative areas of differently oriented facets. See Figs. 5.12 and 5.17.

obtained are not aligned with the electric axes of the specimen. They are shifted approximately 12° therefrom, and in opposite directions for the right and left varieties. Figure 5.11 shows a section of right quartz so positioned on the sliding table that the etch-figure therefrom will be properly aligned with three radial lines of the viewing screen. The section need not have natural faces as here shown. With the section so positioned the sliding table is moved over the right-hand marking template, and the section is marked with three radial lines. These lines on the section then give the approximate direction (within 5°) and the sense of the three electric axes of the quartz, positive X-outward. With left quartz the etch-figure is still aligned with the same lines on the viewing screen, but the section is marked through the left-hand marking template (the marking having the same meaning as be-

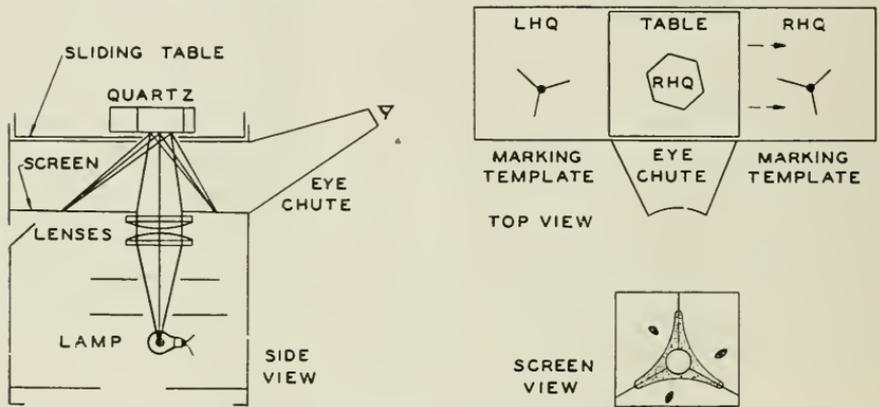


Fig. 5.11—The reflection ORIASCOPE as applied to determining the direction and sense of the X (electric) axes in Z-cut sections. After the etch-figure is aligned on the screen the table and sections are moved over a marking template and the section marked from below with axes.

fore). The section so marked is ready for laying out the approximate cutting directions, the sense of which may be found from Fig. 5.4. The exact cutting directions are obtained by X-rays. It might be noted that ordinarily the handedness of the section is determined in the conoscope (see Section 2.7, Chap. II) before examination on the oriascope. Also the twinning boundaries are previously determined by examination of the etched surface in a spot-light beam.

Figure 5.12A, B show the type of etch-figures obtained on Z-cut sections (in each case the figure is properly aligned with the rulings on the viewing screen). The simpler etch-figure A is obtained on a fine ground (400 carborundum) surface by a weak etch (about 10 minutes in 50% HF). Though the three faint spots, about 40° clockwise from the rulings (for the left-hand

quartz of A) may be used for determining the handedness of the section, it is usually considered more reliable to use the conoscope for handedness determination. The counter-clockwise rotation of these spots in B indicates right-hand quartz. The more complicated etch-figure B, results from etching a fine ground surface too long,¹⁴ or from using a coarse instead of a fine grind. With such figures it is difficult to know which portion of the figure is to be aligned with the screen rulings. Hence the sections must be fine ground and the etching time closely controlled.

The obvious disadvantages of the reflection oriascope (the necessity of predetermining handedness and twinning, and the requirements of fine ground surfaces and closely controlled etching time) are largely overcome by the pin-hole oriascope, later described. However, the reflection oriascope is an

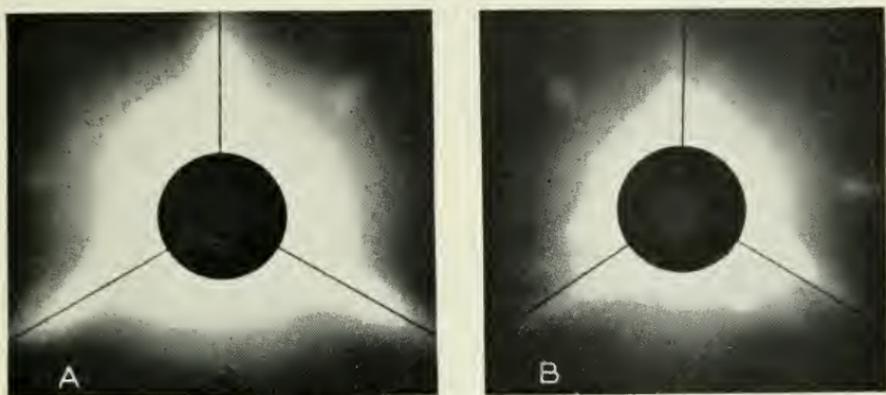


Fig. 5.12—Etch-figures obtained on the reflection oriascope with Z-cut sections (reduced from 11 inches square). A is a good usable figure while B is difficult to use due to its complexity.

excellent explanatory instrument for obtaining experimental etch-figures from surfaces of any orientation, preliminary to devising a special instrument to most advantageously utilize the reflection characteristics found. This fact results from the large and symmetrical screen coverage, and from the fact that only one etch surface is encountered by the light beam (thickness and back surface shape is of no concern).

5.52 THE REFLECTION TWINORIASCOPE

Figure 5.13 shows diagrammatically a reflection "Twinoriascope" designed especially for shop use in detecting and marking twinning boundaries and the sense of orientation in etched AT, BT, CT and DT slabs. When, for ex-

¹⁴ It appears that excessively strong etches (hours long) again give a simple, strong, and reliable figure.

ample, CT slabs are to be examined the tiltable mounting-table is clamped in the 38° position, and the slab placed crosswise on the table (X-axis normal to line of sight, and beveled edge as shown). Upon moving the viewing screen to position 1, only lamp 1 is lighted, and the slab is viewed by reflected light at a preferred angle. If the slab be twinned, one portion of the slab will exhibit a bright sheen while the other portion is dull by contrast, see two examples in Fig. 5.14, Test 1. The twinning boundary is now penciled in. The viewing screen is then shifted to position 2 which lights only lamp 2, and the crystal moved to right or left so that only one twin is illuminated. On the screen¹⁵ will be seen an etch-figure similar to one of the four shown in Fig. 5.14, Test 2. If either of the two positive-cut figures are observed the illuminated portion of the slab is usable, since the CT plate

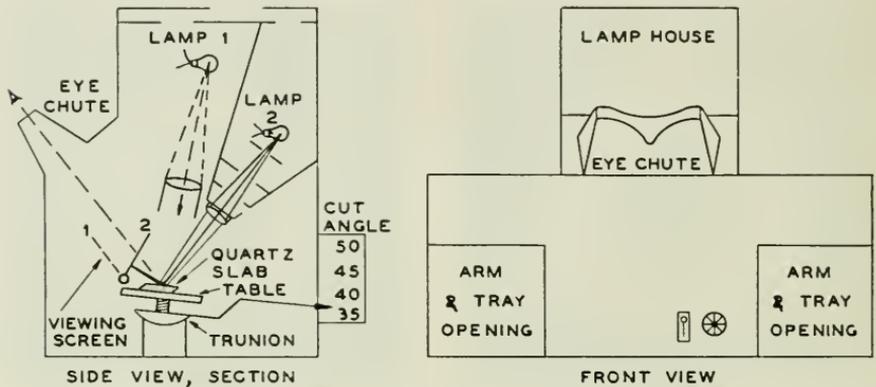


Fig. 5.13—The reflection TWINORIASCOPE for detecting *twinning* (using lamp 1 and no viewing screen, position 1) and for determining the orientation or *sense-of-cut* (using lamp 2 and the viewing screen in position 2), of AT- BT-, CT-, or DT-cut slabs. The “cut angle” is set for a CT slab.

must have a positive 38° orientation. The negative-cut, “golf-club”, figures are produced by the unusable portion of the plate.

The same procedures are followed with the AT, BT and DT plates, in each case resetting the table to the proper tilt, 35° , 49° and 52° , respectively. The reflection view of Test 1 is the same for all cuts, and the etch-figures of Test 2 are nearly the same (being almost identical for the negative-cut portions of the slabs). However, in the case of AT and CT slabs the *positive-figures* represent *good* portions (since these are positive cuts), and in the case of BT and DT slabs, the *negative-figures* represent *good* portions.

The basic principle of this instrument is as described in section 5.41. As here used, the two optical systems (including the eye and the slab) are so disposed as to obtain the best reflection-contrast in Test 1, and the most dis-

¹⁵ An excellent screen consists of two sheets of thin sandblasted cellulose acetate.

tinct portion of the etch-figures in Test 2. That the observations are so similar for this 20° range of cuts indicates that the nature of the etch-pits on these cuts is very similar, (see Fig. 5.6 for the nature of the etch-pits on AT slabs). The angular arrangement of the Test 1 optical system makes use of strongly developed facets which are approximately parallel to the X-axis and inclined at an angle of -57.6° to the Z-axis of the quartz. Within experimental error these facets are parallel to the 01.2 atomic planes and hence are called the 01.2 facets. It is also these facets that give the enlarged

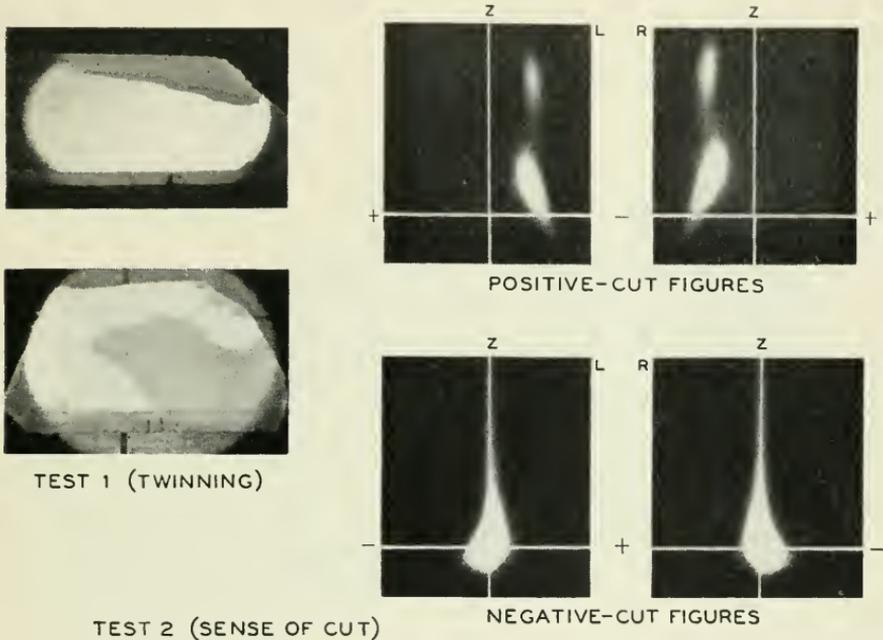


Fig. 5.14—The appearance in the twinoriascope of twinning in Test 1 (two examples) and of the four possible etch-figures in Test 2. The observance (in Test 2) of either of the positive-cut figures indicates that the illuminated portion of the slab is a positive cut, while either negative-cut figure indicates a negative cut. These etch-figures for a CT slab, are not markedly different than those for AT, BT, and DT slabs.

head of the golf-club, negative-cut figures. The right and left handedness of quartz results in two figures each for the positive and the negative orientation. Though it is commonly of no interest, it is possible to determine from the etch-figure observed, both the handedness and the electrical sense of the illuminated portion of the slab. The handedness is as indicated by L and R in each etch-figure of Fig. 5.14, and the electric axis is \pm to the right or left as indicated by the + and - signs.

Best etch-figures are obtained in the twinoriascope with fine ground (400 carborundum) slabs which have been given a strong etch (40 minutes in 50%

HF). Stronger etching is not deleterious. Very strong etching gives moderately good figures with sawn or coarse ground slabs. For Test 1, alone, weaker etches would suffice. Under properly controlled conditions of slab preparation and instrument operation Test 2 might be eliminated, for under such conditions the negative-cut portion of the slab is bright, the positive-cut portion is dark. Under shop conditions this means of detecting sense of cut appears to be not reliable, especially with *untwinned slabs* (which are either all bright or all dark). The addition of Test 2, however, gives

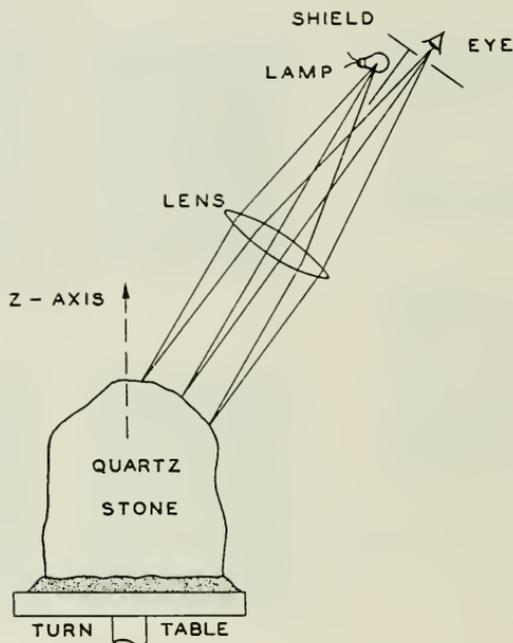


Fig. 5.15—The direction and sense of the electric axes of a sand-blasted and etched raw quartz stone may be determined by reflection of light from the 0.21 facets. These same facets are utilized in Test 1 of the twinoroscope, Figs. 5.13, 5.14.

complete reliability, for if etch-figures are obtained the sense of cut is obvious, if no figures are obtained the slab can be returned for further etching.

The principle of Test 1, above, has been applied by W. L. Bond to a laboratory instrument for determining the direction and sense of the X-axes in raw quartz stones prepared with a sand-blasted and etched surface. With the stone mounted rotateably about its Z-axis (previously determined by conoscope or inspectoscope), and a light beam properly projected onto the stone, reflection of the light beam to an eye piece or viewing screen will occur whenever the 01.2 facets come into proper angular position, see Fig. 5.15. The approximate direction and sense of the electric axis, or the sense of cuts

to be made from the stone, may be determined from these reflecting positions of the stone, and twinning may be partially explored. Thus if the stone appears to be not badly twinned, it may be cut up at once into slabs of proper sense of cut, without previously sectioning for further examination.

5.53 THE PIN-HOLE ORIASCOPE

Figure 5.16 shows a "Basic Pin-Hole Oriascope" with auxillary attachments for shop examination of etched Z-cut sections, and Fig. 5.18 the same

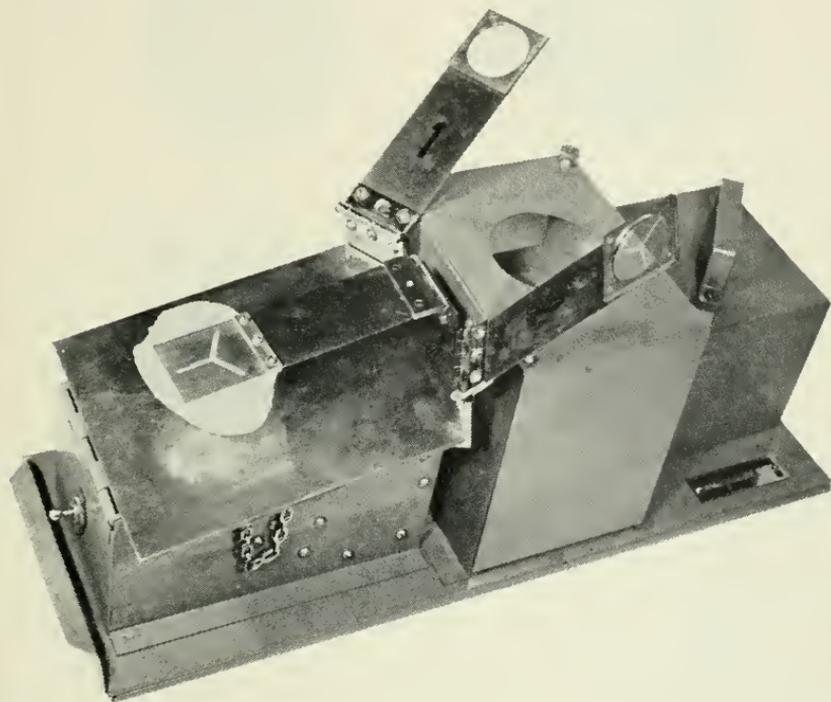


Fig. 5.16—The BASIC PIN-HOLE ORIASCOPE with matching and marking arms for use on Z-cut sections. Twinning, and the direction and sense of the X (electric) axes may be determined and marked on the section.

for X-cut sections. The optical principle of this instrument is according to Section 5.43. Light from a concentrated-filament lamp within the central ventilated housing, is projected horizontally forward by a pair of condenser lenses and reflected upward by a mirror in the forward housing, onto a diffusion-disk placed directly against the pin-hole.¹⁶ The latter is centrally located in the inclined mounting table. Etched quartz sections are placed over this pin-hole and viewed from above. The section may be moved about and examined for twinning boundaries, which are then penciled in.

¹⁶ See footnote 12.

The section is then examined through the ruled window of a matching arm, one of which is shown in use in Fig. 5.18. The section is rotated on the table until the etch-figure seen in the quartz is properly aligned with the lines on the window. Without moving the sections, the viewing arm is replaced with a marking arm, one of which is shown in place in Fig. 5.16. The section is

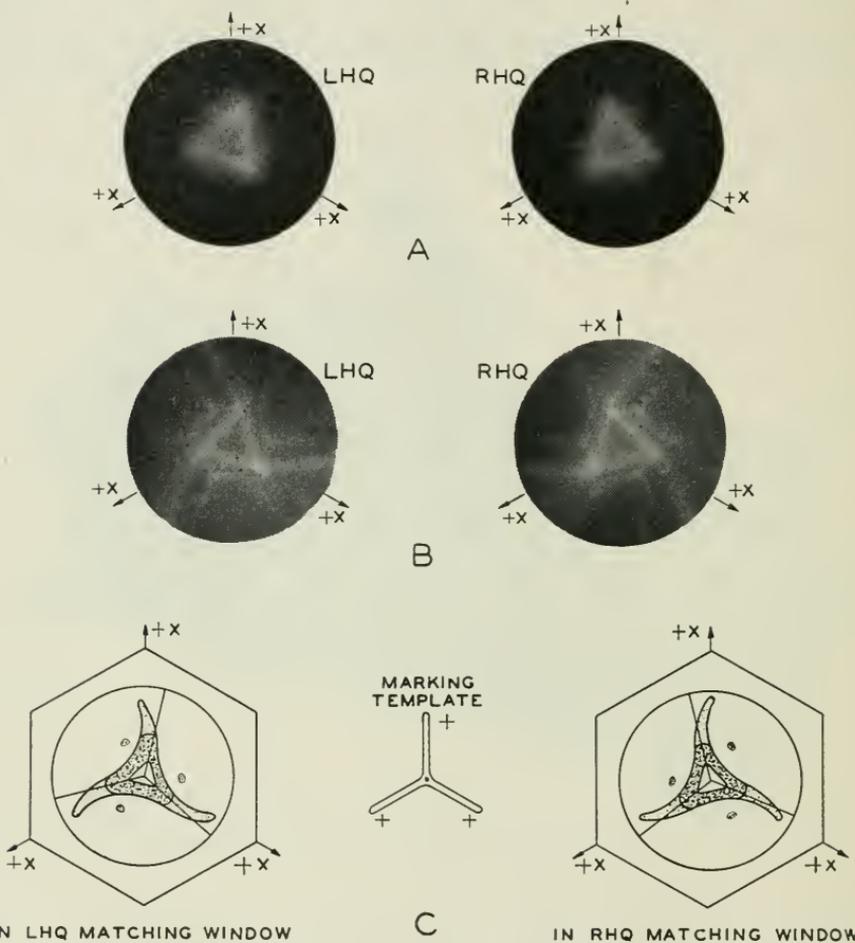


Fig. 5.17—Etch-figures obtained with the pin-hole oriascope in Z-cut sections; A for a fine ground surface and B for a coarse grind. The relation of the etch-figures to the structure orientation of the section is shown in C.

marked through the template of this arm with the desired axes or cutting directions.

Figure 5.17A, B shows the etch-figures obtained with the pin-hole oriascope, on Z-cut sections. Figure 5.17A is for a fine ground surface (600 carborundum) while Fig. 5.17B is for a coarse ground surface (100 carborun-

dum), and in both cases a moderate etch, (20 to 30 minutes in 50% HF). It is noted that the spiralling, outer tails of the etch-figures (as well as other features) denote the handedness of the quartz. Such handedness features are not as marked with fine ground surfaces, nor with weaker etches. The central triangular portion of these figures is used for alignment of the section with the rulings on the marking arm windows. Since this triangular figure is misaligned with the X-axes of the quartz by approximately 12° , and in an opposite sense for the two kinds of handedness, there are provided two matching arms. One is to be used for left quartz and the other for right quartz. The diagram of Fig. 5.17C shows the orientation arrangement of a combination of matching windows and marking template, that results in the section being marked with three radial lines which correspond to the positive X-axes of the quartz. Though this is the most obvious manner of marking Z-cut sections, it is of advantage in practice to obtain a reversed marking on left-hand quartz (by using an oppositely ruled left-hand matching window). By so marking the quartz no further attention need be paid to handedness, see Section 2.4, Chap. II.¹⁷ In either case the relation of the various plate cuts to the axis markings obtained above, may be determined from Fig. 5.4. Since the etch-figures give only approximate orientation X-rays are used for the final determination. That X-rays are not used for the whole determination is as explained in Section 5.1.

With X-cut sections, having a coarse grind (100 carborundum) and a strong etch (30–45 minutes in 50% HF), the etch-figures obtained are like those of Fig. 5.19. Here the positive face of the section gives an entirely different figure than the negative face, as would be expected from the nature of the etch-pits shown in Fig. 5.5. Opposite-handedness gives reversed figures. The four possible figures are oriented with respect to the Z-axis and the major cap face direction of the section "r" as shown in Fig. 5.19A and B. The non-parallelism of the Z-axis and the parallel sides of the etch-figures amounts to three to five degrees. This disposition of figures (relative to quartz axes) is taken into account in the design of the matching and marking arms shown in Fig. 5.18, and diagrammed in Fig. 5.19C. The etched X-cut section is rotated on the mounting table, with the central matching arm in position, until the long straight sides of the "parallelogram" figure, or the long parallel lines of the "H" figure, are parallel to the two parallel-lines ruled on the window of the matching-arm (the parallelogram figure is shown so aligned in C). The figure thus used is compared with the four figures sketched on this matching-arm, to determine which of the two marking arms is to be used for marking (note arrows giving this indication). The proper marking arm is lowered onto the section and used to

¹⁷ The instrument of Fig. 5.16 has a still different arrangement of matching and marking arms.

mark a long line approximately parallel to the optic axis and a short line indicating, in the case shown, the approximate direction and the sense of cut of a BT-plate. It is to be noted, here, that neither handedness nor electrical sense need be individually determined or considered, as such, for the sense of cut is directly obtained.

The size of an etch-figure depends upon the thickness of the section being examined, as explained in Section 5.43. For the etch-figures here presented the size of the figure relative to the thickness of the section, may be estimated

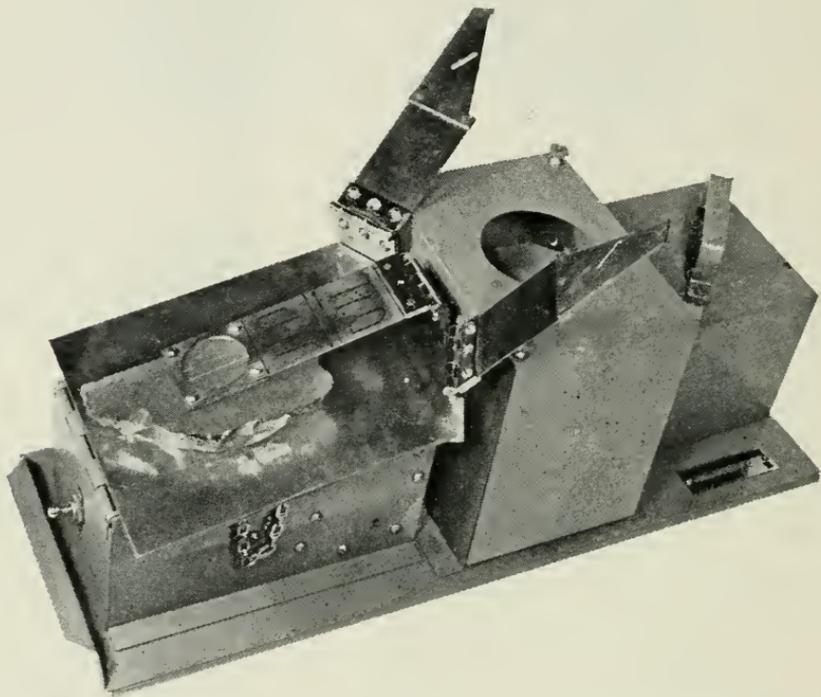
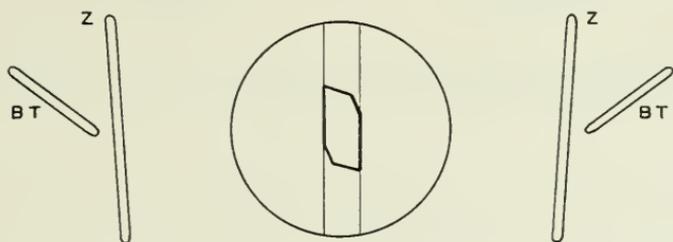
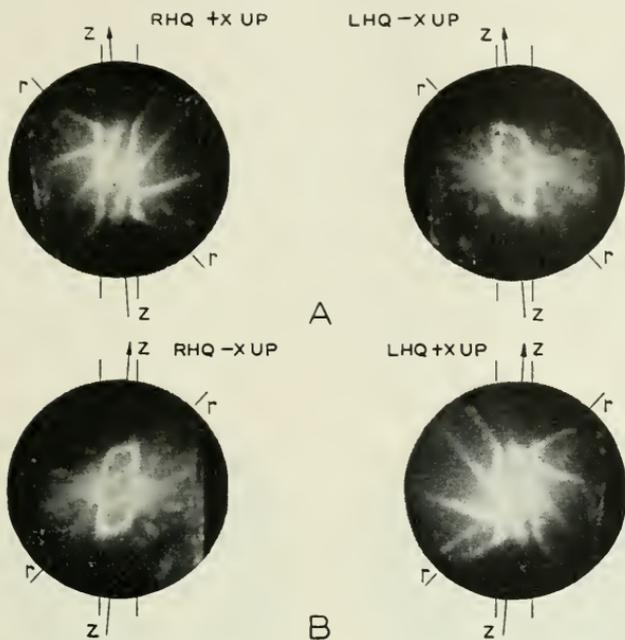


Fig. 5.18—The BASIC PIN-HOLE ORIASCOPE with matching and marking arms for use on X-cut sections. Twinning, and the direction of the Z axis, and the direction and sense of cut may be determined and marked on the section.

from a knowledge of the ratio, N , of the total diameter of the view to the thickness of the section giving that view. For Fig. 5.17A and B, $N = 1.3$; for Fig. 5.19A and B, $N = 2.7$; for Fig. 5.20, $N = 1.7$; for Fig. 5.21, $N = 2.5$.

The pin-hole oriascope may be used in a variety of other ways for examining any crystal cut with at least one etched surface. When used with sections as described above the bottom flat surface may be very small, just large enough to cover the pin-hole. However, this restricts the inspection

to an area directly over the bottom surface. This restriction may be eliminated, and no flat bottom surface need be used at all, if the bottom surface



C

Fig. 5.19—Etch-figures obtained with the pin-hole oriascope in X-cut sections. After an etch-figure is aligned with the rulings on the matching window, as in C, the section is marked thru a marking template (in this case the one on the left) with the direction of the Z axis and the direction of cut of the desired plate (in this case the BT).

of the section be immersed in a transparent dish of immersion fluid (whose refractive index matches that of quartz) placed over the pin-hole. Here the

size of the etch-figure depends on the whole distance from the pin-hole to the etched top-surface, and hence, may be made as large as desired, by raising the section and fluid level. Very thin sections, slabs or plates may be examined similarly, with the bottom surface contacting the immersion fluid, or the plates may be wet with immersion fluid and placed on thick glass plates and placed over the pin-hole. In either case the top etched-surface must be kept dry. By this means the twinoroscope examinations described in Section 5.52 might be performed on the pin-hole oriascope, (a disadvantage being the necessity of using an immersion fluid).

Usually etch-figures are obtained from flat etched surfaces whose orientation is known within 5° . However, if the surface be 10° to 20° off-orientation the etch-figure will be plainly distorted. If now the section be viewed at an angle to the normal position, or if the section be tilted in the fluid-

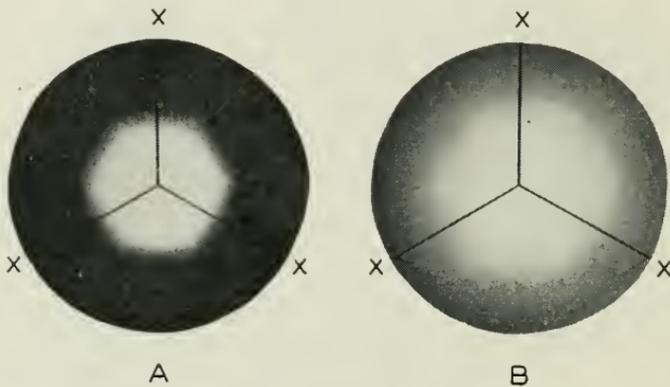


Fig. 5.20—CLEAVAGE-FIGURES may also be observed on the pin-hole oriascope in ground but unetched specimens, in this case a Z-cut section. Here the direction of the X axes but not their sense (nor handedness, nor twinning) may be determined.

both method described above, the undistorted figure may be observed. The direction and amount of misorientation of the surface may be thus estimated. By provision of suitable mounts and scales the misorientation could be measured to 5° .

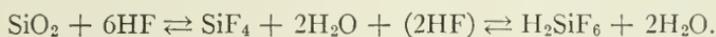
It might be added that in some cases unetched, ground (or sawn) quartz surfaces give "cleavage-figures." Thus with Z-cut sections which have been ground, but not etched, there may be observed on the pin-hole oriascope cleavage-figures like those shown in Fig. 5.20. The difference between the two views is mainly a difference in focusing and in photographic reproduction. The cleavage-figure indicates that there are preferential cleavage planes in quartz, which are parallel to the X-axes, and correspond approximately to the natural cap faces. Further, there is no indicated difference between the major and minor planes. Thus, the cleavage-figure is

six-fold and may not be used to determine electrical sense or twinning. It may, however, be used to determine approximately the orientation of the X-axes. Cleavage-figures are seldom strong, but appear to be best with coarse grinding.¹³

5.6 THE PROCESS OF ETCHING QUARTZ

Few factors related to the chemical process of etching quartz have been extensively studied. Much of the information here presented is taken from preliminary reports of L. Egerton of the Laboratories, who has undertaken an investigation of the etching process. Though the information mainly regards hydrofluoric acid etching, some data is given on etching with hydrofluoric gas, and bifluoride mixtures.

The reaction of quartz, which is silicon dioxide (SiO_2), with hydrofluoric acid (HF) is given by the following equations:



Since the hydrofluoric acid is a solution of HF gas in water, the reaction of the acid with quartz results in a reduction of the concentration of HF. At the same time there is produced silicon tetrafluoride (SiF_4) which reacts with more HF to give fluosilicic acid (H_2SiF_6) in solution. It is common practice to start with about 50% HF acid and to continue etching until the HF concentration is down to 20 or 25%, at which time there should also be a 30% to 35% concentration of H_2SiF_6 , if all the depletion of HF were due to reaction with the quartz. Actually much smaller concentrations of H_2SiF_6 are found, and this discrepancy is mainly due to the large continuous loss of HF from the solution by gassing. Further, the etching power of this used acid is not the same as would be obtained with a solution of 20%-25% HF alone in water. However, this difference is hardly noticeable except with weak etches.

Through the useful life of the acid, starting with 50% HF and depleting to about 20% HF, practically identical etch-figures may be obtained by properly adjusting the etching time. Means of testing the etching power of the acid to determine the proper etching time are complicated by the production of H_2SiF_6 in the solution, and by the irregular loss of HF by gassing. Further, the power of the acid to produce useable etch-figures is not the same as its power to remove quartz, or to etch glass, or as its concentration of HF or H_2SiF_6 . For these reasons any indirect method of measuring etching-power must be correlated empirically with the etching-time required to give the desired etch-figures.

An indirect method of testing the etching-power, developed by Dr. W. Hoff of Western Electric, Hawthorne, involves the etching of sand blasted

¹³ Scrubbing the surface with soap, water, and brush sometimes improves the figure.

microscope slides for a standard length of time. The lead-glass slides become coated with a white lead-fluoride deposit to a depth dependent mainly upon the HF content of the acid. The optical density of this deposit is measured with a specially adapted photometer. The photometer readings are correlated with required etching-times to give the desired etch-figures; a different etching-time being required for different kinds of sections, slabs, etc. Use of this means of controlling the etching time has greatly improved the regularity with which good etch-figures are produced in the shop.

Commercial hydrofluoric acid from a number of different suppliers has been analyzed for purity, and tested for the development of etch-figures. It appears that when such acids are brought to the same concentration (by addition of water if necessary) there is no difference in their effectiveness, nor are they inferior to pure reagent acid. Commonly the acid is supplied as 48% solutions in lead or hard rubber drums, or as 60% in steel drums (usually the concentration is a few per cent higher than labeled). The difference in packaging is of no importance in the results obtained, provided the concentration is properly reduced.

There are two important factors regarding the starting concentration of hydrofluoric acid baths. In the first place, acids stronger than 50%, though reacting vigorously with the quartz (and removing material rapidly), do not give good etch-figures. Secondly, strong acids not contained in sealed containers lose strength very rapidly by gassing of the HF gas. Hence unused fresh acid should be kept well stoppered. Before use the acid should be diluted to a concentration of 45% to 50%. This may be accomplished by adding about $\frac{1}{3}$ volume of water to one volume of 60% acid, or $\frac{1}{6}$ volume of water to one volume of 55% acid.

Concentrated hydrofluoric acid loses HF by gassing more rapidly than it loses water by evaporation. This preferential loss of HF continues until the HF concentration is reduced to 35% or less,¹⁹ and is not completely overcome by covering the bath without sealing. In fact, in practice, it appears that about as much HF is lost by gassing as is used in etching the quartz. Thus the bath should be kept as tightly covered as is practicable.

Whereas, in the past only lead and hard rubber have been used for fabrication of acid baths and racks, it appears that for concentrations not greater than 50% HF, copper, nickel, and brass may be used as well (steel is inferior at low concentrations). Lead-tin solders may not be used, but silver solder is satisfactory. Thus shop acid equipment may be easily fabricated out of common fabricating materials.²⁰

¹⁹ At room temperatures there appears to be a constant-concentration mixture at some concentration below the 35% concentration of the constant boiling mixture, the exact value depending upon the temperature of the solution and the ambient humidity.

²⁰ Polystyrene is a good material for use in fabrication of vessels for handling HF and its reaction products in the laboratory.

While agitation of the acid bath during etching does speed up the removal of quartz from the surface, it does not appear to speed up the development of the etch-figures here considered. However, moderate agitation does improve the uniformity of etch from one crystal to another, and even over the surface of single large surfaces (especially when such surfaces are close together). Uniformity of etch is important in examining for twinning. The surfaces to be etched should never be placed in contact with each other, or with other surfaces, so that the acid cannot flow between them (the separation should be at least $\frac{1}{32}$ of an inch).

The effect of temperature on the etching process appears to be small for the range of room temperatures normally encountered in practice.

A word of caution should be added regarding the handling of hydrofluoric acid and other fluorine etching materials. The dangers are of two kinds. First, fluorine poisoning may result from contact with any fluorine compounds, the effects of which may be cumulative. Special care should be taken to prevent inhalation of vapors from all etching baths containing fluorine. Some persons are especially sensitive to fluorine poisoning. Secondly, hydrofluoric acid baths, or any baths containing free HF, may produce acid burns. Commonly such burns are attended by fluorine poisoning. For these reasons etching with all fluorine compounds is preferably carried out in ventilated hoods (with strong air suction through the door), with continually running water for washing, and with rubber gloves, tongs, racks, etc. for handling the quartz.

Etching compounds other than hydrofluoric acid have been widely used in etching glass, as is evidenced by the variety of formulae presented in the "Chemical Formulary."²¹ Solutions of ammonium bifluoride (NH_4HF_2), with additions of various amounts of free hydrogen fluoride, sodium bifluoride, sugar, and other materials have long been used on glass. One of the possible advantages of such formulae for etching quartz is the elimination of the dangers of acid burns and strong fumes that may be obtained with hydrofluoric acid (care must still be maintained to prevent fluorine poisoning). A number of these formulae have been made up and tested on quartz. The preliminary conclusions are as follows.

The etch-figures that may be developed by the bifluoride compounds on Z and X-cut sections of quartz are not the same as those developed by hydrofluoric acid. The results approach each other, however, for excessively long etching in both cases. To obtain *usable* etch-figures on X-cut sections with the bifluoride requires considerably longer etching time than with hydrofluoric acid, or an elevation of the bath temperature to about 45°C. The addition of hydrofluoric acid to the bifluoride formulae speeds up the development, but partly negates the safety advantage of the bifluoride bath.

²¹ Published by the Chemical Publishing Co., Brooklyn, N. Y.

The figures produced on Z-cut surfaces are small and complex, (hardly usable) unless a considerable amount of free HF acid is added. Etch-figures here considered are those produced on the pin-hole instrument, and are usable only if they have such character as will permit of their use in determining quartz axes. Fig. 5.21 shows the type of usable etch-figure obtained on X-cut sections with an ammonium bifluoride and sugar solution (the sugar is here effective mainly in preventing creepage of the solution). It might

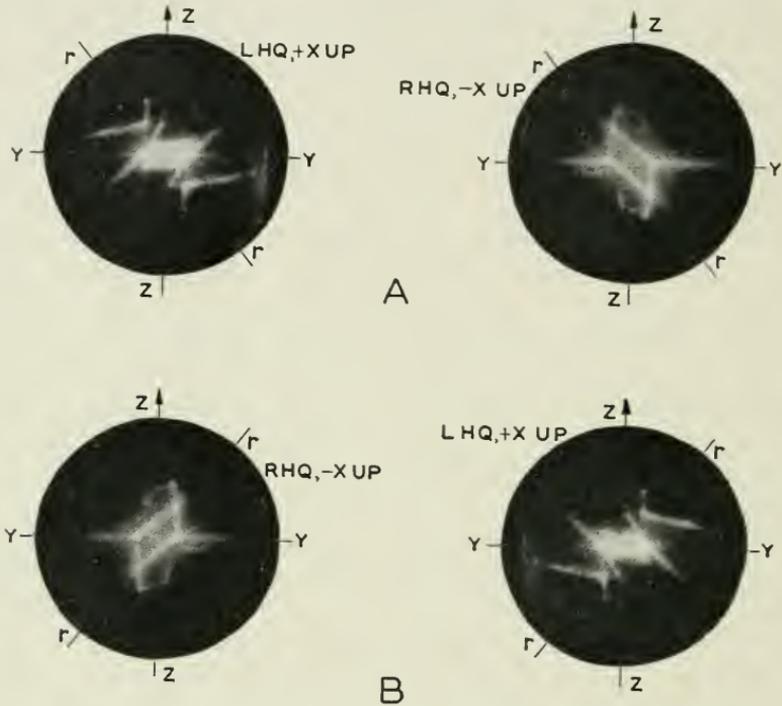


Fig. 5.21—Etch-figures obtained on the pin-hole oriascope with X-cut sections which have been strongly etched in bifluoride mixtures, or excessively etched in hydrofluoric acid. These etch-figures differ from those of Fig. 5.19 (for a moderate etch in hydrofluoric acid) but are obviously usable.

be noted that a similar figure is obtained with hydrogen fluoride gas, and with *excessively* long etching (several hours) in hydrofluoric acid.

When the bifluorides are used only to develop reflection contrast in the detection of twinning, their effectiveness appears to be about the same as hydrofluoric acid, under equivalent process conditions. The etching power of the bifluorides may be maintained nearly constant over a long period of use by maintaining an excess of the salt in solution, a distinct advantage over the acid. The metals copper, nickel, brass and stainless steel may be used in fabricating tanks and racks, lead and steel are inferior.

Finished quartz surfaces are sometimes etched to remove surface debris

(fragments of quartz loosened by grinding, and grinding refuse embedded in microscopic surface irregularities), and to remove predetermined small amounts of the surface for frequency adjustment. It is common for these purposes to use weaker etching solutions, since very small amounts of quartz are to be removed. With hydrofluoric acid, weak solutions (less than 20% HF) have an advantage in that their concentrations are little reduced by exposure to the air. In fact with very weak solution the concentration may increase slightly by exposure, and thus partly compensate for the HF lost by reaction. Weak ammonium bifluoride solutions may also be used, provided no deposit forming material is added.

5.7 THE EFFECT OF TWINNING IN THE FINISHED PLATE

While it is commonly considered that electrical and optical twinning are not allowable in a finished oscillator plate, it cannot be unconditionally stated that small amounts of twinning will too seriously affect the properties of all types of oscillator plates. The allowance of even small amounts of twinning in the finished plate would save quartz and simplify the processing procedures. Hence, consideration must be given to the factors which would affect the utilization of twinned material, and the effect of twinning on the operating characteristics of the finished plate. Consideration will first be given to the nature and distribution of electrical and optical twins²² in the raw quartz.

The analysis of twinning in raw quartz has been carried out by the examination of numerous, etched Z-cut surfaces. By the method to be described it is possible to detect the handedness, and the axial orientation and sense, of each homogeneous portion, twin, appearing at the etched surface of a twinned specimen. Both electrical and optical twins may be analyzed by this method. It might be added that electrical twinning boundaries and orientation are only detectable at an etched surface, and that while interior optical twinning may be detected by polarized light, its exact analysis is only possible at an etched surface.

Figure 5.22 E shows the optical arrangement used for examining twinning in etched Z-cut sections. The sections (prepared with a fine grind and weak etch) were mounted on a turntable, illuminated from an elevation of about 30° to the horizontal etched surface by a spot lamp, and viewed (or photographed) from vertically above the section according to principles of Section 5.41). With the section properly aligned on the table (with the predetermined electric axes parallel to the table-lines joining diametrically opposite fiducial marks), the table was successively turned into positions about 12° to the right or left of the plane of illumination and reflection (as indicated by the R and L marks and the index pointer). Four of these positions of

²² See footnote 4.

illumination of a given section are sufficient to determine the nature of the four possible twins in the section. The four corresponding photographic views of the section have been arranged in a special manner to simplify their explanation. This arrangement, as shown in Fig. 5.22A, B, C and D, is equivalent to what would be observed if one looked down on a single, stationary section, and illuminated the section from the four different direc-

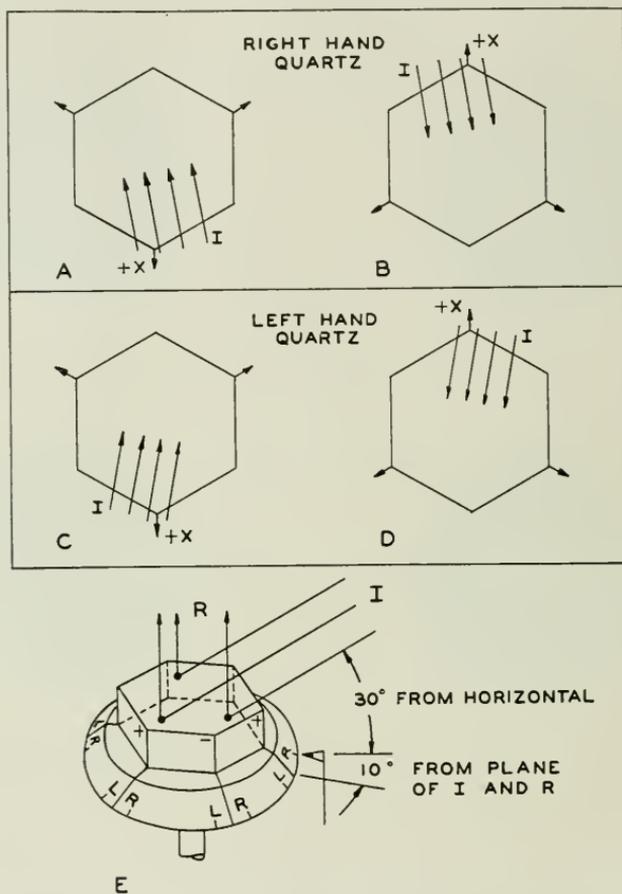


Fig. 5.22—Reflection patterns of the twinned, Z-cut sections shown in Figs. 5.23, 5.24, 5.25 and 5.26 were obtained by the means shown in E. A, B, C, and D are a key to the four equivalent directions of illumination of a single stationary section.

tions shown in the figure. For each direction of illumination there is a corresponding view, the outline of the section (and any cracks, chips or other flaws) being identically positioned in each view. However, when the four types of twins are present in a given section, each view will show a different region, or regions, of brightness. For each view, the interpretation of handedness and electrical sense of the *bright portion* of the view is according

to the labeling of this particular view, only. Thus if a section is entirely right quartz and of the electrical sense shown at A the whole surface of the section will appear bright in view A and dark in all other views. If a section is all right quartz, but partly of the electrical sense shown in A and partly that shown in B, then part of the surface will appear bright in A and the other part will be bright in B (the whole surface will be dark in C and D).

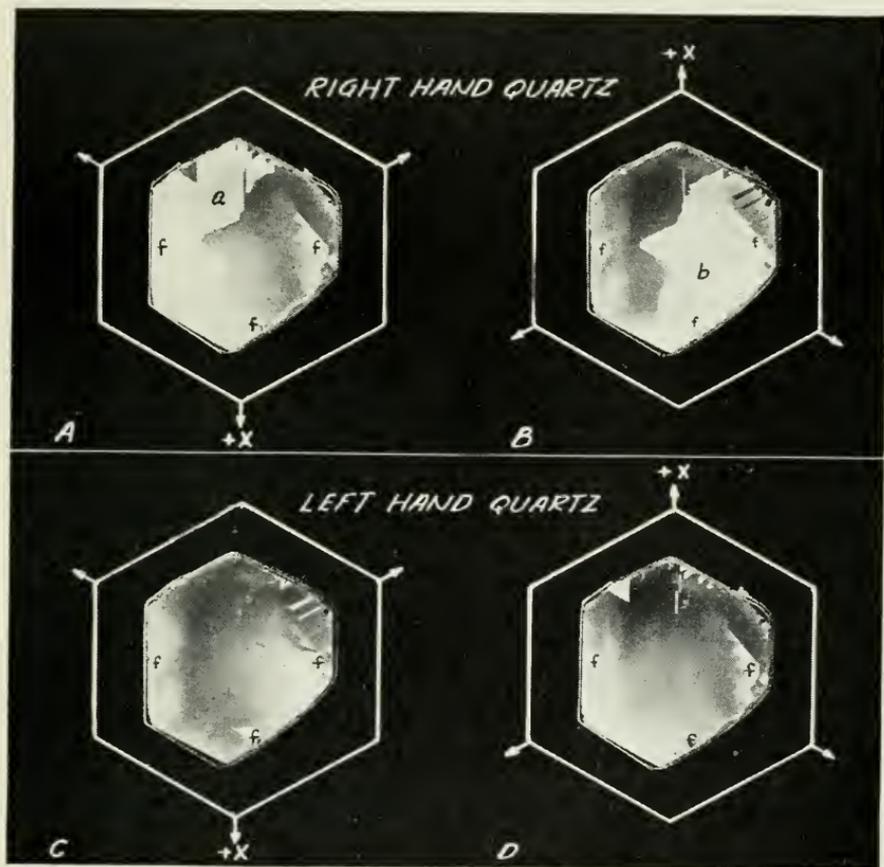


Fig. 5.23—The four possible conditions of handedness and electrical sense in a single section are shown here. In each view the handedness and sense is for only the *bright* portion of that view. The *a* and *b* regions are seen to be both of right quartz but of opposite electrical sense, hence electrical twins. (Flaws indicated by *f* are to be disregarded).

A section containing all four possible twins would exhibit bright regions in each view, and a different bright region in each view. All bright regions would fit together to make a complete map of the surface. Only the bright portion of each view has the handedness and electrical sense indicated for that view.

Figure 5.23 shows a Z-cut section containing twins of the four possible

conditions of electrical sense and handedness. The two large, bright regions a and b (appearing in views A and B respectively) are both right quartz but of opposite electrical-sense. Hence the surface is mainly of electrically twinned right quartz. The small dark regions within the borders of a (view A) are bright in view D. Hence these small, triangular and line regions are left quartz and of opposite electrical sense to the large region a containing them. They are then optical twins of the large a region. Similarly the dark regions of b (view B) are found from view C to be optical twins of the b region. (Flaws labeled f are cracks, chips, etc.) If the whole section were cut up to make AT plates, for example, and at the proper angular sense according to the a portion of the section, then those plates coming from the b region would be of wrong angular sense. Those crossing a boundary between the a and b regions would be of both senses, i.e., electrically twinned. Those few plates which contained some left quartz would be optically twinned. To make the most economical use of this section it should be separated, by cutting along a line approximating the a to b boundary, so that each half of the section may be cut at the correct sense of orientation. Even when so cut, some of the plates will contain optical twinning and remnants of electrical twinning. This section is typical of much of the raw quartz that must be used for manufacturing piezoelectric plates.

Figure 5.24 shows a section which is mainly of left quartz as exhibited by the large bright c and d regions of views C and D. The large c region is optically-twinned to a small extent by the line regions b of view B. One of the d regions is badly optically twinned by the small striated a regions, as seen in A. Such a section would be very uneconomical to process, since separating the larger electrical twins is not feasible. If processed at all, it should probably be entirely cut according to the handedness and sense of the large c portion, the wrong-sensed regions and twinning being cut away at a later stage (after inspection of the slabs in the twinoroscope, for example). It might be noted that only the optical twinning could have been observed in the initial polarized-light, raw quartz inspection, where such a stone would be passed as moderately good.

Fig. 5.25 shows an *unusual* section that is mainly composed of left quartz, regions c and d . The right quartz regions shown in view B are of both opposite-handedness and electrical-sense to the c region inclosing or bordering them. This is the common and expected conditions. The unusual condition is exhibited by the regions c and a , where twins of opposite-handedness but *same* electrical-sense have a common boundary. Since this boundary could be detected by optical means, the a and c regions might be described as optical-twins, of an "uncommon variety". However, by convention *optical twinning* has long been used to denote twinning exhibiting both opposite-handedness and opposite-electrical-sense (crystallographically,

Brazil twinning). Further, twinning exhibiting both opposite-handedness and same-electrical-sense, combines the crystallographic twinning laws of Brazil twinning and Dauphiné (electrical) twinning. Hence this uncommon variety of twinning may preferably be called combined electrical and optical twinning, or just COMBINED TWINNING. Thus, the boundary

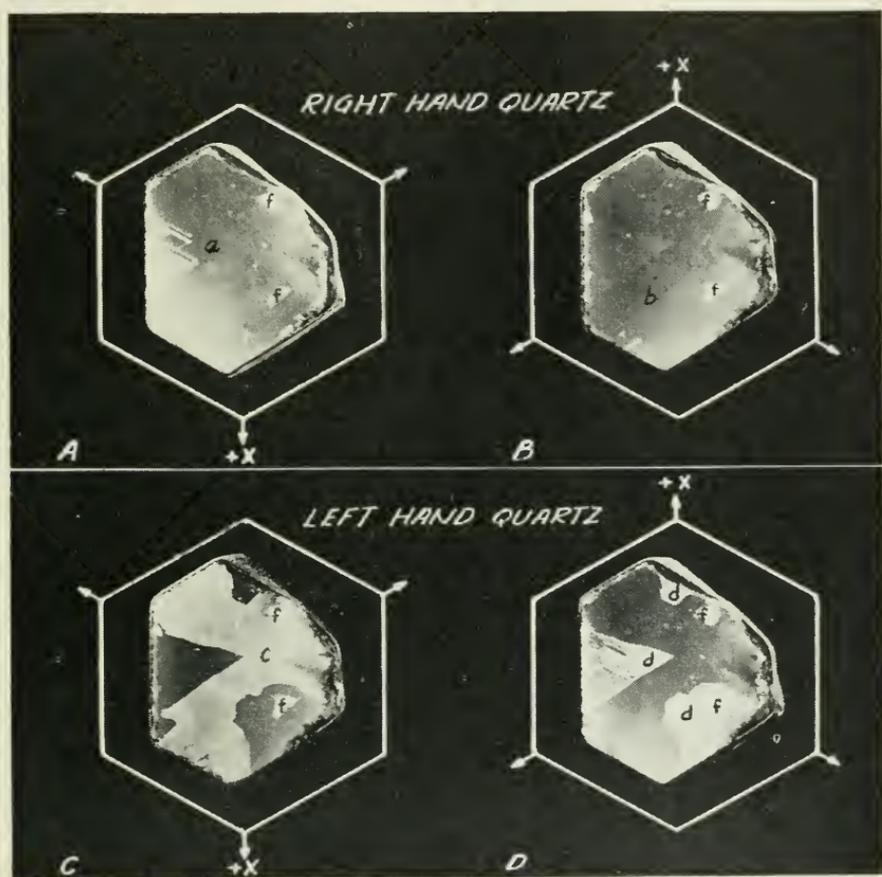


Fig. 5.24—Regions *d* are electrical twins of the region *c*. The striated regions *a* are of opposite handedness and electrical sense to the *d* region enclosing them, hence optical twins of *d*. The *b* regions are small optical twins of *c*, and *f* are flaws.

between the *a* and *c* twins separates combined twins. Note also that the *a* twin bounds the *b* twin and the *b* twin bounds the *c* twin. Thus, *a* and *b* are true electrical twins, and *b* and *c* are true optical twins.²³

²³ It is possible that growth conditions are such that combined twinning cannot occur by itself, without the presence of true optical twinning and true electrical twinning. That is, a region of given handedness and sense can not be entirely bordered by a region of opposite-handedness and same-sense.

Figure 5.26 shows an unusual section which is mainly composed of left quartz, of the electrical sense shown in D, region *d*. The region *c* is an electrical twin of *d*. The region *f* is a flaw in the quartz and is to be disregarded. The region *a* is an optical twin of *d*, and is *uncommonly* large for an optical twin (note: region *a* contains within it, two small optical twins). Since

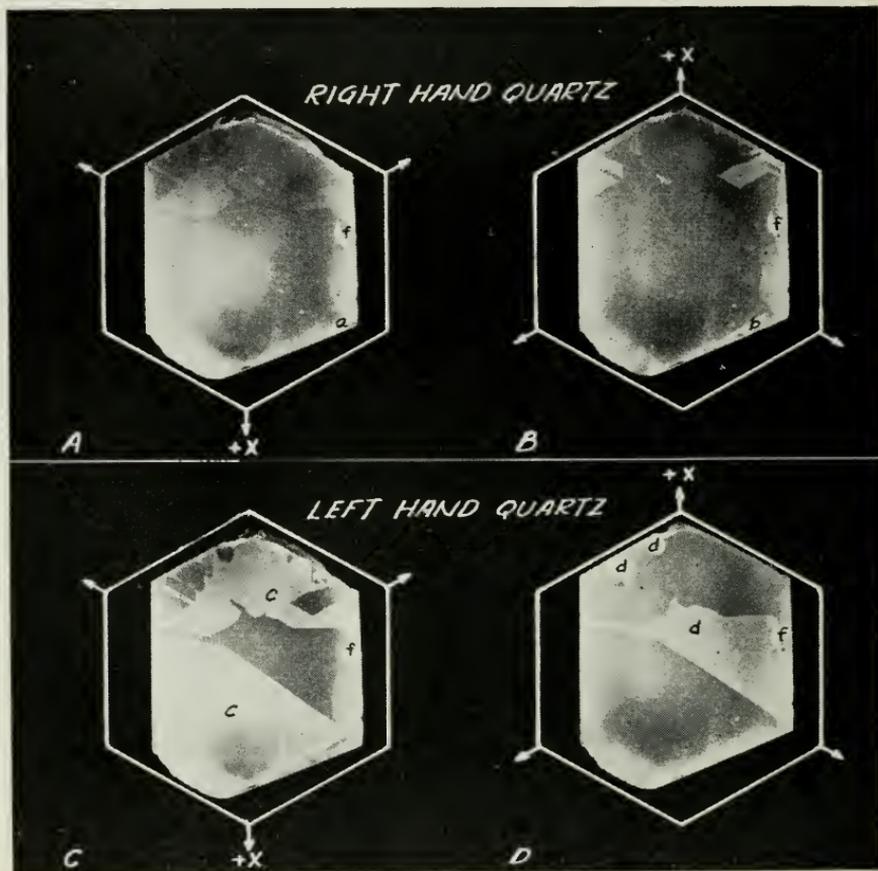


Fig. 5.25—Regions *c* are electrical twins of the adjacent *d* regions, *a* is an electrical twin of *b*, and *a* is also an optical twin of *d*. An uncommon condition of twinning is presented by the adjacent *a* and *c* regions which are of opposite handedness but the same electrical sense, thus exhibiting COMBINED-TWINNING.

optical twins are usually very small (except for the one major surrounding twin), it is seldom possible to cut them apart and use each twin individually.

Figures 5.1, 5.2 and 5.3 were obtained by the means above described, and all sections shown in these figures (except Fig. 5.2A and C) actually exhibited both electrical and optical twinning. Thus Fig. 5.3D was obtained from Fig. 5.24A, and Fig. 5.2F from Fig. 5.24C, etc., by trimming the latter

named figures to give the sections simulated natural faces. Figures 5.2 and 5.3 are of particular use in learning to distinguish between electrical and optical twinning when examining etched surfaces by reflection. Note that electrical twins are usually large and separated by irregular boundaries, Fig. 5.2. Optical twins are usually separated by straight-line boundaries

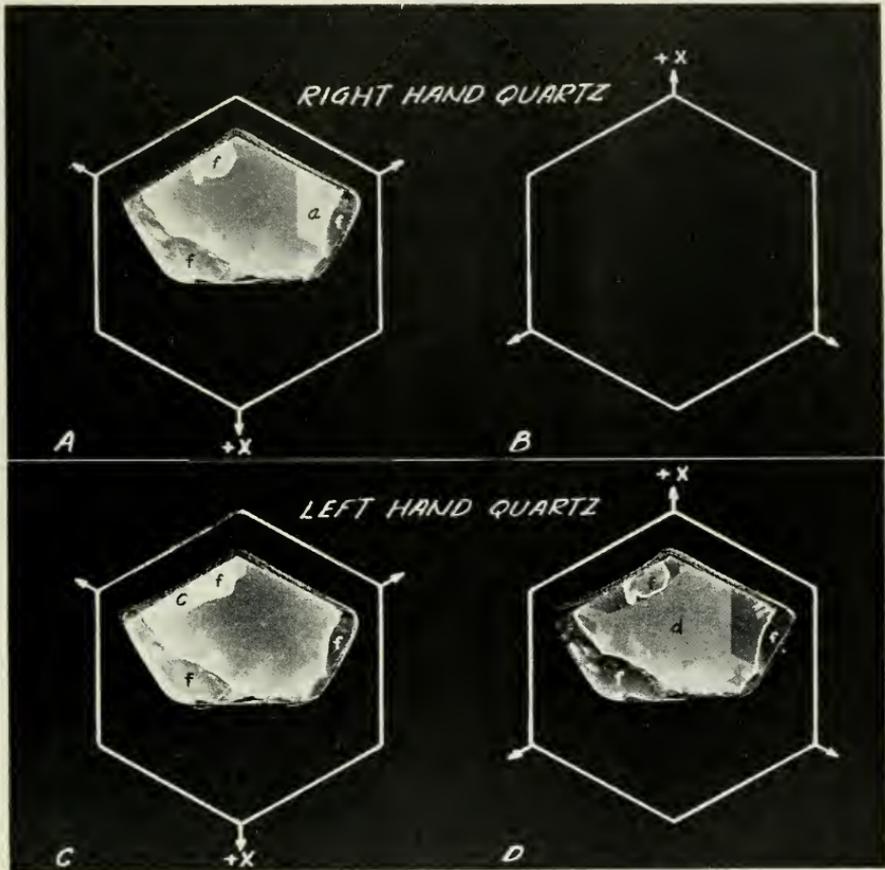


Fig. 5.26—Since this section exhibited no bright regions (except flaws *f*) in view B (i.e. no right quartz of electrical sense B) it was not reproduced in view B. The *c* region is an electrical twin of the adjacent *d* region, while *a* is an optical twin of *d*. It is *uncommon* for a minor optical twin to be as large as *a*.

parallel to natural faces, thus forming triangular, parallelogram, and straight line insets, Fig. 5.3. Optical twins (except for the one major, surrounding twin) are usually very small and often interlayered (with the major twin). Large interlayered regions are entirely unusable and hence are cut away at the earliest possible stage to save the labor of processing worthless material.

Small optical twins and small electrical twins (or remnants of electrical

twins left after cutting electrical twins apart) may be isolated or removed in an intermediate or late stage of processing, where they are detected by the etch technique. Commonly the final rejection of material twinned in either way is delayed until after the final blanks are cut out. These may be etched and examined by reflection, one at a time under a spot lamp, and those showing twinning (and other imperfections) sorted out and rejected.

Another possible method of rejecting twinning which is of sufficient amount to be harmful is by making electrical tests on the finished (or semi-finished) plates, at which time those plates failing to meet the electrical tests for any reason (including twinning), are rejected. While this method of rejection does not assure that twinning will be entirely absent from the accepted plates, neither does any other method assure complete absence of twinning. Further, except for imperfections which may affect the useful life of the plate, acceptance of finished oscillator plates is not illogically based

TABLE I.—*Constants for Plates of Correct and Incorrect Sense of Cut*

Cut, Angle	Frequency Constant (fxd. in Kc. mm.)	Temperature Coefficient (parts/10 ⁶ /C. ^o)
AT +35° (-35°)	1670 (2400)	0 (+30)
CT +38° (-38°)	3080 (2100)	0 (-30)
BT -49° (+49°)	2560 (1880)	0 (-55)
DT -52° (+52°)	2060 (2850)	0 (+45)

upon their meeting the desired electrical operating characteristics, i.e., frequency, temperature-coefficient, activity and internal damping (all determinable by electrical tests).²⁴ It does not appear that twinning will affect the useful life of the plate. Its effect upon the electrical operating characteristics of the plate depend upon many factors.

An important factor regarding twinning in the finished plate is that optical twinning introduces a less important variation in the physical properties of the plate than does electrical twinning. Thus, in the case of optical twinning alone, both portions of the plate are of the same sense of cut, though still being of opposite electrical sense. This may be understood from an examination of Fig. 5.4, the second and third views taken together represent optical twinning. In the case of electrical twinning the two portions of the

²⁴ With filter plates additional operating characteristics must be met. The ratio of capacities (see Chap. I, Appendix A.3) is greatly affected by the opposed electrical sense of twinning.

plate are of both opposite sense of cut and opposite electrical sense, as may be observed from the third and fourth views of Fig. 5.4. The effect of this difference in sense of cut for the two types of twinning is brought out by Table I, which gives the approximate frequency constants and temperature coefficients for the common cuts of oscillator plates, together with those for the analogous, oppositely (and hence wrong) sensed cuts.

In the case of a CT plate, for example, both portions of an *optically twinned* plate (cut at $+38^\circ$) will be of the same $+38^\circ$ orientation. The plate is elastically the same throughout and hence should exhibit the frequency and low temperature-coefficient desired. However, the opposed electrical senses of the two portions will cause a reduction in the electrical activity. The amount of this reduction will depend upon the relative size of the two portions and upon their placement relative to the vibration nodes of the plate.

On the other hand, when a CT plate is *electrically twinned* one portion of the plate will be of the correct $+38^\circ$ orientation while the other portion is of the incorrect -38° orientation. The two portions of the plate have widely different elastic properties, as is exhibited in the table by the different frequency constants and their respective temperature-coefficients. Resulting from this difference alone, the plate will exhibit operating characteristics (if operable at all) intermediate between the two listed in the table (usually near one of these two), and its activity will be reduced. The activity will also be reduced by the opposite electrical senses in the two portions. The degree to which the frequency, temperature-coefficient, and activity are affected, again depends upon the relative sizes of the two portions of the plate and their placement relative to the "nodes" of the plate.

Thus, for equivalent proportions and placement of twinning, electrical twinning will cause a much greater change in the operating characteristics of the plate than will optical twinning.²⁵

A note may be inserted regarding the electrical testing of plates, some of which may be twinned while others may be untwinned but of incorrect sense of cut. As seen from Table I, untwinned plates of the correct sense of cut are easily distinguished from those of the incorrect sense of cut by their frequency. This distinction between sense of cut holds as well for plates containing very little twinning. The presence of appreciable twinning in the plate is easily distinguished by the activity of the plate. While ordinarily a plate would be electrically tested in the mode of vibration it is intended to be operated in, it is sometimes of advantage to test it in a different mode.

²⁵ In the case of the *uncommon* "combined-twinning" the two portions of the plate are of opposite sense of cut but of the same electrical sense. The effect on the operating characteristics will be like that for electrical twinning, except that the activity may not be as greatly reduced.

Thus the high-frequency mode plates (AT and BT) might be tested in their low frequency modes (corresponding roughly to the CT and DT modes, respectively). A further discussion of this matter will be found in a later chapter by I. E. Fair.

5.8 CONCLUSIONS

In the processing of quartz, consideration must be given to the nature of twinning and to its characteristic distribution in the raw stone. There are only two common types of twinning that need be considered, namely *electrical twinning* and (true) *optical twinning* ("combined-twinning" and other types are a rarity). Due to the characteristically large size (and the nature) of *electrical twins*, a stone must be examined for electrical twinning (by the etch technique) at an early stage of processing so that the electrical twins may be observed and cut apart before the angular cuts (AT, BT, CT, DT, etc. slabs, bars, or wafers) are made. Otherwise, some of the large electrical twins will be entirely cut up with the incorrect angular sense, and hence wasted.

On the other hand *optical twins* are characteristically small and interlayered, or small and scattered. The interlayered regions are entirely unusable. Hence processing labor will be saved by inspection of the raw stones (by the polarized light means of Chapter IV), and of the first sections at least (by the etch technique) for large regions of interlayered optical twinning.

Scattered optical twins and small electrical twins, or remnants of electrical twins which have been cut apart, may be cut away in an intermediate processing stage, or in a later stage plates containing such twinning may be separated out. In either case the etch technique may be used to detect the twinning.

An alternative method of eliminating small electrical twins (or remnants thereof) and of small optical twins (most of which are characteristically very small) is by electrical tests on the finished plate. This method has merit in that if the twins are sufficiently small, and not disadvantageously placed in plate, they may not harmfully effect the desired operating characteristics of the plates. The degree of the effect depends not only upon the size of the twin and its location in the plate, but upon whether the twinning is electrical or optical; *optical twinning* being considerably *less harmful* than electrical twinning. The effect of the twinning further depends upon the type of plate being considered, i.e. its size and mode of operation, and use. It is probable that twinning is more tolerable in low-frequency mode oscillators (CT and DT) than in the high frequency modes (AT and BT), and of course more tolerable in plates of low requirements on the operating characteristics (activity, frequency and temperature-coefficient). Twinning is

probably least tolerable in filter plates, which have to meet very special requirements.²⁶ Detailed experimental studies of allowable amounts of twinning are of little value since to use the results in a manufacturing process would require a careful inspection of each plate and a difficult classification into groups depending upon the variety, amount, and placement of the twinning. Acceptance or rejection of finished plates on the basis of their final electrical operating characteristics appears to be the only practical means of separating useably twinned plates from unusably twinned plates. This method of selection does not determine whether the rejected plates contain twinning or other imperfections (or are misoriented or misdimensioned) and is therefore of little use in analyzing the processing methods to determine best practices. This disadvantage may be eliminated by etching the *rejected* plates and examining them for twinning (and such other imperfections as show up best after etching).

The effects of crystal imperfections other than twinning were discussed in Chapter IV, Section 4.9.

²⁶ See footnote 24.

CHAPTER VI

Modes of Motion in Quartz Crystals, the Effects of Coupling and Methods of Design

By R. A. SYKES

6.1 INTRODUCTION

WITH the recent extended use of Quartz crystals in oscillators and electrical networks has come a need for a comprehensive view of the various types of crystal cuts. In addition there has been a need for illustration of some of the methods employed in choosing the proper cut for a given requirement, the manner in which quartz crystals vibrate and the basic principles governing the choice of a design to use certain cuts most advantageously. In particular one of the greatest problems associated with the recent large scale production of crystals for oscillator purposes has been that of obtaining crystals the activity and frequency of which would not vary to any large degree over a wide range in temperature.

It is the intention of this chapter to present a physical picture of the manner in which quartz crystals vibrate in their simplest forms and then to show what has been learned from these simple forms that will apply to the more complex combinations of motion. The motion of a bar or plate is determined almost wholly by its dimensions and the particular type of wave generated, or frequency applied, and very little upon the driving system if the coupling to the driving system is small. In the case of quartz the coupling between the electric and mechanical system is small and hence we may study the motion of rods and plates without always considering the effect of changes due to the method of excitation (i.e., piezo-electric). However the ease of exciting and measuring a particular mode does depend on the piezo-electric constant driving it. Basically only three types of motion will be considered; flexural, extensional and shear. These three types of motion or combinations of these can be considered to represent most of the cases with which we will concern ourselves. In addition, the frequency equations will be given for common types of motion and the effect of coupling between various modes of motion. Finally the general rules relating to the dimensioning of oscillator plates will be presented.

6.2 TYPES OF MOTION IN QUARTZ RODS AND PLATES

6.21 *Flexural*

The motion associated with flexure will be discussed first because this is the type of motion that we see more commonly in nature. This motion is

the type which presents itself in the xylophone, the chime type door bell, and various other vibrating reeds or bars. Fig. 6.1 shows the general type of motion of a bar free to vibrate in flexure. The displacement takes place in the direction of W and the wave is propagated along the length. A flexure mode is one in which the center line does not change length. The type of motion associated with the first order, or fundamental, of a bar free to vibrate on both ends is shown in Fig. 6.1 with a dotted figure superim-

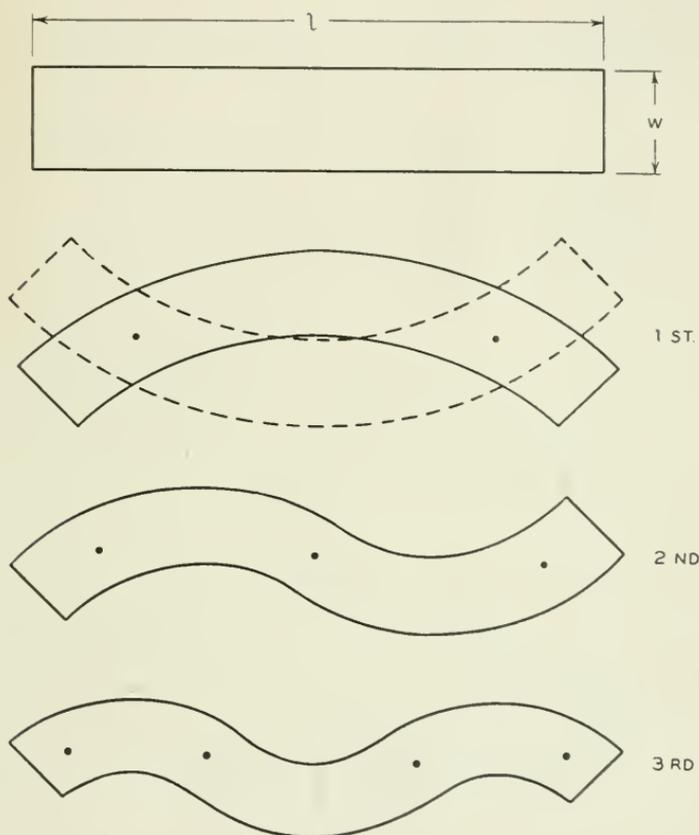


Fig. 6.1—Motion of a bar in free-free flexure.

posed to show the motion in the opposite phase. The straight bar then would be distorted first in one direction and then in the direction of the dotted figure. In the case of the second mode of vibration, it will be noticed that it consists essentially of two of the fundamental mode types joined end to end. This is not strictly the case, but serves to illustrate the motion. The dots shown at various points on the bar show positions of zero motion or nodes. In the case of the fundamental mode, there are two nodes and in the second and third there are three and four respectively. One point of

interest in flexure vibration as seen in Fig. 6.1 is that the ends of the bar will be vibrating in the same direction for odd order modes and the motion of the two ends will be in opposing directions for even order modes. The frequency of a bar vibrating in flexure may be easily computed for low orders when the width is small in comparison with the length. When the width is appreciable other factors must be considered as will be shown later. In general, the flexure frequency of a bar will be the lowest frequency of vibration.

In the case of a plate where we are concerned with flexural vibrations propagated along the length with motion in the direction of the thickness it

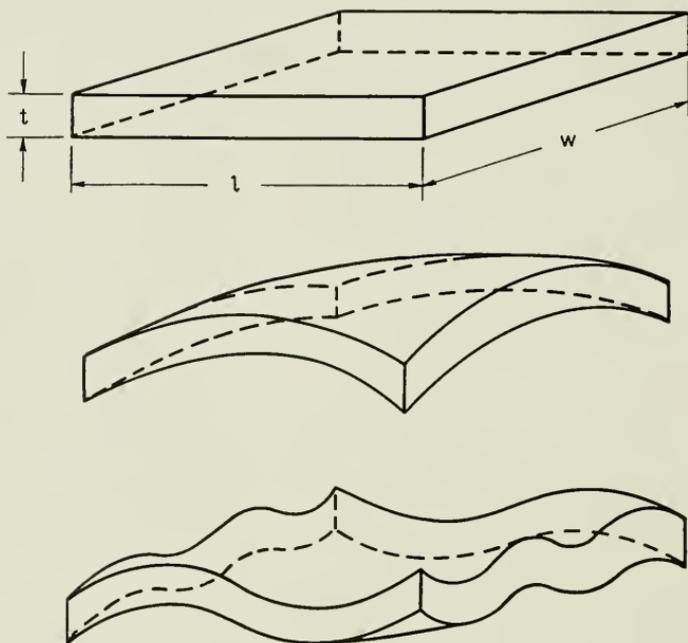


Fig. 6.2—Motion of a plate in free-free flexure.

is necessary to consider also the width. As noted in Fig. 6.1, our concern was only for a bar of small third dimension. When considering the case of a plate in flexure along its length and thickness, then the third dimension must also be considered for more complicated types of motion. In a manner somewhat similar to the vibration of a bar, we can consider a plate vibrating in its thickness-length plane. Since a plate also has width, we must also consider this dimension. The simplest type of motion would be that of a simple flexure which would bend the plate into the shape of an arch. If now, the third dimension is permitted to flex, the distortion of a plate shown in Fig. 6.2 could be illustrated by a flexure in the l - t plane and in the

$w-t$ plane. Considering the motion of the plate as a flexure vibration along the length vibrating in the thickness, then we may also have a distortion along the width and thickness corresponding to similar or higher types of flexure motion. The illustration at the bottom of the figure shows a plate vibrating in its second order flexure along the length and thickness and the fourth order flexure along the width and thickness. The effect of these higher orders in the $w-t$ plane is to slightly modify the frequency of the $l-w$ mode.

A thorough treatment of this type of double flexure in plates will be given in Chapter VIII by H. J. McSkimin.

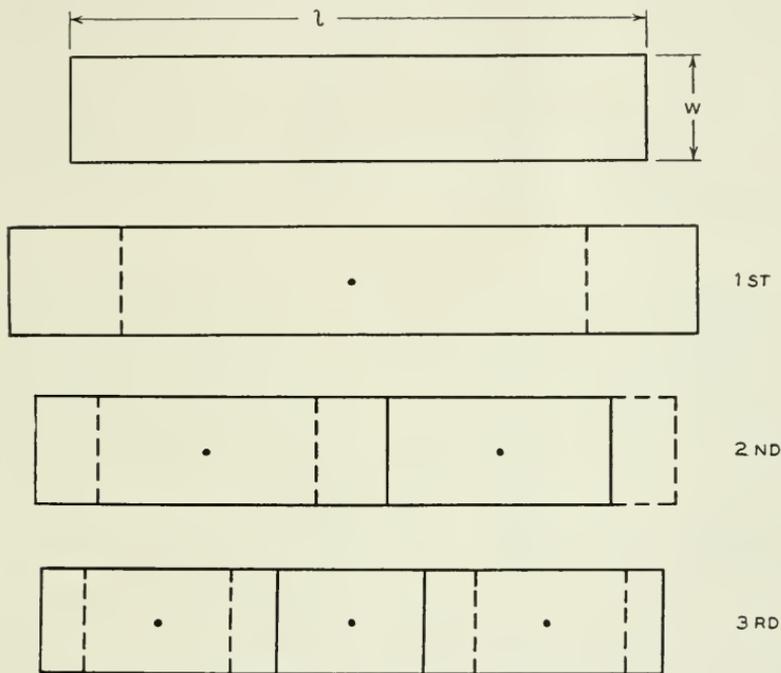


Fig. 6.3—Motion of a bar in free-free extension.

6.22 Extensional

The extensional or sometimes termed longitudinal motion of a bar free to vibrate is shown on Fig. 6.3. This motion is somewhat simpler than the flexure motion and consists simply of a displacement in the direction of the length of the bar of a wave propagated along the length. This means that the first mode of vibration will be simply an expansion and contraction of all points with respect to the center of the bar. This motion will be along the length. The displacements along the bar will then be in proportion to the sine of the angular distance from the center. The distortion of a free bar in its simplest mode is then illustrated in Fig. 6.3 labeled 1st. Since the

motion must be dynamically balanced, a node will appear at the center of the bar, and the bar will grow longer and shorter as shown by the solid and dotted lines. In the case of the second order of motion, as shown in Fig. 6.3, it consists essentially of two 1st order modes joined together at their ends and of opposite phase. That is to say, when one half of the bar is expanding, the other half is contracting. In the case of the 3rd mode, as can be seen from Fig. 6.3, the central element is contracting while the external elements are expanding. From this we may state generally, that for odd order types of motion, the extreme ends of the bar will be expanding or contracting in phase and for even order modes, the extreme ends will be expanding or contracting in opposite phase. Fig. 6.3 illustrates extensional motion in its simplest form. In a practical case an extension in one direction is accompanied by a contraction in one or both of the other two dimensions. This of course is due to elastic coupling and will be considered more in detail later. If we consider a rectangular plate it is not difficult to imagine that it would have three series of extensional modes of vibration due to the three principal dimensions.

6.23 *Shear*

The low frequency of face shear type of motion of a plate is somewhat more complicated than either the flexure or longitudinal and, as shown in Fig. 6.4, consists simply of an expansion and compression in opposite phase along the two diagonals of the plate. This motion is shown in Fig. 6.4 labeled $m = 1, n = 1$. The two phases are shown, one a solid curve and the other a dotted curve to illustrate the distortion with respect to the original plate. One peculiarity of shear motion in plates is that it may break up into motions similar to its fundamental along either the length or the width. For example, if we take the motion associated with $m = 1, n = 1$, and superimpose two of these in opposite phase on the same plate, we would get the type of motion illustrated by $m = 2, n = 1$. In a similar manner, the motion may reverse its phase any number of times along either the length or the width. One particular case is shown for $m = 6, n = 3$. As can be seen from the case of $m = 1, n = 1$, the distortion is not that of a parallelogram as it is in the static case because here we are concerned only with the dynamic case. While the equation of motion of a free plate vibrating in shear has not been completely solved, a microscopic analysis indicates that the actual motion of the plate edges appear to be somewhat as shown for the case $m = 1, n = 1$ when driven in this mode.

The shear mode of motion in the case of a thin plate is somewhat different for the high frequency case than for the low frequency case. In the case of high frequency shear modes of motion in thin plates, the motion of a particle is at right angles to the direction of propagation which in this case would be

the thickness. The simplest type of motion for high frequency shear is shown in Fig. 6.5 where the top of the plate is displaced in the direction along ℓ with respect to the bottom of the plate. This would then be termed the length-thickness shear. When viewed from the edge of the plate, the motion is very similar to that shown in Fig. 6.4 for the case of $m = 1, n = 1$. In a manner similar to the previous case of shear the front edge of the plate may be divided into segments along ℓ and along t . For example, we may get

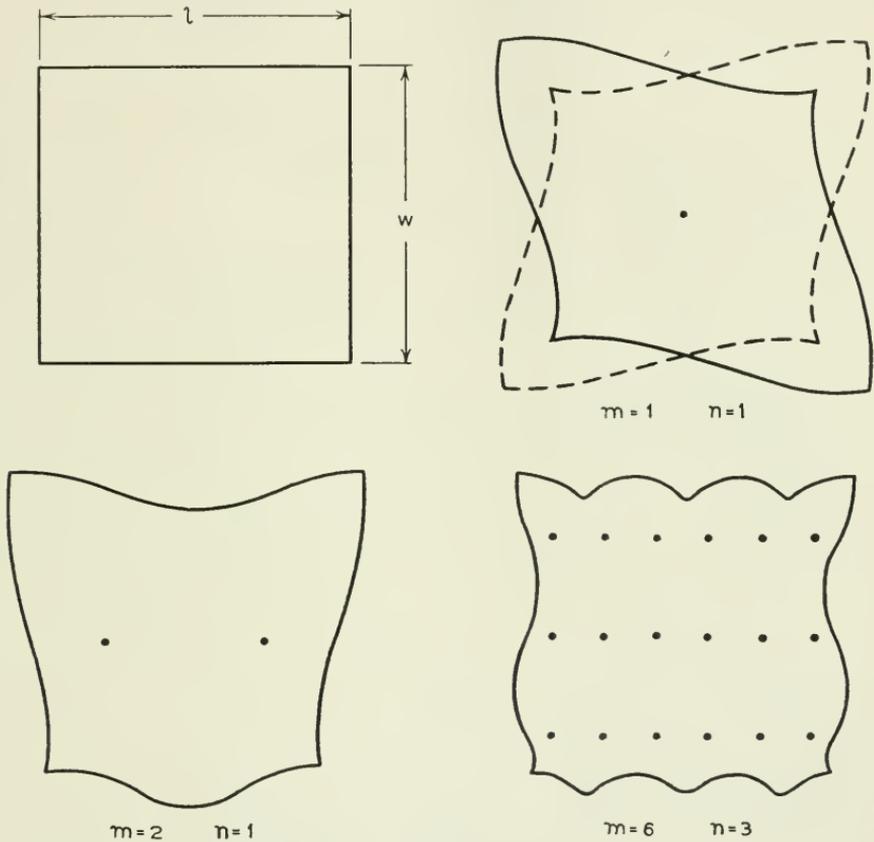


Fig. 6.4—Motion of a plate in low frequency shear.

a double shear along ℓ with a single shear along t . This case is illustrated in Fig. 6.5 for $m = 1, n = 2$ and $p = 1$. In general, m and n may assume any integral value. As in the case of flexure we must also consider the third dimension. The motion associated with the third dimension may be represented by simple reversals of phase as before. For example, in Fig. 6.5 the case for $m = 1, n = 1, p = 2$ is shown which simply means that the high frequency shear on the front half of the plate is out of phase with that of the

back half of the plate. This discussion relates only to the case of the high frequency shear commonly assumed to be a single shear along the length and thickness of the plate. Similar statements can be made if we consider the high frequency shear as being along the width and thickness.

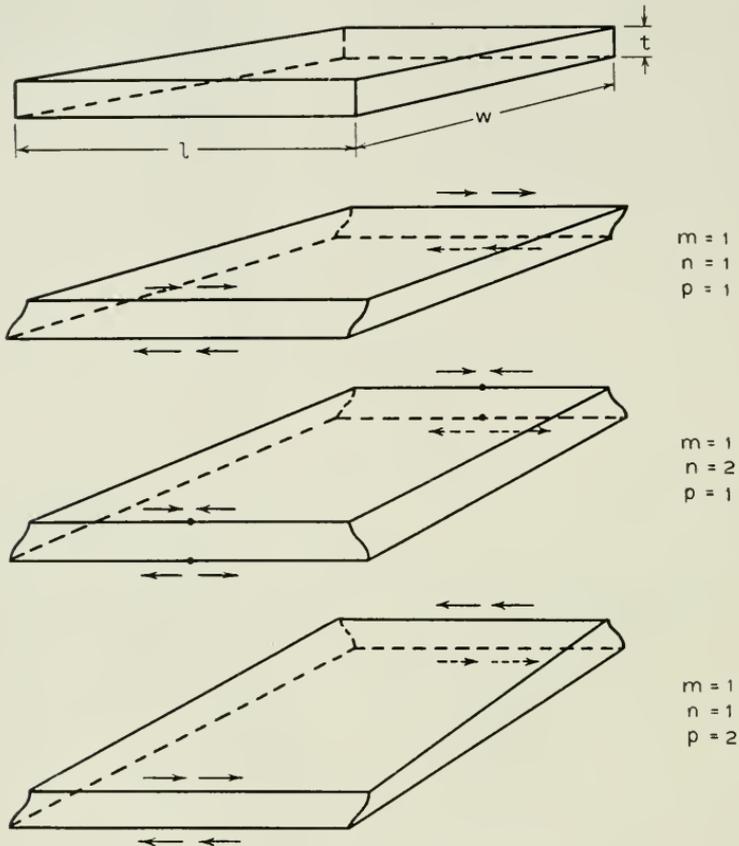


Fig. 6.5—Motion of a plate in high frequency shear.

6.24 Type of Motion for Some Standard Filter and Oscillator Plates

To get a more complete picture of the applications of the various types of motion, we will now take specific cases. The various crystals as commonly used for oscillators or filters are shown in Fig. 6.6. At the top of Fig. 6.6 are shown the various types of shear plates with their relative position with respect to the crystallographic axis.

The *AT* and *BT* plates are termed high frequency shear plates and the motion associated with them is that of a length-thickness shear as shown in Fig. 6.5. Their use is found for the control of radio frequency oscillators in

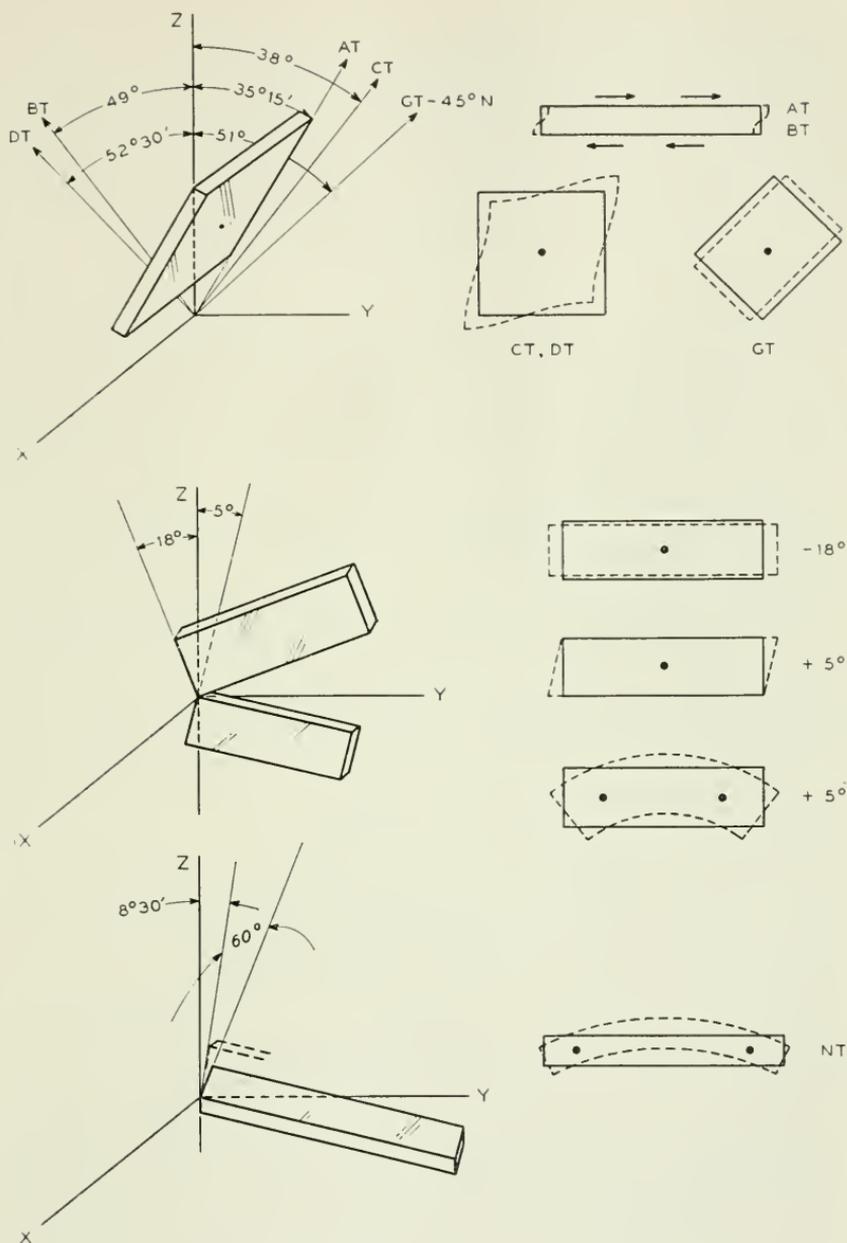


Fig. 6.6—Motions of typical cuts of quartz.

the range from 1 to 10 megacycles. The *AT* is most useful in the lower range and the *BT* in the upper range since it has a higher frequency constant.

Considerable use for the *AT* plate has been found for filters on pilot channels for the coaxial telephone system.

The *CT* and *DT* are analogous to the *AT* and *BT* but are termed low frequency shear plates. The motion associated with these cuts is that of a face shear as illustrated in Fig. 6.4. The *CT* and *DT* cuts are useful for both filter and oscillator applications in the frequency range from 60 kilocycles to 1000 kilocycles. Here again the *DT* would be most useful in the lower range and the *CT* the upper range due to the higher frequency constant for the *CT* cut.

The *GT* is similar to the *CT* except that it is rotated by 45° about the normal to the plate so that instead of a face shear type of motion there are two extensional modes similar to that shown in Fig. 6.3. These two modes are coupled to each other resulting in one of them having a zero temperature coefficient over a wide range of temperature. This crystal is most useful in the range from 100 kilocycles to 500 kilocycles for a primary standard of frequency and in filter networks having extreme phase requirements.

The filter plates commonly called the -18° cut and 5° cut are shown with their relation to the crystallographic axes in the central part of Fig. 6.6. The -18° cut commonly used in filters employs a simple extensional motion along its length with small coupling to an extensional motion along its width and practically zero coupling to a face shear type of motion. Since the width is usually the order of half the length these modes are not troublesome. The $+5^\circ$ cut is useful in filter work because it has a low temperature coefficient and in spite of its strong coupling to the plate shear, it has been found quite useful in both its extensional mode and its flexure mode. The -18° cut is used over the frequency range from 60 kilocycles to 300 kilocycles and forms the basic crystal used in the channel filters of the coaxial telephone system. When driven in flexure the 5° cut may be made to operate as low as 5 kilocycles and is used in oscillator and filter circuits.

The *NT* cut is shown at the bottom of Fig. 6.6 with its relation to the crystallographic axis. This is obtained by a rotation of $+8.5^\circ$ about the *X* axis with a second rotation of $\pm 60^\circ$ about the resulting *I'* axis. The purpose of the second rotation is to give the shear modulus a positive coefficient. This modulus enters into the equation for the flexure frequency and therefore the effect of the second rotation is to change the temperature coefficient of the flexure mode from a negative value to zero. This crystal has been used to some extent as a low frequency oscillator. Its main purpose so far has been for the control of frequency modulation broadcast transmitters and for low frequency pilot channel filters.

Another crystal called the *MT* which is cut in a manner similar to the *NT* but with angles of 8.5° and 36° respectively has been used for filter work where an extensionally vibrating crystal of zero temperature coefficient is

required. The motion associated with this crystal is similar to that shown for the $+5^\circ$ cut of Fig. 6.6. The low temperature coefficient is obtained through coupling to, and the effects of, a shear mode of positive temperature coefficient. Its use has been mainly for pilot channel filters of rather narrow frequency bands.

6.3 FREQUENCY EQUATIONS FOR FLEXUREL, EXTENSIONAL AND SHEAR MOTIONS

In determining the motion and resonant frequencies of a particular type of vibrating system it is customary to consider an isolated type of motion in order that the solution shall be in a simple enough form to be practical even though it may not be too accurate. The more accurate type of solution is often so complex that its use for practical solutions might be small. Since any solutions so far obtained are not complete in every detail, it is usually necessary to resort to experimentally determined frequencies in any case, and the solution can only be regarded as a guide to the complete result. In the following treatment it will be assumed that the frequency equations are given for isolated modes of motion and it will be later shown which of these forms are coupled and the effect of the coupling.

6.31 FLEXURAL RESONANT FREQUENCIES

The simplest equation relating the resonant frequencies of a rod vibrating in flexure is given by¹

$$f = \frac{m^2}{2\pi} \frac{k}{\ell^2} \nu \tag{6.1}$$

where ν = velocity of extensional propagation = $\sqrt{Y_0/\rho}$

k = radius of gyration of cross section

Y_0 = Young's modulus

ℓ = length

$m \doteq (n + 1/2)\pi$ for free-free modes

$\doteq (n - 1/2)\pi$ for clamp-free modes ($n > 1$)

n = order of mode (1, 2, 3, etc.)

This equation holds only for the case of a long thin rod. Measurements of the resonant frequencies of a quartz crystal vibrating with both ends free has shown the above equation to be true where m is defined approximately as $(n + 1/2)\pi$ provided $\frac{m\omega}{\ell}$ is less than .1. For values greater than this the measured values are somewhat lower than that predicted. When the dimension in the direction of vibration is appreciable in comparison with the

¹ Rayleigh, Theory of Sound, Vol. 1, Chapter VIII.

length, Mason² has shown that it is necessary to consider the effects of rotary and lateral inertia. His solution leads to the same frequency equation as 6.1 but with a different evaluation of the factor m which is obtained from the transcendental equations

$$\left. \begin{aligned} \tan m X &= K \tanh mX' \text{ for even modes} \\ \tan m X &= -\frac{1}{K} \tanh mX' \text{ for odd modes} \end{aligned} \right\} \quad 6.2$$

where

$$\begin{aligned} X &= 1/2 \left[\left(1 + \frac{m^4 k^4}{4\ell^4} \right)^{1/2} + \frac{m^2 k^2}{2\ell^2} \right]^{1/2} \\ X' &= 1/2 \left[\left(1 + \frac{m^4 k^4}{4\ell^4} \right)^{1/2} - \frac{m^2 k^2}{2\ell^2} \right]^{1/2} \\ K &= \left[\left(1 + \frac{m^4 k^4}{4\ell^4} \right)^{1/2} + \frac{m^2 k^2}{2\ell^2} \right]^2 \frac{X'}{X} \end{aligned}$$

Equation 6.2 holds only for the case of a rod free to vibrate on both ends. The case of a clamp-free rod is somewhat more complicated since it cannot be given by separate solutions for the even and odd modes. The interpretation of m given in equations 6.2 will result in the same value as before [$m = (n + \frac{1}{2})\pi$] for values of $\frac{m\omega}{\ell}$ less than .05 but decrease considerably for larger values and ultimately as the bar becomes wider the effects of rotary inertia result in the flexure frequency approaching the extensional mode as an asymptote. As stated before measurements on quartz bars vibrating in flexure departed from that predicted by the simple definition of m when the width of the bar was such that $\frac{m\omega}{\ell} > .1$. By using the value of m defined by

equation 6.2 it is possible to predict the frequency for widths as great as $\frac{m\omega}{\ell} = .5$. For widths greater than this, experiment shows a frequency lower than that predicted by equation 6.2. This then leads one to believe that the effect of shear plays an important part in the flexure of bars with appreciable width. An investigation of the effect of shear on the flexure frequencies of beams has been made by Jacobsen³ and his results lead to the same frequency equation as 6.1 and to the same transcendental equations derived by Mason (6.2) but with different values of X , X' and K to account for the shearing

² W. P. Mason, "Electromechanical Transducers and Wave Filters," Appendix A. D. Van Nostrand Company, Inc.

³ *Jour. Applied Mechanics*, March 1938.

effect. These values are given by

$$\begin{aligned}
 X &= \frac{1}{2} \left[\left(1 + \frac{m^4 k^4}{4\ell^4} \left(\frac{1}{c_{jj} s_{ii}} - 1 \right) \right)^{\frac{1}{2}} + \frac{m^2 k^2}{2\ell^2} \left(\frac{1}{c_{jj} s_{ii}} + 1 \right) \right]^{\frac{1}{2}} \\
 X' &= \frac{1}{2} \left[\left(1 + \frac{m^4 k^4}{4\ell^4} \left(\frac{1}{c_{jj} s_{ii}} - 1 \right) \right)^{\frac{1}{2}} - \frac{m^2 k^2}{2\ell^2} \left(\frac{1}{c_{jj} s_{ij}} + 1 \right) \right]^{\frac{1}{2}} \\
 K &= \left[\left(1 + \frac{m^4 k^4}{4\ell^4} \left(\frac{1}{c_{jj} s_{ii}} - 1 \right) \right)^{\frac{1}{2}} + \frac{m^2 k^2}{2\ell^2} \left(\frac{1}{c_{jj} s_{ii}} - 1 \right) \right]^2 \frac{X'}{X},
 \end{aligned} \tag{6.3}$$

where c_{jj} is the shear constant in the plane of motion s_{ii} is the elastic constant in the direction of propagation. While it is true that these values will result in a lower value of m than those associated with equation 6.2 and hence fit the actual measured results more closely for bars wider than $\frac{n\omega}{\ell} = .5$, there is some doubt in the minds of various investigators as to the actual amount of correction necessary to apply to compensate for the shear. The solution of equation 6.2 using the constants of equation 6.3 is a lengthy process and could only be applied to a given orientation since the elastic constants vary with direction in quartz. While the results of Jacobsen's work are difficult to handle for intermediate values of $\frac{n\omega}{\ell}$ where the correction of rotary and lateral inertia do not fit the measured results it does imply that for large values of $\frac{n\omega}{\ell}$ that the flexure frequencies will be mainly a function of the length alone. Therefore when we are concerned with very high orders of flexure in plates such as the case of high frequency *AT* and *BT* shear crystals we may assume the interfering modes due to flexures will be essentially harmonic in nature. Restating the general problem of determining flexure frequencies in quartz rods or plates we may assume that the ratio of width to length is the controlling factor in deciding which method of attack is to be employed. For values of $\frac{n\omega}{\ell}$ less than .1 equation 6.1 will give quite accurate results. For values of $\frac{n\omega}{\ell}$ up to .5 equation 6.1, using the values of m determined by equation 6.2 will give satisfactory results. While the values of m determined by using equation 6.3 will give more accurate results for the range .4 to .6, it is not desirable to carry it further because, while 6.2 does take into consideration the effect of shear it does not account for coupling to the shear mode of motion. Hence for values of $\frac{n\omega}{\ell} > .6$

it is best to depend upon experimental measurements if accurate results are a factor.

6.32 Extensional Frequencies

The resonant frequencies of a bar vibrating along its length, commonly called an extensional mode of motion is derived quite easily from the wave equation in one dimension and is given by

$$f = \frac{n}{2\ell_i} \sqrt{\frac{1}{s_{ii}\rho}} \quad 6.4$$

where ℓ = length

s_{ii} = elastic constant in the direction of propagation

ρ = density

$n = 1, 2, 3, 4, \text{etc.}$

This is the case when the length is the greatest dimension. When we consider extensional modes along the thickness of a plate, it can be shown that the c constants be employed to account for the lateral inertia in the two directions at right angles to the direction of propagation, (provided that the resulting motion is nearly along the thickness direction). Hence, for thin plates

$$f = \frac{n}{2t_i} \sqrt{\frac{c_{ii}}{\rho}} \quad 6.5$$

As an example of the use of the above equation an X -cut bar vibrating along its length would result in a series of resonant frequencies defined by equation 6.4. An X -cut plate vibrating along its thickness would result in a series of frequencies defined by equation 6.5. Applying the appropriate constants

$$\begin{aligned} f_\ell &= \frac{n}{2Y} \sqrt{\frac{10^{14}}{127.9 \times 2.65}} \\ &= \frac{272}{Y(\text{cm})} n \text{ kilocycles} \end{aligned} \quad 6.6$$

and

$$\begin{aligned} f_t &= \frac{n}{2X} \sqrt{\frac{86.05 \times 10^{10}}{2.65}} \\ &= \frac{285}{X(\text{cm})} n \text{ kilocycles} \end{aligned} \quad 6.7$$

This shows that although Young's Modulus is the same in the two directions the resulting frequency constants are different because of the conditions at the boundaries.

6.33 Shear Resonant Frequencies

As shown in section 6.23 the low frequency face type shear mode results in a doubly infinite series of frequencies due to the manner in which the plate may break up into reversals of phase along its length and width. While a solution for the low frequency shear motion that satisfies the boundary condition of a free edge has not yet been accomplished, several approximate solutions for the frequencies are available. A modification of the equation developed by Mason⁴ will give results which verify experimental data.

$$f = \frac{1}{2} \sqrt{\frac{1}{\rho s_{jj}}} \sqrt{\frac{m^2}{\ell^2} + k^2 \frac{n^2}{w^2}} \quad 6.8$$

- where ρ = density
- s_{jj} = shear modulus in ℓw plane
- $m, n = 1, 2, 3, \text{ etc.}$
- ℓ = length of plate
- w = width of plate

The value of k so far remains experimental and for low orders of m and n may be assumed unity. Its use is mainly for high orders of m and n where Young's modulus is different in the ℓ and w directions. Experimental data in the case of *BT* plates indicates that it should be 1.036 to account for the difference in velocity in the two directions. When m or n is large the velocity component, namely $\sqrt{\frac{1}{\rho s_{jj}}}$ should be replaced by $\sqrt{\frac{c_{jj}}{\rho}}$ for reasons explained for the extensional case. Equation 6.8 holds for the case of a plate vibrating in low frequency shear in regions where no highly coupled extensional or flexural resonant frequencies exist. As will be shown later, these regions are few. By assuming the frequencies are given by these equations and then applying the normal correction for coupled modes, a fairly accurate result will be obtained.

The high frequency case of a plate vibrating in shear is somewhat similar to the face shear or low frequency case with the exception that three dimensions must be considered since two are large compared to the third (the main frequency controlling dimension). An experimental formula for this case is given by

$$f = \frac{1}{2} \sqrt{\frac{c_{jj}}{\rho}} \sqrt{\frac{m^2}{\ell^2} + k \frac{n^2}{\ell^2} + k_1 \frac{(p-1)^2}{w^2}} \quad 6.9$$

- where c_{jj} = shear modulus in plane of motion
- ρ = density
- ℓ, w, t = length, width and thickness

⁴ "Electrical Wave Filters Employing Quartz Crystals as Elements," W. P. Mason, *B.S.T.J.* July, 1934.

m , n and p represent reversals of phase along the three directions and may be termed overtones. The values of k and k_1 are inserted to correct for the change in shear velocity resulting from a change in Young's modulus in the three directions. For most work with oscillator crystals where the length and width are large compared to the thickness, the following simplification of equation 6.9 is most useful.

$$f = \frac{m}{2l} \sqrt{\frac{c_{ij}}{\rho}} \quad 6.10$$

When high frequency shear type crystals are used in connection with selective networks, it is necessary to make use of equation 6.9 to determine where the next possible pass regions will occur.

6.34 *Effects of Rotation About the Crystallographic Axes on the Resonant Frequencies and Coupling between Modes of Motion*

Several of the elastic constants have been used in equations expressing the resonant frequencies. Since most of the crystal cuts now in use are rotated at some particular angle about the X crystallographic axis, it is of interest to know the effect of this rotation upon the elastic constants since they determine the resonant frequencies and the coupling between certain of the modes of motion. The general stress-strain equations for an aeotropic body are given in equation A.1 of Appendix A together with their definitions. In the case of quartz where the axes of the finished plate are aligned with the crystallographic axes the constants reduce to 7 and are shown in equation A.8. Examination of these equations shows that there are extensional and shearing strains resulting from dissimilar extensional and shearing stresses through the elastic constants s_{ij} and c_{ij} . This results in coupling between modes of motion where a so-called cross strain exists. These couplings may be made zero or small by proper orientation of the crystal plate about the X crystallographic axis. The mathematics of this operation is simplified by the use of matrix algebra⁵. Upon performing this operation a new set of elastic constants are obtained and are plotted graphically together with the piezoelectric constants on Fig. 6.7. From this figure we may see that the coupling resulting from the s'_{24} constant will be zero if the crystal plate is orientated by -18.5° about X with respect to the crystallographic axis. This constant determines the coupling between the extensional mode along the length (Y' dimension) and the face shear mode ($Y'X'$ dimensions). This analysis resulted in the use of the -18.5° cut in the channel filters of the coaxial system. Two other crystal cuts resulting in low coupling between different modes of motion are the AC and

⁵ "The Mathematics of the Physical Properties of Crystals," W. L. Bond, *B.S.T.J.*, Jan. 1943.

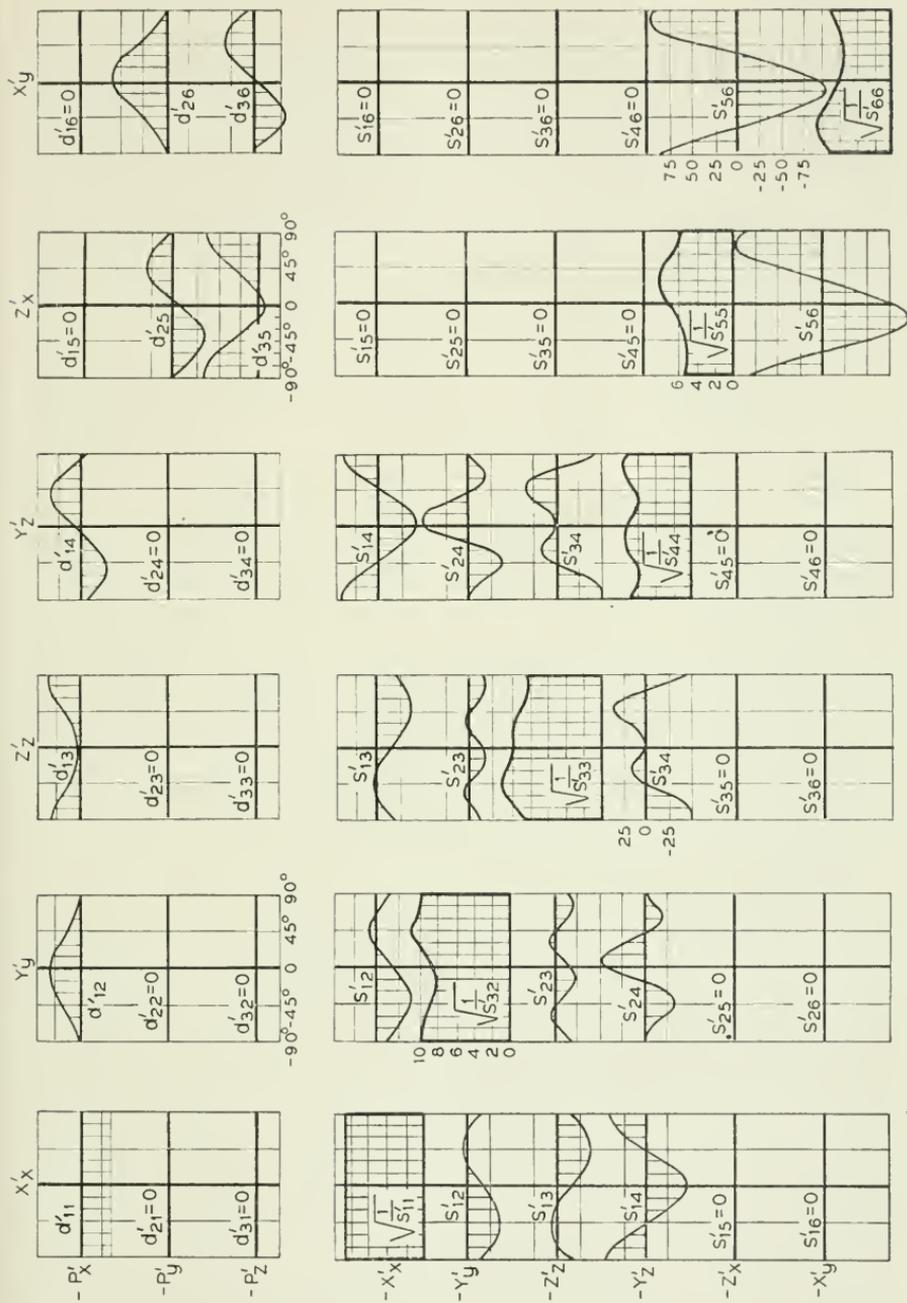


Fig. 6.7—Variation of piezoelectric and elastic constants of quartz with rotation about the X' crystallographic axis.

BC cuts. The s'_{56} constant determines the coupling between the face and thickness shear modes. As shown in Fig. 6.7 this constant passes through zero at two values, namely $+31^\circ$ and -59° and the resulting angles have been termed the *AC* and *BC* cuts. These angles are very close to the *AT* and *BT* cuts and hence they also possess the benefits of low coupling between modes. In addition to making the cross coupling constants zero, a rotation of the crystal plate with respect to the crystallographic axes also results in a change in the extensional and shear elastic constants. Notice that these pass through maxima and minima at the zero values for the cross coupling constants. This of course affects the resonant frequencies of isolated modes. Changes as great as 50% increase in frequency constants may be obtained by choosing the proper rotations. The equations relating the elastic constants as functions of orientation are given in appendix B for more complete use.

6.4 COUPLING BETWEEN MODES OF MOTION

As pointed out in the previous section, the frequency equation of a given mode of motion will give accurate results only in the case where the mode of motion is isolated. This is very rarely the case since most quartz crystals in common use are in the form of plates where the frequency determining dimension is not large in comparison with all other dimensions. Only in the case of a long thin rod vibrating in length-thickness flexure of the first order would this be true. It was also shown that the coupling between different modes of motion could be related to the mutual elastic constants (s_{ij} and c_{ij}) and that some of these could be made zero by the proper choice of orientation of the finished crystal plate. The elastic constants s_{ij} and c_{ij} only relate to the coupling between the extensionals, the shears and the extensional to the shear. For example s_{23} relates to the coupling between the extensional modes along the *Y* and *Z* axes, s_{56} relates to the coupling between the low and high frequency shear modes of a *Y* cut plate and s_{24} relates to the coupling between an extensional mode along the *Y* axis and a shear mode in the *YZ* plane. One other important coupling condition occurs and that is between the flexure and the shear modes. There is at present no mathematical theory relating this form of coupling except from simple assumptions that may be drawn from the fact that the shear modulus enters as a controlling factor in determining the frequency of a bar vibrating in flexure and from the similarity of the two types of motion near the boundaries. Since it is possible to have a definite coupling between extensional and shear modes there must be coupling between the extensional and flexure modes. It would be expected that it would be proportional to the coupling between the extensional and shear modes.

6.41 Extensional to Shear and Extensional to Flexure Coupling

The coupling between the extensional and shear motion can best be illustrated by taking the case of an *X* cut plate the length of which lies along the *Y* axis and the width along the *Z* axis. This is shown in Fig. 6.8 together with two other cases, one in which the plate is rotated about the *X* axis by -18° and the other a similar rotation but $+18^\circ$. Also in Fig. 6.8 is shown an enlarged view of the change in the elastic constants and frequency constants as a function of the rotation of the plate about the electric or *X* axis. For the case of an *X* cut plate the strains resulting from an applied exten-

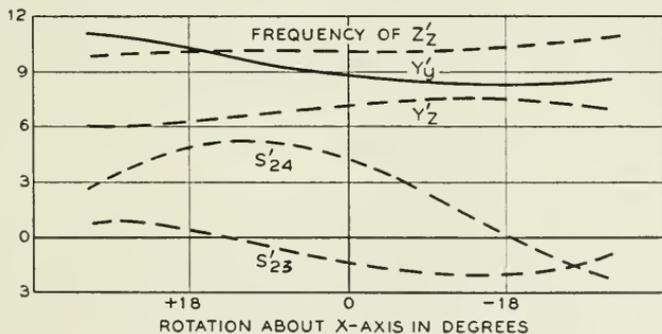
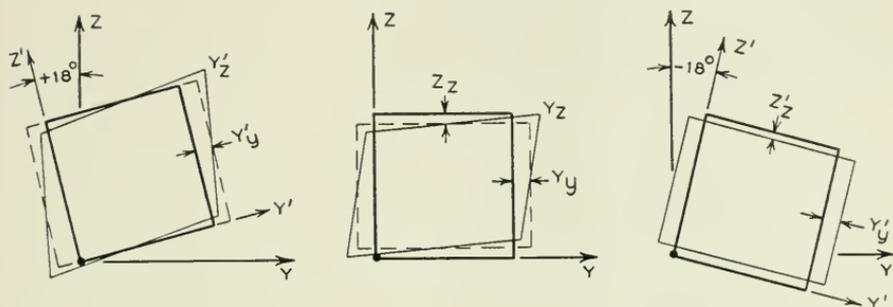


Fig. 6.8—Motion in an *X* cut plate for different orientation about the *X* crystallographic axis.

sional stress along the length according to equation A.8 would be

$$\begin{aligned} x_z &= s'_{12} Y_v \\ y_v &= s'_{22} Y_v \\ z_z &= s'_{23} Y_v \\ y_z &= s'_{24} Y_v \end{aligned}$$

6.11

where x_z is an extensional strain along the thickness
 y_v " " " " " " length
 z_z " " " " " " width
 y_z " a shear strain in the length-width plane

If the plate is thin we may neglect the x_x strain as far as its effect on the resonant frequencies associated with the length and width are concerned. From the plot of the elastic constants on Fig. 6.8 we may determine the strains resulting from a stress along the length of an X cut plate for various orientations about the X axis. In addition to the expected extension along the length we have for a $+18^\circ$ cut, a large amount of length-width or y'_z shear strain due to s'_{24} and very little width or z'_z strain. For the 0° cut there is also large length-width or y'_z shear strain and a width or z'_z strain. In the case of the -18° cut the shear strain vanishes due to s'_{24} being zero, leaving in addition to the expected length or y'_y strain a width or z'_z strain. These relationships are more clearly shown if we plot the resonant frequencies resulting from the three modes of motion namely, the extensional modes along the length and width and the shear mode in the length-width plane

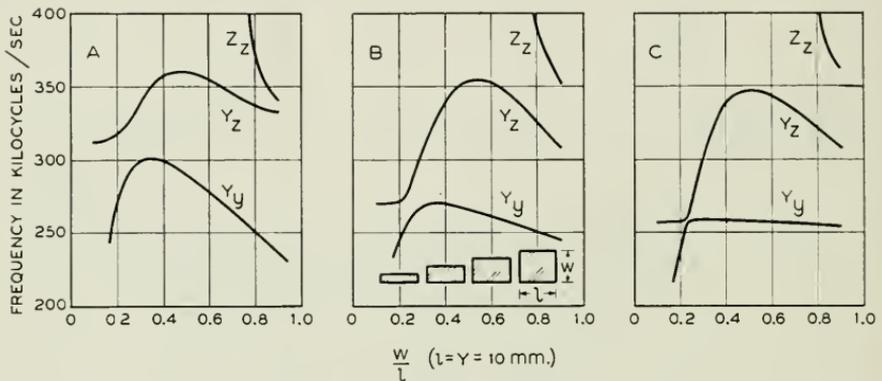


Fig. 6.9—Effect of rotation about the X axis on the resonant frequencies of an X cut plate.

A plot of measured resonances is shown in Fig. 6.9 for the above described three cases as a function of the change in width. The resonant frequencies for these three types of motion are given in section 6.3 as

$$f_{y'_y} = \frac{1}{2\ell} \sqrt{\frac{1}{\rho s'_{22}}}, \text{ extensional along } \ell \quad 6.12$$

$$f_{z'_z} = \frac{1}{2w} \sqrt{\frac{1}{\rho s'_{33}}}, \text{ extensional along } w \quad 6.13$$

$$f_{y'_z} = \frac{1}{2} \sqrt{\frac{1}{\rho s'_{44}}} \sqrt{\frac{1}{\ell^2} + \frac{1}{w^2}}, \text{ shear in } \ell w \text{ plane} \quad 6.14$$

These equations specify only the uncoupled modes and do not take into consideration the effect of coupling to other modes of motion. In the case of Fig. 6.9 it is shown that when only the width is changed the extensional

mode along the length (the y'_y mode) is unaffected only in the case of the -18° cut. The effect of coupling between the extensional and shear is clearly shown in the case of the 0° cut by the change in the length-extensional frequency. This is more pronounced in the $+18^\circ$ case not because of more coupling but because the frequency constants of the two modes are more nearly alike as indicated in Fig. 6.8.

The mode of motion associated with the line intersecting the extensional y'_y mode is that due to the second length-width flexure mode. As mentioned before it is strongly coupled to the shear mode in the same plane. The coupling between this flexure and the extensional mode is directly related to the coupling between the shear and the extensional mode. This is borne out by Fig. 6.9, for in the case of the -18° cut, s'_{24} is zero and as can be seen the change in frequency of the extensional mode is very slight even when the flexure mode is nearly identical in frequency.

We may state generally that the change in frequency of a particular mode of motion from that of its uncoupled state is dependant on two factors; the coupling to and the proximity to other forms of motion. This follows well established mathematical procedures but to solve the case just discussed would require the solution of a four mesh network with mutual impedances the values of some of which are at best only approximate. This will serve to illustrate that the use of formulae such as given in section 6.3 may be used more as a guide in establishing certain modes of motion rather than for accurate determinations of resonant frequencies.

6.42 Flexure to Shear Coupling

1. Low Frequency Shear

As previously indicated there is no simple means of mathematically determining the coupling between flexure and shear types of motion as there is between the extensional and extensional to shear modes. Here we must base our assumptions upon observed experimental evidence and simple reasoning. The relation between flexure motion and shear motion can be illustrated by the figures associated with Fig. 6.10. The forces that are necessary to produce flexure and shear motion are shown by arrows in Fig. 6.10. When the two arrows point toward each other, it indicates a compression and when the arrows point away from each other, it indicates tension. The diagrams on the left of Fig. 6.10 illustrate the conditions for flexure motion and the diagrams on the right indicate the conditions for shear motion. Notice that in the case of the first flexure and the second shear that the forces applied to the top and bottom of the plate are similar. Also in the case of the second flexure and third shear, they are similar. Here again we have certain similarities which in this case are important to remember.

The motion of the ends of the plate in the case of the first flexure are similar to those of the second shear. In the case of the second flexure the similarity is observed in the case of the third shear. The end motion in the case of the third shear is also the same in the case of the first or any odd shear. Likewise, the end motion of the first flexure is similar to the second shear or any even shear. We may then generalize and say that it is very likely that an odd order flexure would be coupled to an even shear; and also an even flexure would be coupled to an odd shear.

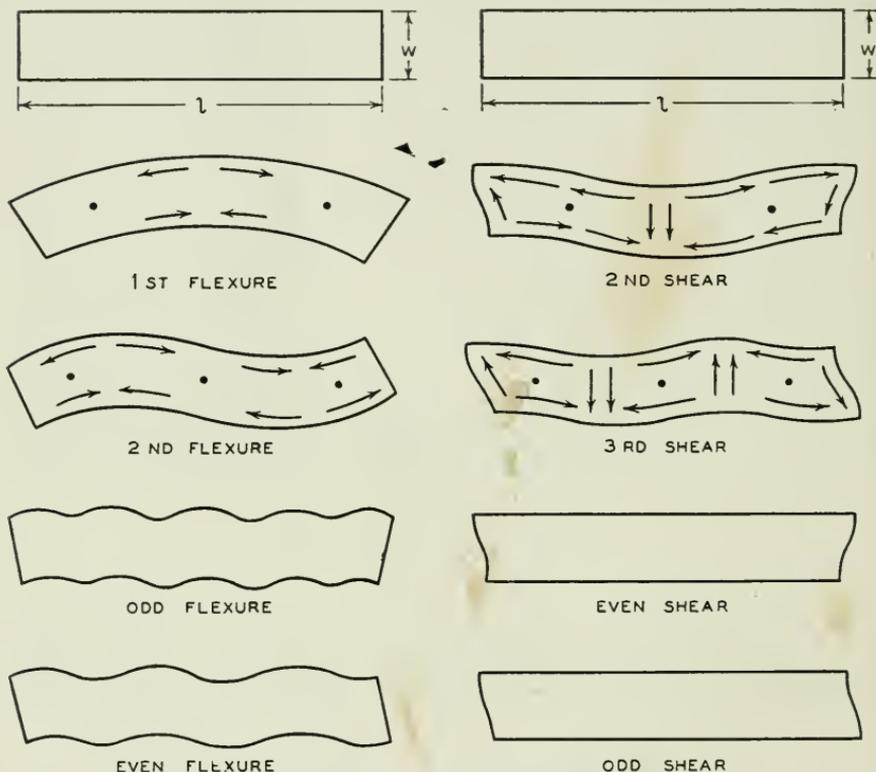


Fig. 6.10—Similarities in shear and flexure motions in a bar.

To illustrate the coupling between flexure and shear type motions, the frequencies of flexure and shear modes in a *Z*-cut quartz plate as shown in Fig. 6.11 have been measured. These measured frequencies are shown by the solid lines for various widths of the plate. It will be seen that there are no observed resonances following an unbroken continuous line to represent the shear frequency, but they are interrupted by several other frequencies which we must interpret as being various even modes of the flexure in the plane of the plate. It is clearly shown here that only even order flexures are

strongly coupled to the fundamental or odd shear. The strong coupling shown between the X_y shear and the second X_y flexure explains why the frequency equations given in section 6.3 for the frequency of flexure and

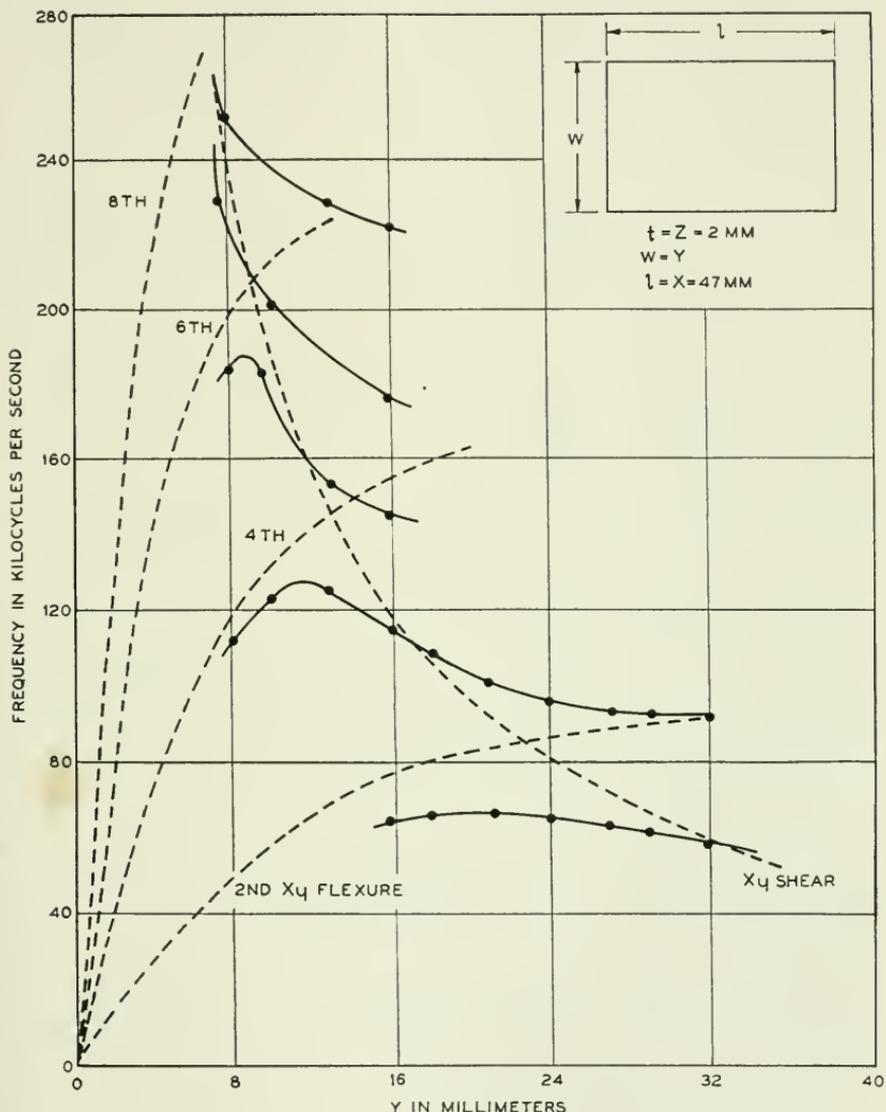


Fig. 6.11—Shear and flexure resonances in a Z-cut quartz plate.

shear modes will not give even approximate results if applied to this case for a square crystal. It will be shown later that if account is taken of coupling, the shear mode for a square crystal of this type may be more accu-

rately determined. Fig. 6.12 is a more detailed representation of the conditions shown broadly in Fig. 6.11 except in this case an *AC*-cut quartz plate was used and most of the observable resonant frequencies are shown

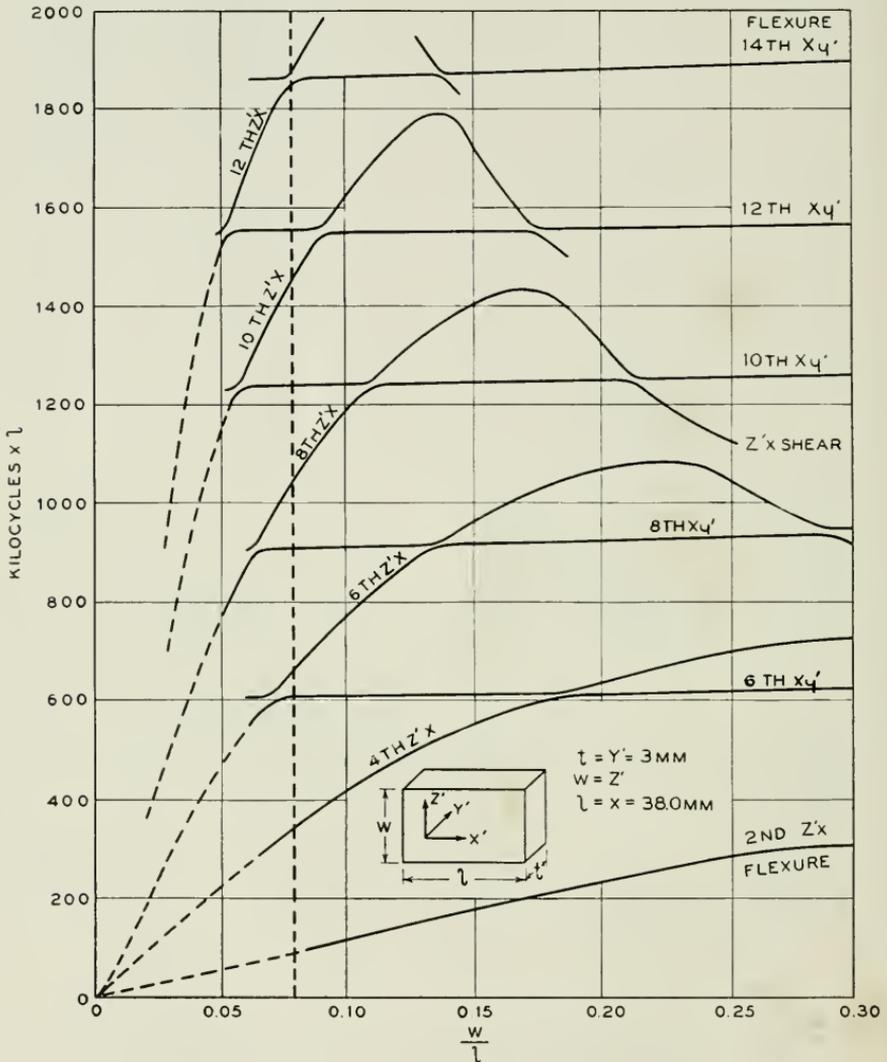


Fig. 6.12—Shear and flexure resonances in an *AC*-cut quartz plate.

for various values of $\frac{w}{l}$. The plate shear is labeled Z'_x shear and occurs at the frequencies predicted by equation 6.8 except in the regions where a flexure in the same plane exists. This is the type of motion shown in Fig. 6.4 for the case of $m = 1, n = 1$. It can be seen that as the difference in order of modes becomes greater the effect on the shear frequency is less

except where they are coexistent. We can then state generally that even though there is coupling between particular modes of motion, if the difference in order is great, the approximate frequencies may be computed as though they were isolated. This is more clearly shown in the case of thickness shear modes. The modes that are shown coupled to the face shear mode are Z'_x flexures propagated in the direction of the length or X axis. The lower orders can be shown to follow the general frequency equation discussed in section 6.3 but the higher orders for a given $\frac{w}{\ell}$, it will be noticed, are regularly spaced in frequency and show the effect of shear. The X'_y flexure modes determined by the length and thickness are shown as nearly horizontal lines since only the width was changed. Since these two groups of flexure modes are propagated in the same direction, it would be expected that the difference in frequency for the same ratio of dimension (i.e., $\frac{w}{\ell} = \frac{t}{\ell}$) would be due to the differences of the shear coefficients in the two planes of motion. The vertical dotted line indicates the ratio of thickness to length. When the ratio of width to length is equal to this value it can be seen that the flexure modes in the width-length plane are in all cases higher than the same order flexures in the thickness-length plane. An examination of Fig. 6.7 shows that for an AC -cut crystal the shear modulus in the width-length plane ($\frac{1}{\sqrt{s_{55}'}}$) is greater than that in the thickness-length plane ($\frac{1}{\sqrt{s_{66}'}}$). This is in agreement with the observation made above. One other generality may be drawn from the experimental data shown in Fig. 6.12. The coupling between flexure modes and shear modes in planes at right angles to each other is very small in comparison with that between modes in the same plane.

As mentioned before the effect of coupling between modes of motion is greatest when the orders are more nearly similar. In this particular crystal this effect can be shown between the fundamental width-length Z'_x shear and the second order width-length Z'_x flexure. This is shown in Fig. 6.13 which is an extension of the data shown in Fig. 6.12 for a crystal nearly square and shows the frequency range covered only by the second flexure and the fundamental plate shear. A computation of the uncoupled second flexure mode propagated along the length and the first plate shear mode are shown by the solid lines f_f and f_s respectively. Inserting the appropriate constants the formulae of section 6.3 become

$$f_f = \frac{1}{2\pi} \sqrt{\frac{7.85 \times 10^{11}}{12 \times 2.65}} m^2 \frac{Z'}{X^2} \tag{6.15}$$

$$f_s = \frac{1}{2} \sqrt{\frac{71.8 \times 10^{10}}{2.65}} \sqrt{\frac{1}{X^2} + \frac{1}{Z'^2}} \tag{6.16}$$

In evaluating m , account was taken only of the rotary and lateral inertia so that some error is expected at the larger ratio of axes. The curve of flexure crosses the shear curve at $\frac{w}{l} = .76$, a condition which we know to be non-compatible since these two motions are coupled. From the theory of coupled circuits we can determine the displacement of two uncoupled frequencies as a result of the coupling, through the relation

$$f_{1,2}^2 = \frac{1}{2}[f_s^2 + f_f^2 \pm \sqrt{(f_s^2 - f_f^2)^2 + 4k^2 f_s^2 f_f^2}] \quad 6.17$$

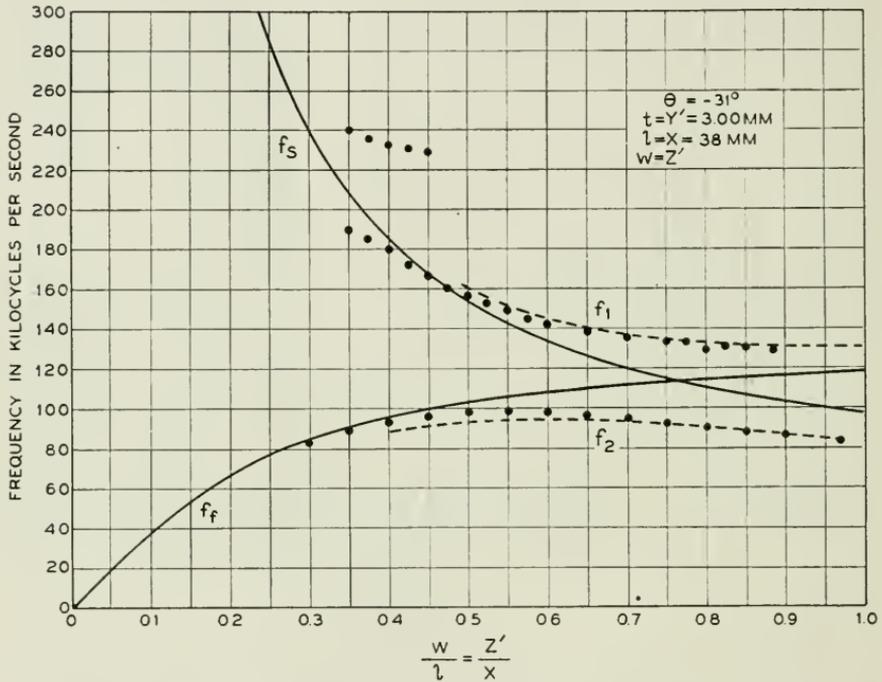


Fig. 6.13—Effect of coupling on the plate shear and the second flexure mode in an AC-cut quartz plate.

where f_s = uncoupled shear frequency,
 f_f = " flexure "
 k = coefficient of coupling.

The coefficient of coupling in this case may be defined as the ratio of the mutual to the square root of the self compliances of the two vibrating systems. As mentioned before no derivation has yet been made to indicate the relation between the coupling between these two forms of motion and the physical constants of the medium in which the vibration occurs. It is necessary to assume some coupling factor which will produce that observed

by experiment. Applying a coupling coefficient of 35% and computing the values of f_1 and f_2 from equation 6.17 the results are the dotted curves shown in Fig. 6.13. The observed points follow the computed values to a fair degree of accuracy for all frequencies below 180 kilocycles. Above this range there is a strong coupling to the fourth flexure and this would require separate consideration. Based upon these results the equation for the low frequency or face shear given in section 6.3 would not give the observed results for a nearly square plate because of the high coupling to the second flexure mode. For an approximately square plate, cut near the AC -cut the plate shear frequency including the effect of coupling would be given by

$$f = \frac{.849 \sqrt{2}}{2d} \sqrt{\frac{1}{\rho_{55}}}, \tag{6.18}$$

where

$$d = \frac{1}{2}(X + Z')$$

and .849 is the factor resulting from the use of equation 6.17. For crystal cuts far different from the above it would be necessary to consider the flexure and shear as uncoupled and then apply equation 6.17 to determine the appropriate factor for square plates.

2. High Frequency Shear

The motion associated with flexure has been shown in Fig. 6.1 and in order to determine the frequency of higher order flexures, measurements were made on an AC -cut crystal. The results of these measurements are shown in Fig. 6.12. The first flexure motion to be expected with this crystal would be a flexure in the plane of the length and width. The various orders of these flexures are shown by the curved lines labeled second z'_x fourth, sixth, etc., all radiating from zero frequency (Primed values of z and y indicate that these are not crystallographic axes). The equation commonly determining the frequency of flexure states that the frequency should be proportional to the width and inversely proportional to the square of the length. If this were true, these curved lines representing the resonances of this type flexure shown on Fig. 6.12 would then be straight lines. Since the actual conditions show a wide departure from this, we must assume that this departure is due to rotary and lateral inertia and the effects of shear. It will be noticed that as we progressively increase the order of the harmonic, that the actual frequency spacing for a given value of $\frac{w}{l}$ is very nearly linear instead of a square law. This point is more clearly seen when we examine the frequency of higher orders of the flexures in the length thickness or xy' plane. As shown on Fig. 6.12 these frequencies

labeled 6th x_v , etc., change very little and are nearly horizontal straight lines. Here again they appear to be simple harmonics of some common low frequency. Also it will be noted that the coupling between the z'_x flexures and the z'_x shear is quite appreciable and in general decreases as the difference in order of the two modes becomes greater. This plot of the various flexure frequencies tells us a great deal about the behavior of progressively higher order of flexure type motion. The important effect to be noticed is that for high orders, and a fixed ratio of $\frac{w}{\ell}$, the flexure may be treated as though it were harmonic so far as frequency is concerned. Some variations to this rule will be observed and special cases will be discussed. So far we have discussed the case of flexure modes of relatively low order. In the case of high frequency shear modes of motion, we would expect that the order of flexure which would interfere with this type of motion would be rather high.

Figure 6.14 shows a plot of these flexure modes as observed in an *AT*-cut plate. These are shown by dashed lines. The dots indicate actual measured resonances. This figure also shows the various other resonant frequencies observed in this type of plate as discussed in section 6.2. The solid lines labeled *mnp* represent the type of shear motion shown in Fig. 6.5. Here again we may observe certain statements made before with respect to the coupling between shear and flexure type motions. Notice in this case that the coupling between an even order flexure and an odd order shear is high and increases as the orders more nearly approach each other. For example, the 38th flexure mode is coupled to the fundamental shear labeled $m_1n_1p_1$ has very little coupling to the second order shear $m_1n_2p_1$, and again is strongly coupled to the third shear $m_1n_3p_1$ and correspondingly higher coupling to the fifth shear. When we speak of higher order shears, such as $n_2n_3n_5$, they are not higher order in the sense of harmonics, but do differ by a small amount in frequency. In the case of a plate where ℓ is not great compared to t , these differences will be greater.

In actual practice in the case of *AT* plates, we are usually concerned mainly with the fundamental high frequency shear and high even order flexures along the length. This case is shown in Fig. 6.15 which gives experimental results of measurements on actual *AT* plates. It will be noticed that the flexure frequencies show a rather regular displacement as the ratio of the length of the plate to its thickness is changed. In this case only the odd order modes of shear and the even modes of flexure are shown. It will be observed that as the ratio of the length to thickness decreases, the coupling between these modes is quite high. This some state of affairs is illustrated again in the case of the third harmonic of high frequency shear and is shown in Fig. 6.16. The near vertical dashed lines represent even order

flexure frequencies and the curve labeled m_3n_1 and the curve labeled m_3n_3 correspond to two different values of the high frequency shear near its commonly called third harmonic.

An examination of Figs. 6.14 and 6.15 indicates that a regular pattern is formed of the ratios of axes at which the high frequency shear and succes-

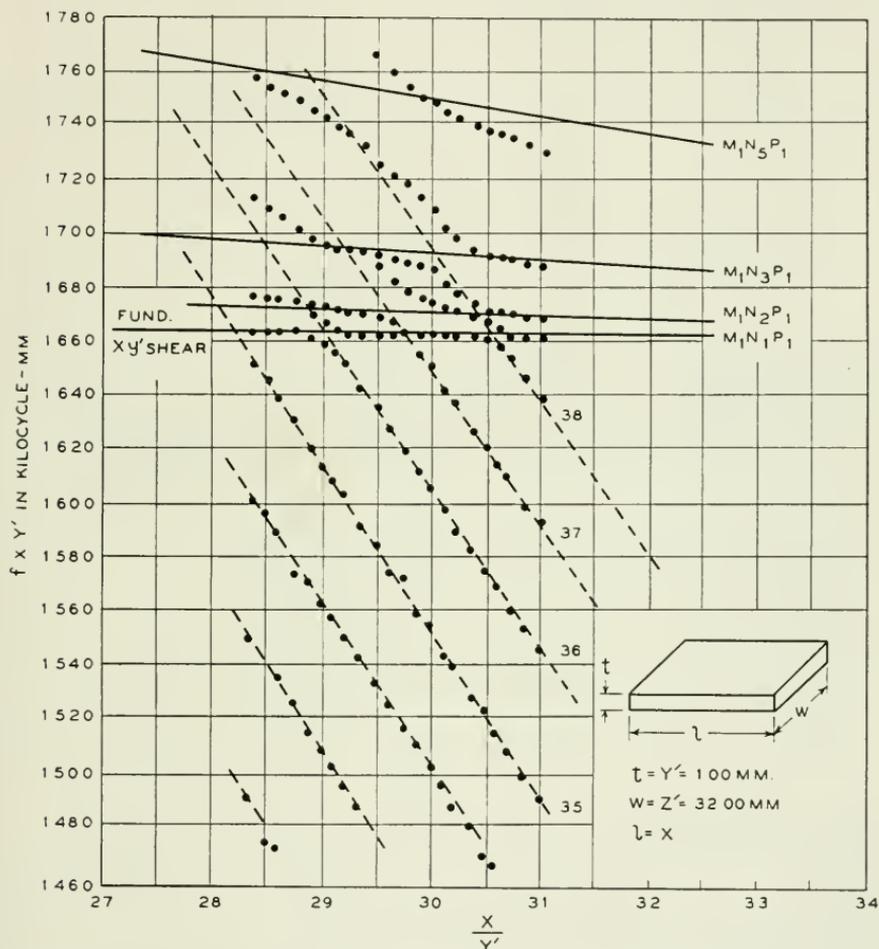


Fig. 6.14—High frequency flexure and shear resonances in an AT-cut quartz plate.

sive even orders of the length-thickness flexure coincide. Rather than define these points on the basis of specific ratios of axes it is more convenient to place them on a frequency basis. Therefore we may say that for a given size plate there will be specific frequencies at which some mode of the flexure motion along the length will be the same as the high frequency thickness

shear. For the case of *AT* plates experiment has shown these to be given by

$$f_{xf} = \frac{1338.4}{X} n_{xf}, \text{ kilocycles} \quad 6.19$$

where X = length of X axis in millimeters,

n_{xf} = order of flexure along X axis

= 1, 2, 3, 4, etc.

In this equation as well as those of a similar nature to follow it is assumed that the thickness is such as to result in the same frequency for the high

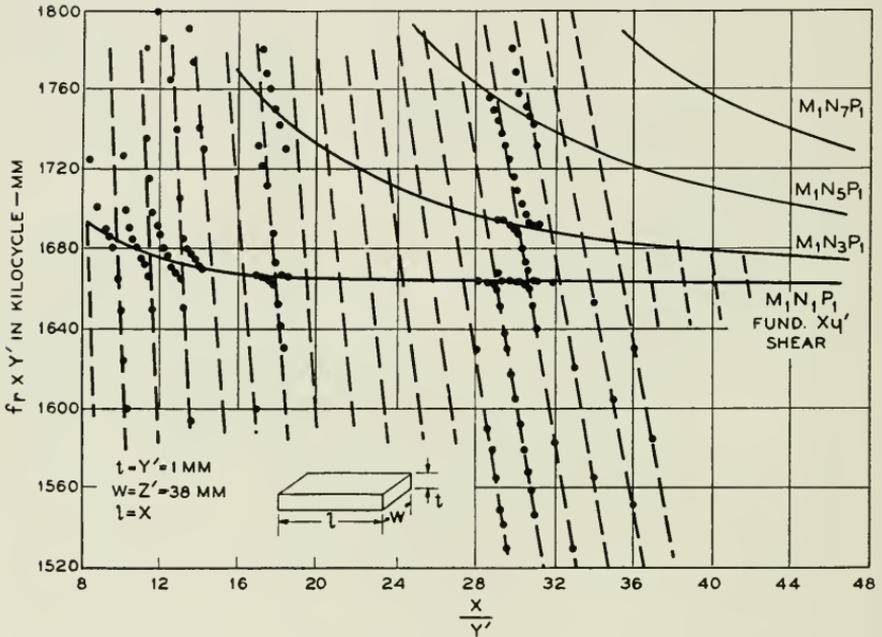


Fig. 6.15—High frequency flexure and shear resonances in an *AT*-cut quartz plate.

frequency X_v shear mode. As shown in Fig. 6.14 only the even orders are strongly coupled to the fundamental thickness shear.

The coupling between high even orders of the flexure along the X axis and the high frequency shear in the case of *BT*-cut plates is similar to that for *AT*-cut plates. Fig. 6.17 shows the various resonant frequencies observed in a *BT*-cut crystal as a result of changing the ratio of the length or X axis to the thickness or Y' axis. The curve m_1n_1 represents the high frequency X_v shear. Curves m_1n_3 , m_1n_5 , m_1n_7 and m_1n_9 represent other X_v shear modes as discussed in section 6.23 resulting from higher orders along the length or X axis. The dashed lines represent even order flexure modes along the X axis. The same regularity is observed here as in the case of the

AT-cut. When placed on a frequency rather than a ratio of axis basis the frequencies at which flexure modes along the *X* axis would coincide with the

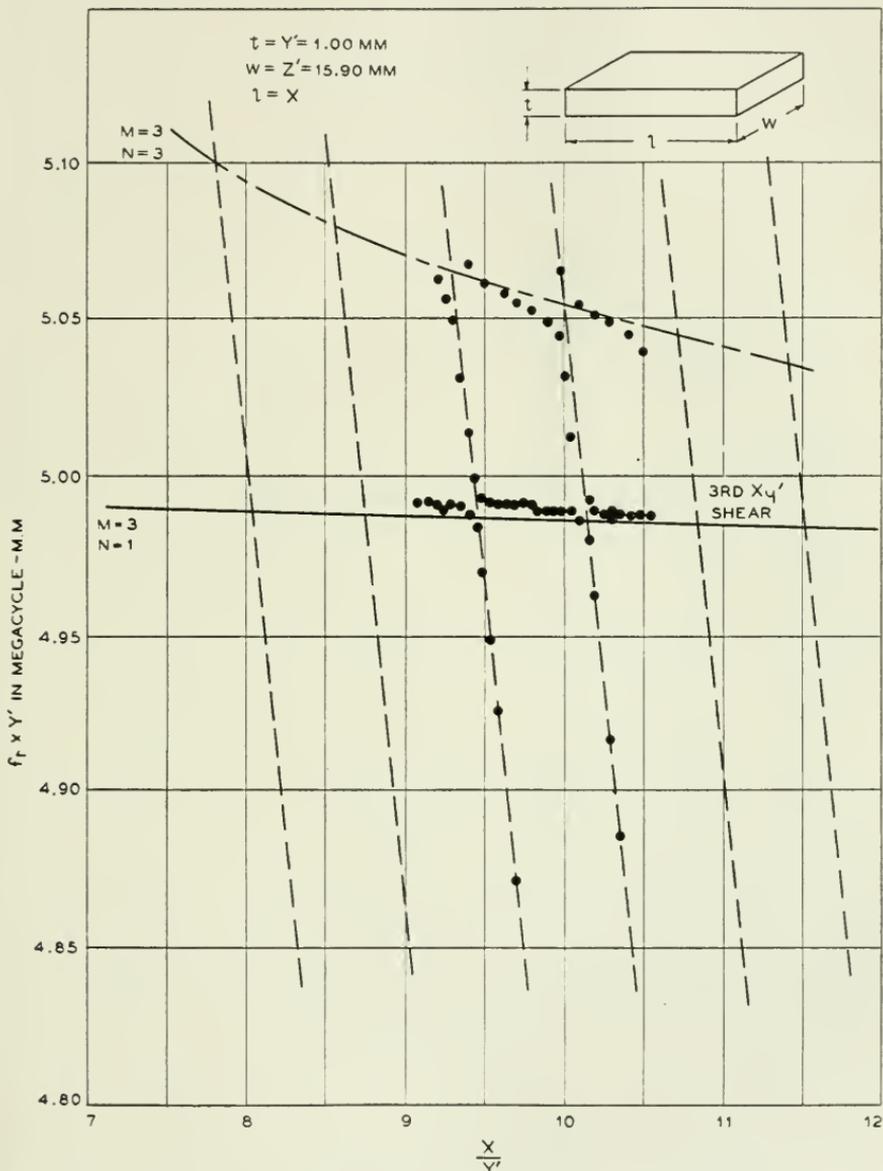


Fig. 6.16—High frequency flexure and shear resonances in an *AT*-cut quartz plate near the third harmonic shear mode.

fundamental $X_{y'}$ shear mode are experimentally given by

$$f_{xf} = \frac{1818}{X} n_{xf} \text{ kilocycles} \quad 6.20$$

where X is given in millimeters. In this case it will be noticed also that only even order flexures are strongly coupled to the fundamental Xy' shear.

The dependence of the flexure frequency on the shear coefficient can be seen from these two cases. The direction of propagation is the same in both cases (along the X axis) but the direction of particle motion is nearly at right angles. It would be expected then that the frequency constant would be highest for the case of the highest shear coefficient. Examination of equa-

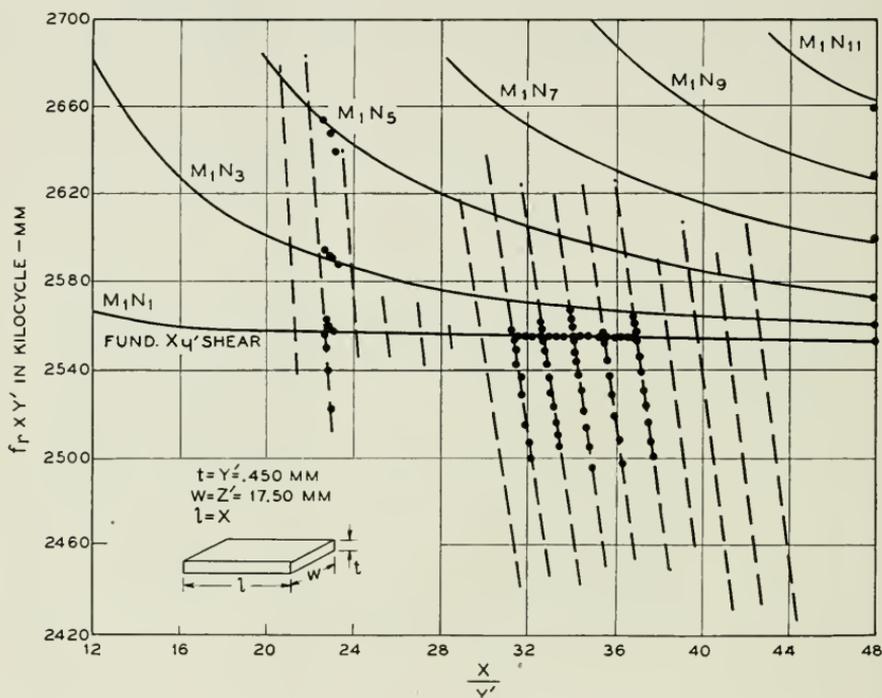


Fig. 6.17—High frequency flexure and shear resonances in a BT -cut quartz plate.

tions 6.19 and 6.20 shows this to be true. In addition, the change in the frequency constant is about the order of magnitude of the change in the shear modulus in the respective planes of motion.

6.43 Coupling between Low Frequency Shear and High Frequency Shear

From an examination of Fig. 6.7 it can be seen that the coupling between the low frequency shear (Z'_x) and the high frequency shear Xy' is related by the s'_{56} constant. In the AC and BC -cuts this reduces to zero but for the AT and BT -cuts it has a finite small value. According to section 6.3 the frequencies of the plate shear modes are given by equation 6.8 but this holds only for the case where m and n are small. When the third dimension

becomes appreciable in comparison with a half wave length along w or l it becomes necessary to use the c constants. When considering high orders of the low frequency shear equation 6.8 is modified to

$$f = \frac{1}{2} \sqrt{\frac{c_{jj}}{\rho}} \sqrt{\frac{m^2}{l^2} + k^2 \frac{n^2}{w^2}} \tag{6.21}$$

Equation 6.21 shows that high orders of the low frequency or plate shear are dependent upon both the length and width dimensions and it might be assumed that this would lead to very complicated results in so far as analysis of experimental data is concerned. The coupling between these modes and the high frequency shear is a result of coupling in the mechanical as well as

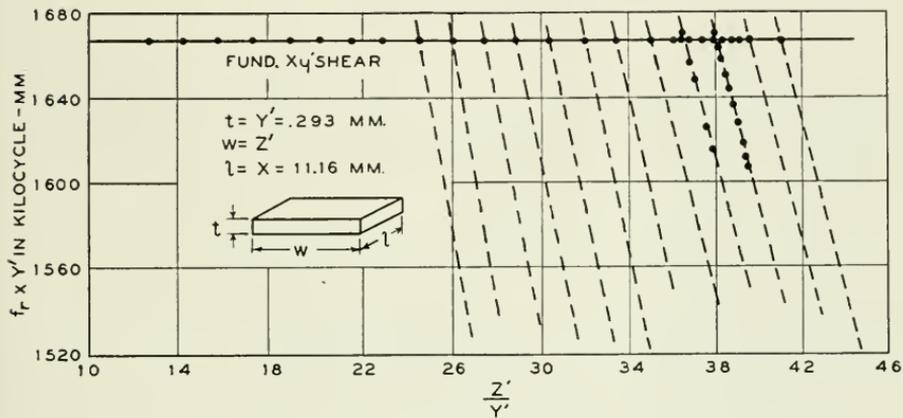


Fig. 6.18—High frequency shear resonances in an AT-cut plate.

the electrical systems. The strongest coupling with reference to the length axis would then be for high odd orders of m and unity for n with successively smaller coupling for higher orders for n if the driving potential extends over the complete surface of the crystal. In a similar manner when considering high orders of plate shear along the width axis the highest coupling will result from unit order for m . Based on these assumptions then to a first approximation we can assume these modes to be functions of length and width alone. Equation 6.21 then reduces to

$$f_{\ell_s} = \frac{1}{2} \sqrt{\frac{c_{jj}}{\rho}} \frac{n_{\ell_s}}{l} \tag{6.22}$$

$$f_{w_s} = \frac{k}{2} \sqrt{\frac{c_{jj}}{\rho}} \frac{n_{w_s}}{w} \tag{6.23}$$

where $n_{s\ell}$ = order of shear mode along ℓ axis,
 n_{sw} = order of shear mode along w axis.

These modes have been measured in *AT* and *BT*-cut crystals. Fig. 6.18 shows the points at which these modes intersect the fundamental high frequency shear mode in *AT*-cut plates. This is the case for high orders along the Z' or width axis. A similar set of resonances can be shown to exist when the X or length axis is varied. Experiment has shown that these frequencies of coincidence between high order plate shear modes and the fundamental high frequency $X_{y'}$ shear mode for the case of *AT*-cut plates is given by

$$f_{zs} = \frac{254.2}{X} n_{zs} \text{ kilocycles} \quad 6.24$$

$$f_{z's} = \frac{254.0}{Z'} n_{z's} \text{ kilocycles} \quad 6.25$$

where X and Z' are given in centimeters. Only odd orders are strongly coupled if the crystal plate has a symmetrical contour with respect to an applied equipotential electrode. Upon substitution of the value of c'_{55} for an *AT*-cut crystal in equation 6.22 there results

$$f_s \times \ell = \frac{1}{2} \sqrt{\frac{c'_{55}}{\rho}} = \frac{1}{2} \sqrt{\frac{67 \times 10^{10}}{2.65}} = 251.0 \text{ kilocycle} - \text{cm.} \quad 6.26$$

which is within 1 per cent of that found experimentally. Since Young's modulus is nearly the same along the X and Z' axis the value of k in equation 6.23 is essentially unity. Fig. 6.19 shows measured values of high order Z'_z shear modes near the high frequency $X_{y'}$ shear mode in a *BT*-cut crystal for various values of the width or Z' axis. More detailed measurements have been made of the high order Z'_z plate shear modes in *BT*-cut plates along the X axis. Fig. 6.20 shows both the shear and flexure modes along the X axis near the vicinity of the high frequency $X_{y'}$ shear mode. Since the frequency constant for the Z'_z shear modes is different from that for the $X_{y'}$ flexures there are regions where, if no coupling existed, all three modes would be at the same frequency. It is obvious from Fig. 6.20 that this is not the case. Therefore, we must assume that not only are the high order Z'_z shears and $X_{y'}$ flexures coupled to the high frequency $X_{y'}$ shear but that they are coupled to each other.

While it is difficult to see from Fig. 6.20 the relative coupling of flexures to the $X_{y'}$ shear, experiment has shown the flexure modes along X to have the greater coupling to the $X_{y'}$ shear. This is true when the ratio $\frac{X}{Y'}$ is such that the flexure modes along X and high order Z'_z shear modes along X have their maximum separation. When these modes approach each other

and the X_y' shear such as is shown in Fig. 6.21 at $\frac{X}{Y'} = 31.35$ the relative coupling of each to the X_y' shear is about equal. This arises from the fact that the mutual coupling between them increases the apparent coupling

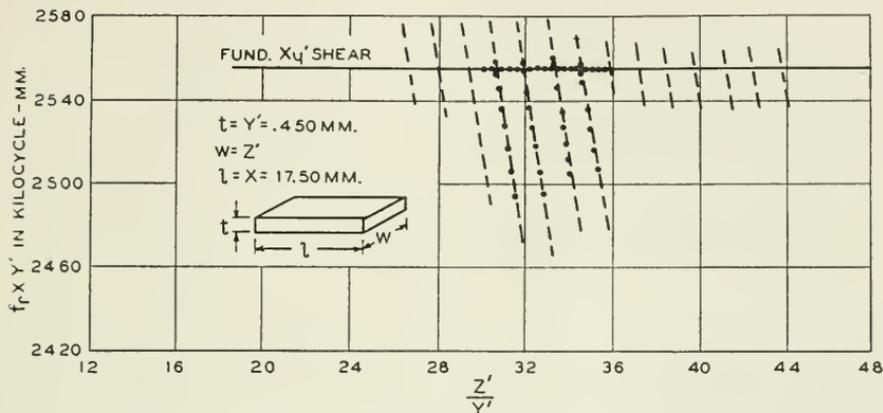


Fig. 6.19—High frequency shear resonances in a BT-cut plate.

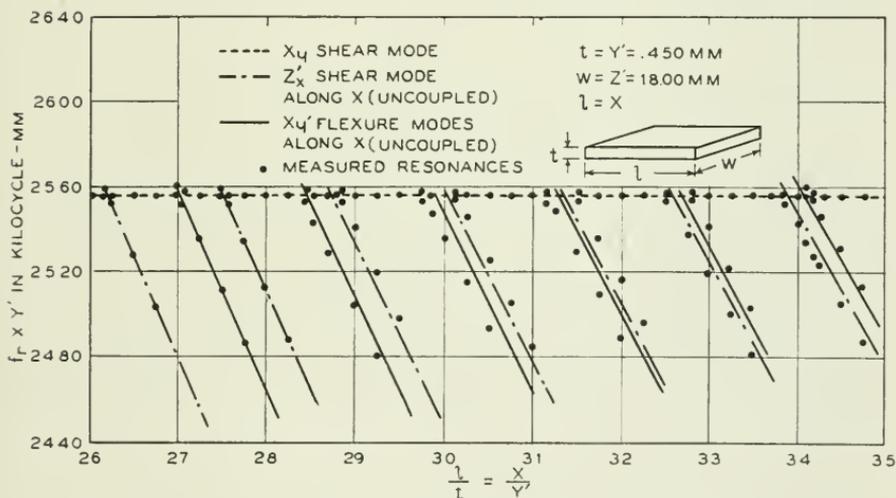


Fig. 6.20—High frequency thickness shear and flexure and shear resonances along the X axis in a BT-cut quartz plate.

between the X_y' shear and high orders of Z_x' shear along X. From this it would appear advisable to avoid such regions in the dimensioning of crystals for oscillator use over wide temperature ranges. Determination of the flexure as well as high order Z_x' shears then must be made in regions where

they are spaced so that the effect of coupling between them will not influence the frequency constant that is determined experimentally. These regions have been investigated and the result for the flexure modes is that shown

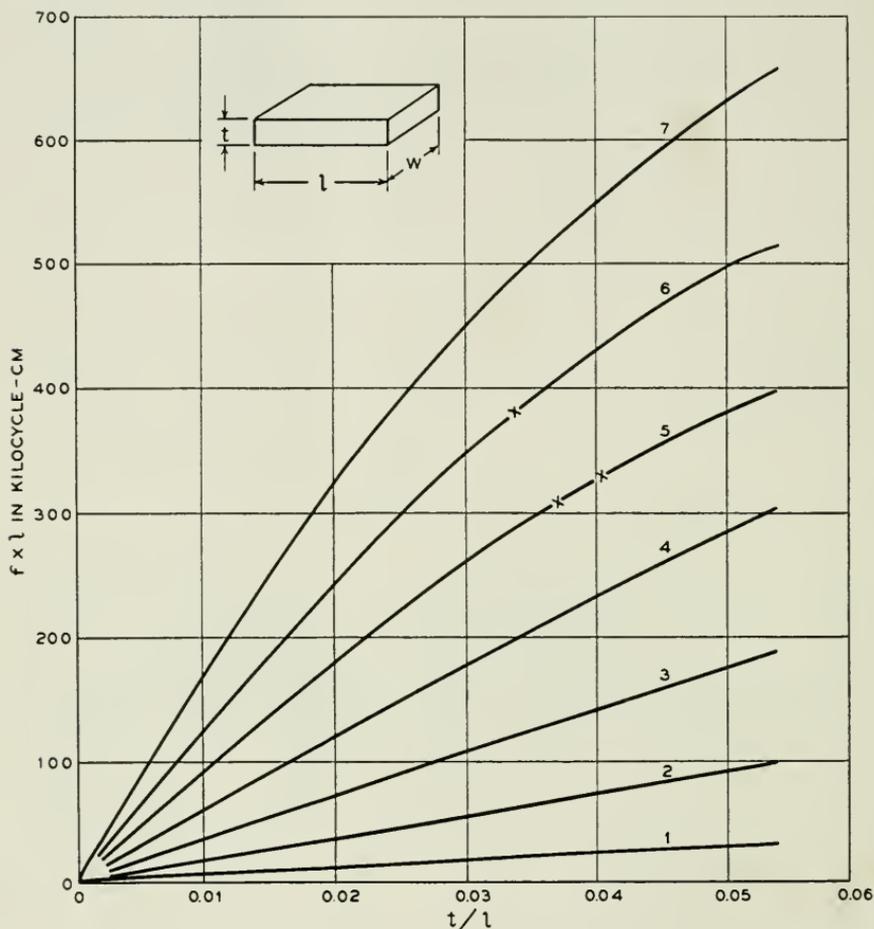


Fig. 6.21—Flexure resonances in a GT -cut quartz plate.

in equation 6.20. From Fig. 6.19 the high order Z'_x shears along Z' will be coincident with the high frequency X'_y shear at frequencies given by

$$f_{z's} = \frac{166.45}{Z'} n_{z's} \text{ kilocycles} \quad 6.27$$

From Fig. 6.20 high orders of the same Z'_x shear along X will be coincident with the high frequency X'_y shear at frequencies given by

$$f_{zs} = \frac{163.514}{X} n_{zs} \text{ kilocycles} \quad 6.28$$

Upon substitution of the value of c'_{55} for a *BT*-cut in equation 6.22 there results

$$f_{ts} \times \ell = \frac{1}{2} \sqrt{\frac{c'_{55}}{\rho}} = \frac{1}{2} \sqrt{\frac{30.3 \times 10^{10}}{2.65}} = 169.0 \text{ kilocycles} - \text{cm.}$$

which is 3.3% greater than that observed in equation 6.28 and 1.6% greater than that shown in equation 6.27. The apparent difference in the observed shear modulus in the *X* and *Z'* directions for the *BT*-cut can be explained from the fact that Young's modulus is quite different in the two directions for the *BT*-cut while it is nearly the same for the *AT*-cut as verified by equation 6.24 and 6.25.

From the discussion in this section it can be seen that a single theory that would relate all the now known resonances in quartz plates together with the effects of coupling would be prodigious indeed. In order to reduce the design of quartz plates to a simple engineering basis it is necessary to take specific examples and investigate the region in the vicinity of the frequency to be used based on general theory and then apply approximations that fit the specific cases.

6.5 METHODS FOR OBTAINING ISOLATED MODES OF MOTION

6.51 *GT* Type Crystals

In the case of *GT* type crystals the modes that cause the greatest concern are flexure modes in the two planes of the length and thickness and the width and thickness. The desired mode is that of an extensional mode along the width. To produce a low temperature coefficient it is also necessary that this mode be coupled to an extensional mode along the length, a fixed frequency difference from it. Therefore it will be necessary to prevent flexure modes from occurring at either of these two frequencies. Fig. 6.21 shows the frequency of various flexure modes that would be observed in *GT*-cut plates for different ratios of thickness to length. In the case of the *GT*-cut the elastic constants in the length and width directions are the same and therefore it is only necessary to determine the flexures in one plane to get a determination in both. From the plot of frequencies shown in Fig. 6.21, it would be very easy to determine the proper thickness for any given *GT* plate. Since in all practical cases there is a definite relation between the length and width of this type of plate, it would be necessary to examine the flexures in these two directions as a function of the change in thickness.

Fig. 6.22 shows a plot of this for the case of a *GT* crystal designed to operate at 164 kilocycles. All the information shown in this figure is obtained directly from Fig. 6.21. Since a change in thickness will not have any effect upon the length and width extensional modes of vibration and only

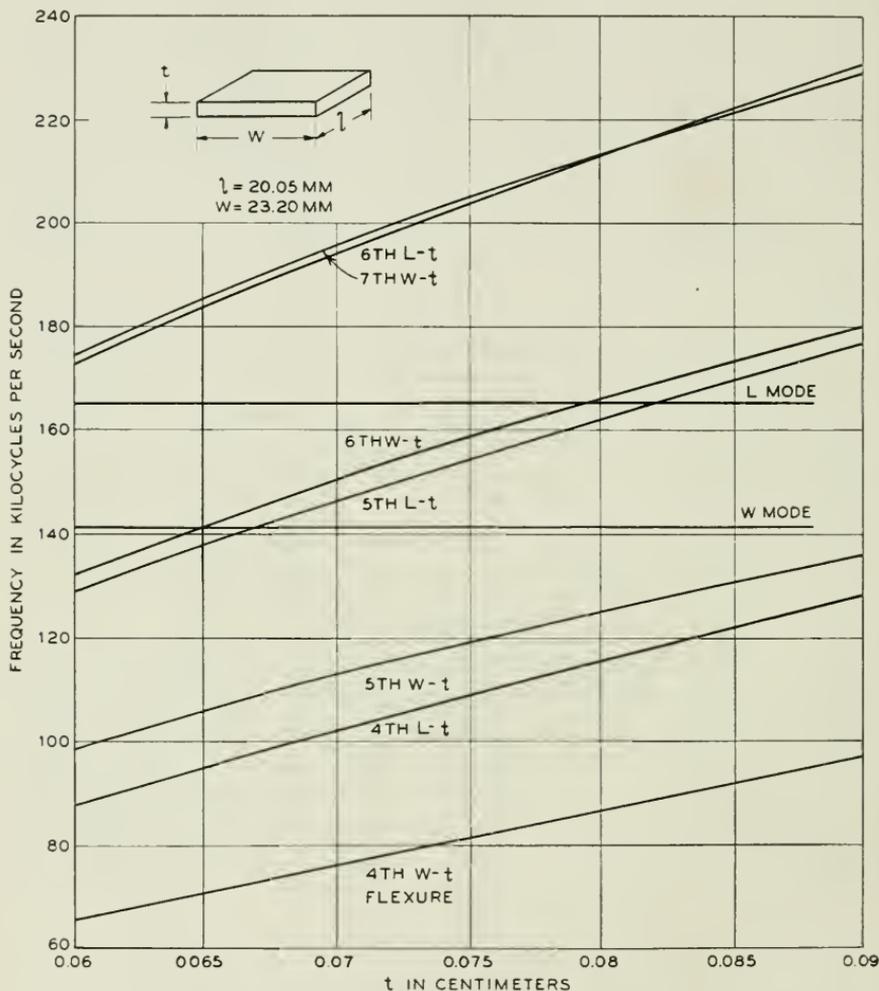


Fig. 6.22—Flexure and extensional resonances in a 164 kc *GT*-cut quartz plate.

changes the flexure frequencies, it would be reasonable to suppose that some thickness could be obtained where no flexure along the length or width would be of the same frequency as the length or width extensional mode. Examining the curves of Fig. 6.22, we find that a thickness of .06 cm., .075 cm. or .085 cm. would meet these conditions.

6.52 *BT Type Crystals*

As discussed in Section 6.4 the modes showing the greatest coupling to the high frequency thickness shear are of two types: high orders of X_y , flexure propagated along the X axis and high order Z'_z shears along the X and Z' axes independently. Complex orders of the flexure and plate shear as illustrated in Fig. 6.2 and Fig. 6.4 do cause considerable difficulty and their analysis calls for special treatment and is not within the scope of this text. For the case of the *BT*-cut the three primary interfering series of modes are given by

$$\begin{aligned} f_{zf} &= \frac{181.8}{X} n_{zf} \text{ kilocycles} \\ f_{zs} &= \frac{163.514}{X} n_{zs} \text{ kilocycles} \\ f_{z's} &= \frac{166.45}{Z'} n_{z's} \text{ kilocycles} \end{aligned} \tag{6.30}$$

where X and Z' are given in centimeters and f_{zf} is the frequency at which integral orders of flexure modes along the X axis would coincide with the high frequency thickness shear mode. In a similar manner f_{zs} and $f_{z's}$ relate the same conditions for integral orders of the plate shear modes. These equations are true only in the case where the thickness is of such a value as to place the high frequency thickness shear mode at the same frequency as the computed interfering mode. In most practical cases for oscillator use the electric field is applied to the crystal by means of a flat electrode on each side of the crystal plate. Under this condition only odd order X_y , shear modes along the X axis are excited and hence the strongest couplings to the X_y , flexure modes will be only for even order values of n_{zf} in equation 6.30. In a similar manner the greatest interference between the X_y , shear mode and high orders of the Z'_z shear modes along both X and Z' will occur for odd orders. Therefore the strongest interference from these modes will occur only for odd integers of n_{zs} and $n_{z's}$ in equation 6.30. These assumptions of only even flexures and odd shears showing appreciable coupling are based upon a crystal plate cut precisely along its proper axis and of uniform contour assembled in a holder using electrodes of uniform air gap. Deviations from these conditions will of course alter the ideal results dependent upon the amount and type of deviation.

The relationships shown in equation 6.30 may be more clearly seen when plotted graphically. Assuming a *BT*-cut crystal plate 1 centimeter square we may determine the frequencies at which an interfering mode will coincide with the high frequency X_y , shear by assigning even integers to n_{zf} and odd

integers to n_{x_s} and $n_{z'_s}$. Fig. 6.23 shows a plot of these three types of interfering modes on a folded frequency scale covering the range from 5 to 15 megacycles for a plate 1 centimeter square. Each abscissae covers a range of one megacycle with dots at three levels. The first level shows the frequencies at which successive even orders of flexure along the X axis occurs. The second level shows successive odd Z'_x shear modes along X and the third level successive odd Z'_x shear modes along Z' . The circles shown on the three levels indicate the results of actual measurements on BT -cut crystals as resonating elements. It will be noticed that the circles and dots coincide for most frequencies, the regions of departure occur only when a high order shear mode and a high order flexure mode along the X axis approach each

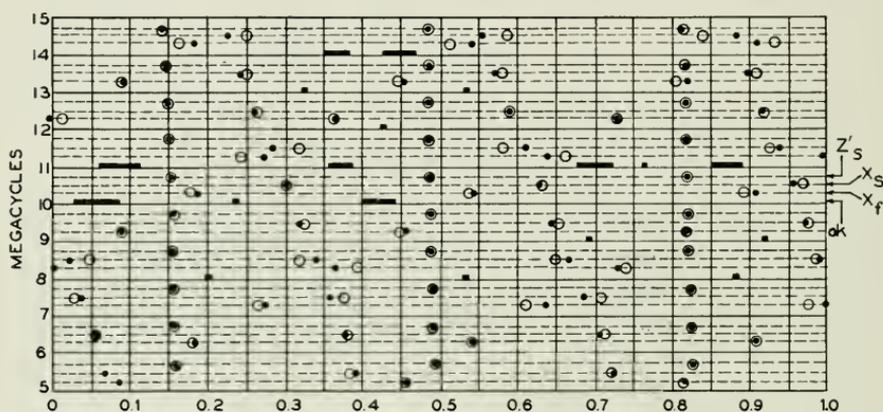


Fig. 6.23—Frequencies at which the Z'_z shear along X , the Z'_z shear along Z' and the X_y flexure along X coincide with the high frequency X_y shear in BT -cut crystals.

other in frequency. The reason for this is obvious from the previous discussion on the coupling between flexure and shear modes of motion.

The chart of Fig. 6.23 is of course not limited to a crystal 1 centimeter square or for that matter even a square crystal. In reality it relates the product of the frequency and X and Z' dimensions. For example a flexure mode interferes with the high frequency shear mode at a frequency of 9.45 megacycles for a plate with X dimension equal to 1 centimeter. If the X and Z' dimensions were doubled the same situation would exist at one half the frequency. In determining the dimensions for a crystal at a given frequency we know that the product of the frequency and X dimensions as well as Z' dimension must not result in a frequency close to those given by the circles of Fig. 6.23. In addition other interfering modes as previously mentioned must be avoided. These at present may be determined experimentally by choosing regions on the chart clear of the known flexure and shear modes.

On the abscissae are shown certain discreet frequencies as well as frequency ranges which have been found to result in crystal units having no serious dips in activity over a wide range in temperature. These are for square crystals in the 18 millimeter size range and have been obtained by Mr. G. M. Thurston of the Bell Laboratories and Mr. F. W. Schramm of the Western Electric Company. It will be noted that no so-called ok regions have been found at the frequencies of the three principal coupled modes.

While the use of the chart shown in Fig. 6.23 will often lead directly to the proper X and Z' dimensions for a given oscillator it cannot be overemphasized that only the three principal interfering modes are shown and only the odd orders for the shears and only the even orders for the flexure modes. Since the even order shear modes are excited due to slight variations which would produce wedge shaped air gaps or quartz blanks, it is advisable to avoid these regions also. Complex combinations of the three principal modes as shown in Figs. 6.2 and 6.4 are also driven. Therefore when it is necessary to produce a crystal unit possessing the highest activity for a given area of quartz plate over an extended temperature range it is necessary to scan the supposed desirable regions shown in Fig. 6.23 by complete measurements on finished units of a given size and varying frequency or of constant frequency and varying size. As an illustration the region shown in Fig. 6.23 between 10.025 and 10.080 megacycles was determined in this manner with the use of crystal plates approximately 18 millimeters square. The use of crystals with other than square dimensions could undoubtedly have increased the range of this region but their use is undesirable from a manufacturing standpoint. Assuming that the electrodes and crystal holder permit a variation in size of the quartz plate from 17.20 millimeters to 18.20 millimeters this approved region will immediately specify the dimensions of crystals to cover the frequency range from 5508 to 5727 kilocycles. This also assumes crystal blanks cut to precise orientations with controlled contours and electrodes of uniform flatness and constant airgap. While the theory would indicate that the frequency range given above could be expanded to considerably higher values by utilizing a smaller crystal blank this has not been proven so far since most crystals produced by the Western Electric Company require large area plates to meet high activity requirements.

As an illustration of the effect on the behavior of oscillators of changing the X and Z' dimensions of BT -cut quartz plates measurements have been made of the activity, in a conventional tuned plate circuit with the crystal connected between grid and cathode of quartz plates of constant thickness and varying X and Z' dimensions. Fig. 6.24 shows the effect of changing the X dimension of a quartz plate on its activity as an oscillator. By taking the product of the frequency and dimension we can determine the dimen-

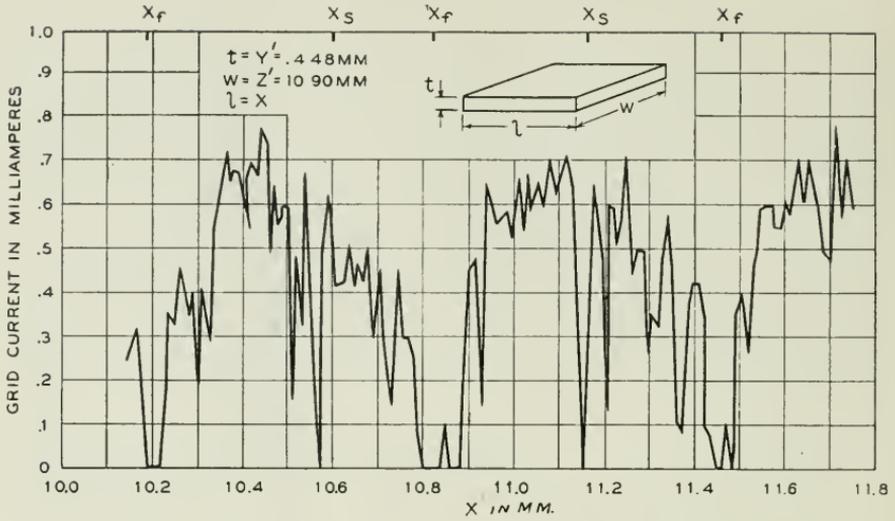


Fig. 6.24—Effect of change in X dimension on the activity of a BT -cut quartz plate in an oscillating circuit.

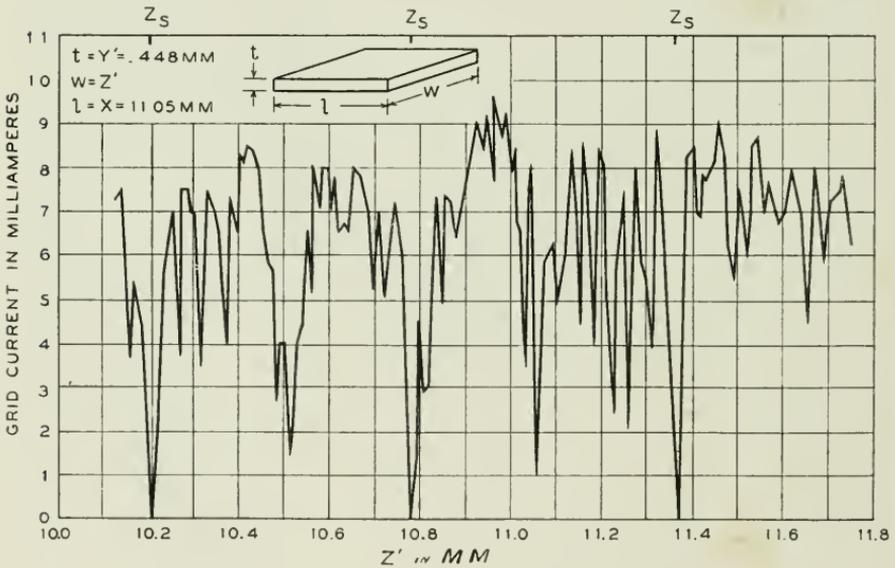


Fig. 6.25—Effect of change in Z' dimension on the activity of a BT -cut quartz plate in an oscillating circuit.

sions from Fig. 6.27 for this case where the X_y flexures and Z'_x shears will interfere to produce poor characteristics. These are shown in Fig. 6.24 for flexure modes as X_f and for the shear modes as X_s and do in general cor-

respond to the dimensions resulting in low or no activity. This illustrates quite clearly the necessity for grinding the edges of plates not dimensioned for a specific frequency. Fig. 6.25 shows the same conditions when only the Z' dimension is changed. In this case the dimensions shown at regular intervals as Z_s were derived from Fig. 6.25 as before and correspond to the zero activity dimensions found experimentally. It will be noticed that low activity regions are found halfway between the dimensions designated as Z_s . These correspond to even orders of the Z'_z shear and are the result of a slight wedge in the airgap. This was intentional to show the existence of this condition.

Figures 6.24 and 6.25 show the necessity for avoiding certain dimensions for oscillator plates at specific frequencies. This can be accomplished by

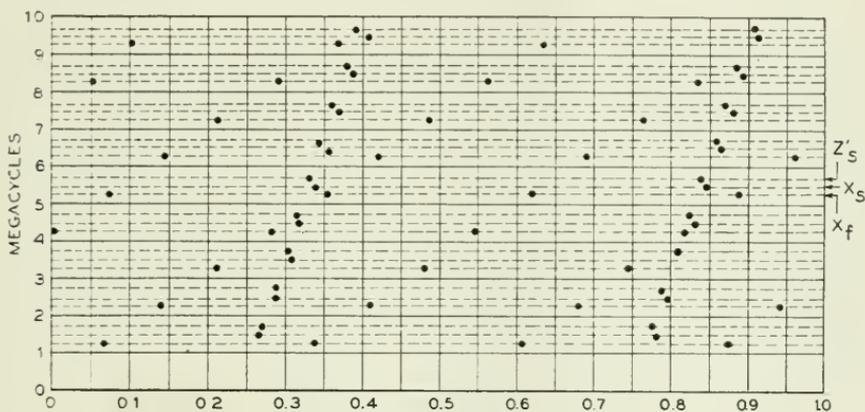


Fig. 6.26—Frequencies at which the Z'_z shear along X , the Z'_z shear along Z' and the X_y flexure along X coincide with the high frequency X_y shear in AT -cut plates.

individually adjusting the X and Z' dimensions by hand grinding of each plate or by predetermining the proper dimensions and using mass production methods of precise machine grinding. The advantages of predimensioned crystal units is the insurance of proper operation over a wide temperature range and uniformity of activity. The experience of most manufacturers of low frequency crystal units in the broadcast range and high frequency crystals requiring high activity over a wide temperature range has been that it is necessary to use specific dimensions to insure low rejects in the final tests.

6.53 AT -Type Crystals

The modes of motion encountered in the AT -cut crystal are the same as that of the BT -cut. The effects of coupling between most modes is greater

due to the increased piezo electric constant for this particular cut, and the frequency constants are different due to the change in angle with respect to the crystallographic axes. The three series of interfering modes as described for the *BT*-cut have been measured for this crystal and as shown in Section 6.4 are

$$\begin{aligned} f_{xf} &= \frac{133.84}{X} n_{xf} \\ f_{xs} &= \frac{254.20}{X} n_{xs} \\ f_{z's} &= \frac{254.00}{Z'} n_{z's} \end{aligned} \quad 6.31$$

In a manner similar to the *BT* case a chart has been developed of a folded frequency scale showing the frequencies at which even order X_y' flexure modes propagated along X and odd order Z_x' shear modes along X as well as odd order Z_x' shear modes along Z' will interfere with the high frequency X_y' shear mode for a crystal 1 centimeter square. This is shown in Fig. 6.26. Its use is the same as that described for the *BT* case. Insufficient experimental work has been done to indicate the relative shift in the flexure and shear modes along the X axis when they approach each other in frequency. Also, most of the use of square plates and experimental work has been confined to the *BT*-cut crystals and hence no ok regions are shown for this chart.

APPENDIX B

Equation of elastic and piezoelectric constants for rotation of axes about the X axis. ($s = \sin \theta$; $c = \cos \theta$)

$$\begin{aligned} c'_{11} &= c_{11} \\ c'_{22} &= c_{11}c^4 + c_{33}s^4 + 2(2c_{44} + c_{13})s^2c^2 + 4c_{14}sc^3 \\ c'_{33} &= c_{11}s^4 + c_{33}c^4 + 2(2c_{44} + c_{13})s^2c^2 - 4c_{14}s^3c \\ c'_{44} &= c_{44} + (c_{11} + c_{33} - 4c_{44} - 2c_{13})s^2c^2 - 2c_{14}(c^2 - s^2)sc \\ c'_{55} &= c_{44}c^2 + c_{66}s^2 + 2c_{14}sc \\ c'_{66} &= c_{44}s^2 + c_{66}c^2 - 2c_{14}sc \\ c'_{12} &= c_{12}c^2 + c_{13}s^2 - 2c_{14}sc \\ c'_{13} &= c_{12}s^2 + c_{13}c^2 + 2c_{14}sc \\ c'_{14} &= c_{14}(c^2 - s^2) + (c_{12} - c_{13})sc \end{aligned}$$

$$c'_{23} = c_{13}(c^4 + s^4) + (c_{11} + c_{33} - 4c_{44})s^2c^2 - 2c_{14}(c^2 - s^2)sc$$

$$c'_{24} = c_{14}(4s^2 - 1)c^2 + [c_{11}c^2 - c_{33}s^2 - (2c_{44} + c_{13})(c^2 - s^2)]sc$$

$$c'_{34} = -c_{14}(4c^2 - 1)s^2 + [c_{11}s^2 - c_{33}c^2 + (2c_{44} + c_{13})(c^2 - s^2)]sc$$

$$c'_{56} = c_{14}(c^2 - s^2) + (c_{66} - c_{44})sc$$

$$c'_{16} = c'_{16} = c'_{25} = c'_{26} = c'_{35} = c'_{36} = c'_{45} = c'_{46} = 0$$

$$s'_{11} = s_{11}$$

$$s'_{22} = s_{11}c^4 + s_{33}s^4 + (s_{44} + 2s_{13})s^2c^2 + 2s_{14}s^3c$$

$$s'_{33} = s_{11}s^4 + s_{33}c^4 + (s_{44} + 2s_{13})s^2c^2 - 2s_{14}s^3c$$

$$s'_{44} = s_{44} + 4(s_{11} + s_{33} - s_{44} - 2s_{13})s^2c^2 - 4s_{14}(c^2 - s^2)sc$$

$$s'_{55} = s_{44}c^2 + s_{66}s^2 + 4s_{14}sc$$

$$s'_{66} = s_{44}s^2 + s_{66}c^2 - 4s_{14}sc$$

$$s'_{12} = s_{12}c^2 + s_{13}s^2 - s_{14}sc$$

$$s'_{13} = s_{12}s^2 + s_{13}c^2 + s_{14}sc$$

$$s'_{14} = s_{14}(c^2 - s^2) + 2(s_{12} - s_{13})sc$$

$$s'_{23} = s_{13}(c^4 + s^4) + (s_{11} + s_{33} - s_{44})s^2c^2 - s_{14}(c^2 - s^2)sc$$

$$s'_{24} = s_{14}(4s^2 - 1)c^2 + [2(s_{11}c^2 - s_{33}s^2) - (s_{44} + 2s_{13})(c^2 - s^2)]sc$$

$$s'_{34} = -s_{14}(4c^2 - 1)s^2 + [2(s_{11}s^2 - s_{33}c^2) + (s_{44} + 2s_{13})(c^2 - s^2)]sc$$

$$s'_{56} = 2s_{14}(c^2 - s^2) + (s_{66} - s_{44})sc$$

$$s'_{16} = s'_{16} = s'_{25} = s'_{26} = s'_{35} = s'_{36} = s'_{45} = s'_{46} = 0$$

$$d'_{11} = d_{11}$$

$$d'_{12} = -(d_{14}s + d_{11}c)c$$

$$d'_{13} = (d_{14}c - d_{11}s)s$$

$$d'_{14} = d_{14}(c^2 - s^2) - 2d_{11}sc$$

$$d'_{25} = -(d_{14}c + 2d_{11}s)c$$

$$d'_{26} = (d_{14}s - 2d_{11}c)c$$

$$d'_{35} = -(d_{14}c + 2d_{11}s)s$$

$$d'_{36} = (d_{14}s - 2d_{11}c)s$$

$$d'_{16} = d'_{16} = d'_{21} = d'_{22} = d'_{23} = d'_{24} = d'_{31} = d'_{32} = d'_{33} = d'_{34} = 0$$

$$e'_{11} = e_{11}$$

$$e'_{12} = -(2e_{14}s + e_{11}c)c$$

$$e'_{13} = (2e_{14}c - e_{11}s)s$$

$$e'_{14} = e_{14}(c^2 - s^2) - e_{11}sc$$

$$e'_{25} = -(e_{14}c + e_{11}s)c$$

$$e'_{26} = (e_{14}s - e_{11}c)c$$

$$e'_{35} = -(e_{14}c + e_{11}s)s$$

$$e'_{36} = (e_{14}s - e_{11}c)s$$

$$e'_{15} = e'_{16} = e'_{21} = e'_{22} = e'_{23} = e'_{24} = e'_{31} = e'_{32} = e'_{33} = e'_{34} = 0$$

Response of a Linear Rectifier to Signal and Noise*

By W. R. BENNETT

WHEN the input to a rectifier contains both signal and noise components, the resultant output is a complicated non-linear function of signal and noise. Given the spectra of the signal and noise input waves, the law of rectification, and the transmission characteristics of the input and output circuits of the rectifier, it should, in general, be possible to describe the spectrum of the resultant output wave. Before discussing the solution of the general problem, we shall derive some results of a simpler nature, which do not require a consideration of the distribution of the signal and noise energies as functions of frequency.

I. DIRECT-CURRENT COMPONENT OF OUTPUT

A quantity of considerable importance is the average value of the output amplitude. This is the quantity which would be read by a direct-current meter. Calculation of the average or d-c response can be performed in terms of the distribution of instantaneous output amplitudes in time. The distribution of output amplitude can be computed from the distribution of instantaneous input amplitudes and the law of rectification.

As an example, we shall compute the average current obtained from a linear rectifier when the input to the rectifier consists of a sinusoidal signal with random noise superposed upon it. The probability density function of the signal voltage is first determined, and the result given in (3). The corresponding probability density for the voltage of the noise is well known and is given in (4). The distribution of occurrence of the resultant instantaneous amplitudes of the combined noise and signal voltages is then computed by the rules of mathematical probability, and the result is shown in (7). The assumption that the rectifier is linear then leads directly to an integral which yields the average current obtained from the rectifier.

Let the signal voltage, E_s , be given by

$$E_s = P_o \cos \omega t. \quad (1)$$

The possible angular values of ωt are uniformly distributed throughout the range 0 to 2π . The range E_s to $E_s + dE_s$ corresponds to the range of values of ωt comprised in the interval.

$$\arccos \frac{E_s}{P_o} < \omega t < \arccos \frac{E_s + dE_s}{P_o} \quad (2)$$

*Published in *Acous. Soc. Amer. Jour.*, Jan., 1944.

The angular width of this interval is $(P_o^2 - E_s^2)^{-1/2} dE_s$. There are two such intervals in the range $0 < \omega t < 2\pi$. Values of E_s outside the range $-P_o$ to P_o do not exist. Hence, the probability that the signal voltage lies in the interval dE_s at any particular E_s is given by

$$\Phi_s(E_s)dE_s = \left\{ 0, |E_s| > P_o \right. \\ \left. 2(P_o^2 - E_s^2)^{-1/2} dE_s/2\pi, |E_s| < P_o \right\} dE_s \quad (3)$$

Random noise as discussed in this section may be characterized by the fact that the instantaneous amplitudes are normally distributed in time; that is, if $\Phi_n(z) dz$ is the probability that the noise amplitude lies in the amplitude interval of width dz at z ,

$$\Phi_n(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-z^2/2\sigma^2} \quad (4)$$

where σ is the root mean square noise amplitude. The mean noise power dissipated in unit resistance is given by $W_n = \sigma^2$. The corresponding mean signal power is given by $W_s = P_o^2/2$. Let $\Phi_r(z)$ represent the probability density function of the instantaneous sum of the signal and noise amplitudes. Then

$$\Phi_r(z)dz = dz \int_{-\infty}^{\infty} \Phi_s(\lambda) \Phi_n(z - \lambda)d\lambda \quad (5)$$

or

$$\Phi_r(z) = \frac{1}{\pi\sigma\sqrt{2\pi}} \int_{-P_o}^{P_o} \frac{e^{-(z-\lambda)^2/2\sigma^2} d\lambda}{\sqrt{P_o^2 - \lambda^2}} \quad (6)$$

By the substitution $\lambda = P_o \cos \theta$, we may convert the integral to the form

$$\Phi_r(z) = \frac{1}{\pi\sigma\sqrt{2\pi}} \int_0^\pi e^{-(z-P_o \cos \theta)^2/2\sigma^2} d\theta \quad (7)$$

Suppose we insert a half-wave linear rectifier in series with the source of signal and noise, so that the current I is given in terms of the resultant instantaneous voltage E by

$$I = \begin{cases} 0, & E < 0 \\ \alpha E, & E > 0 \end{cases} \quad (8)$$

Then the average value of current flowing in the circuit is

$$\bar{I} = \alpha \int_0^\infty z \Phi_r(z) dz \\ = \frac{\alpha}{\pi\sigma\sqrt{2\pi}} \int_0^\infty z dz \int_0^\pi e^{-(z-P_o \cos \theta)^2/2\sigma^2} d\theta \quad (9)$$

The value of this integral is shown in Appendix I to be

$$\bar{I} = \alpha \sqrt{\frac{W_n}{2\pi}} e^{-W_s/2W_n} \left\{ I_0(W_s/2W_n) + \frac{W_s}{W_n} \left[I_0\left(\frac{W_s}{2W_n}\right) + I_1\left(\frac{W_s}{2W_n}\right) \right] \right\} \quad (10)$$

This form is particularly convenient for calculation since Watson's Theory of Bessel Functions, Table II, gives $e^{-z}I_0(z)$ and $e^{-z}I_1(z)$ directly.

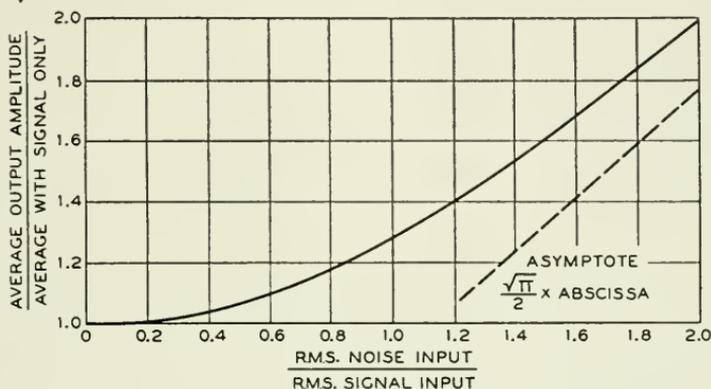


Fig. 1—Variation of direct-current component in response of linear rectifier with ratio of noise input to signal input.

Limiting forms of this equation may be expressed in terms of series in powers of W_s/W_n when the signal power is small compared with the noise power and in powers of W_n/W_s when the noise power is small compared with the signal power. The ascending series for small signal is:

$$\bar{I} = \alpha \sqrt{\frac{W_n}{2\pi}} \left[1 + \frac{1}{2(1!)^2} \frac{W_s}{W_n} + \frac{1(-1)}{2^2(2!)^2} \left(\frac{W_s}{W_n}\right)^2 + \frac{1(-1)(-3)}{2^3(3!)^2} \left(\frac{W_s}{W_n}\right)^3 + \dots \right] = \alpha \sqrt{\frac{W_n}{2\pi}} {}_1F_1\left(\frac{-1}{2}; 1; -\frac{W_s}{W_n}\right) \quad (11)$$

The asymptotic series, which is available for computation when the signal is large, is

$$\bar{I} \sim \frac{\alpha \sqrt{2W_s}}{\pi} \left[1 + \frac{(-1)^2 W_n}{1! 4W_s} + \frac{(-1)^2 \cdot 1^2}{2!} \frac{(W_n)^2}{(4W_s)} + \frac{(-1)^2 \cdot 1^2 \cdot 3^2}{3!} \frac{(W_n)^3}{(4W_s)} + \frac{(-1)^2 \cdot 1^2 \cdot 3^2 \cdot 5^2}{4!} \frac{(W_n)^4}{(4W_s)} + \dots \right] \quad (12)$$

Curves of \bar{I} have been plotted in three ways. Fig. 1 shows the ratio of \bar{I} to $\bar{I}_{so} = \alpha P_o/\pi$, the average current in the absence of noise, as a function

of ratio of rms noise input to rms signal input. Figure 2 shows the ratio of \bar{I} to $\bar{I}_{n_0} = \alpha\sigma/\sqrt{2\pi}$, the average current in the absence of signal, as a function of ratio of rms signal input to rms noise input. Figure 3 shows

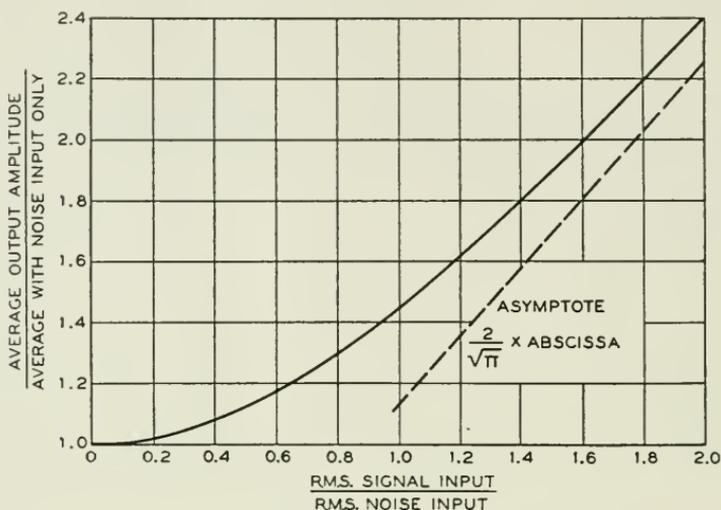


Fig. 2—Variation of direct-current component in response of linear rectifier with ratio of signal input to noise input.

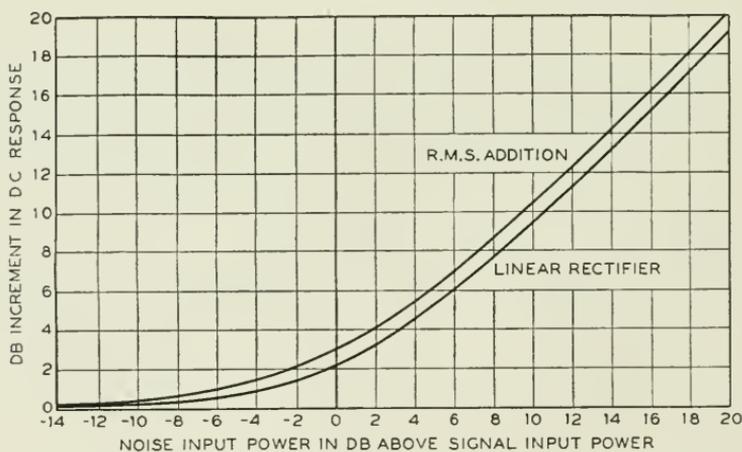


Fig. 3—Variation of direct-current component expressed in decibels, showing comparison between linear rectification and power addition of signal and noise.

the increment in d-c power output in decibels as varying amounts of noise expressed in decibels relative to the signal are added. The corresponding result for power addition is given for comparison.

II. SPECTRUM OF OUTPUT

A much more powerful method of attack on this problem is obtained by the use of multiple Fourier series. In this section we shall use Fourier analysis to obtain not only the direct-current output of the rectifier, but also the spectral distribution of the sinusoidal components in the output of the rectifier. We represent the input spectrum by

$$E = P_0 \cos p_0 t + \sum_{n=1}^N P_n \cos p_n t \quad (13)$$

This representation is more general than that given by (4) in that a frequency spectrum as well as an amplitude distribution is defined; it may be shown that the probability density for the sum of N sinusoidal waves with incommensurable frequencies approaches (4) when N is large. The first term represents the sinusoidal signal; the mean power which would be dissipated by this signal in unit resistance is

$$W_s = P_0^2/2. \quad (14)$$

The noise is represented by a large number N of sinusoidal components with incommensurable frequencies (or commensurable frequencies with random phase angles) distributed along the frequency range f_1 to f_2 in such a way that the mean noise power in band width Δf is:

$$w(f)\Delta f = \frac{1}{2} \sum_{n=\nu(f-f_1)}^{\nu(f+\Delta f-f_1)} P_n^2 = \nu \Delta f P^2(f)/2 \quad (15)$$

Here ν is the number of components per unit band width and $P(f)$ represents the amplitude of a component in the neighborhood of frequency f . Note also that the mean total noise input power, W_n , is given by

$$W_n = \int_0^\infty w(f) df = \frac{\nu}{2} \int_0^\infty P^2(f) df \quad (16)$$

The linear rectifier is specified by the current-voltage relationship (8), which is equivalent to

$$I = -\frac{\alpha}{2\pi} \int_C e^{iEz} \frac{dz}{z^2} \quad (17)$$

where C is an infinite contour going from $-\infty$ to $+\infty$ with an indentation below the pole at the origin. We may expand I in the multiple Fourier series¹

¹Bennett and Rice, "Note on Methods of Computing Modulation Products," *Phil. Mag.*, Sept. 1934. The present application represents an extension to N variables of the theory there given for two.

$$I = \sum_{m_0=0}^{\infty} \sum_{m_1=0}^{\infty} \cdots \sum_{m_N=0}^{\infty} a_{m_0 m_1 \cdots m_N} \cos m_0 x_0 \cos m_1 x_1 \cdots \cos m_N x_N \quad (18)$$

where

$$x_k = p_k t, \quad k = 0, 1, 2, \cdots N \quad (19)$$

$$a_{m_0 m_1 \cdots m_N} = \frac{\epsilon_{m_0} \epsilon_{m_1} \cdots \epsilon_{m_N}}{\pi^{N+1}} \int_0^\pi dx_0 \int_0^\pi dx_1 \cdots \int_0^\pi I \cos m_0 x_0 \cos m_1 x_1 \cdots \cos m_N x_N dx_N \quad (20)$$

$$\epsilon_j = \begin{cases} 2, & j \neq 0 \\ 1, & j = 0 \end{cases} \quad (21)$$

The response of the rectifier is thus seen to consist of all orders of modulation products of signal and noise. In a typical case of interest the band of input frequencies is relatively narrow and centered about a high frequency while the output band includes only low frequencies. In such a case the important components in the output are the beats between signal and noise components and between noise components. The *d-c* component is present in the output only if the pass band of the system actually includes zero frequency; we have already computed its value in Section I, but we will derive it again by the method used here as a check.

The amplitude of the *d-c* component is in fact:

$$a_{00 \cdots 0} = -\frac{\alpha}{2\pi} \int_c \frac{J_0(P_0 z) \prod_{n=1}^N J_0(P_n z)}{z^2} dz, \quad (22)$$

on substitution of the expression for *E* in the integral representation of *I*, substituting the result in (20) and interchanging the order of integration. When *N* is large, *P_n* is small, hence the principal contribution to the integral occurs near small values of *z*, where *J₀*(*P_nz*) is nearly equal to unity, since the product of a large number of factors, all less than unity, will be small indeed unless each factor is only slightly less than unity. We therefore replace *J₀*(*P_nz*) by a function which coincides with it near *z* = 0 and goes rapidly to zero as we depart from this region. Such an approximation (Laplace's process²) is

$$J_0(P_n z) \doteq e^{-P_n^2 z^2/4} \quad (23)$$

² Watson, "Theory of Bessel Functions," p. 421.

which is correct for the first two terms in the Taylor series expansion near $z = 0$. Therefore, when P_n approaches zero as N approaches infinity,

$$\begin{aligned} a_{00\dots 0} = \bar{I} &= -\frac{\alpha}{2\pi} \int_c J_0(P_0 z) e^{-\sum_{n=1}^N P_n^2 z^2/4} \frac{dz}{z^2} \\ &= -\frac{\alpha}{2\pi} \int_c J_0(P_0 z) e^{-W_n z^2/2} \frac{dz}{z^2} \end{aligned} \tag{24}$$

The contour integral cannot be replaced by a real integral directly because the integrand goes to infinity at the origin. However, since

$$\frac{J_0(u)}{u^2} = -\frac{J_1(u)}{u} - \frac{d}{du} \frac{J_0(u)}{u} \tag{25}$$

$$\frac{J_0(Pz)}{z^2} = -\frac{J_1(Pz)}{Pz^2} - \frac{d}{d(Pz)} \frac{J_0(Pz)}{Pz} = -\frac{J_1(Pz)}{P^2 z^2} - \frac{1}{P^2} \frac{d}{dz} \frac{J_0(Pz)}{z} \tag{26}$$

we can substitute (26) in the integral and perform an integration by parts to give the result.

$$\begin{aligned} \bar{I} &= \frac{\alpha}{\pi} \int_0^\infty e^{-W_n z^2/2} \left[\frac{P_0 J_1(P_0 z)}{z} + W_n J_0(P_0 z) \right] dz \\ &= \alpha \sqrt{\frac{W_n}{2\pi}} \left[{}_1F_1\left(\frac{1}{2}; 1; -\frac{W_s}{W_n}\right) + \frac{W_s}{W_n} {}_1F_1\left(\frac{1}{2}; 2; -\frac{W_s}{W_n}\right) \right] \end{aligned} \tag{27}$$

by Hankel's formula.³ But it may be shown that (see Appendix II)

$${}_1F_1\left(\frac{1}{2}; 1; -u\right) = e^{-u/2} I_0\left(\frac{u}{2}\right) \tag{28}$$

$${}_1F_1\left(\frac{1}{2}; 2; -u\right) = e^{-u/2} [I_0(u/2) - I_1(u/2)] \tag{29}$$

Hence,

$$\begin{aligned} \bar{I} &= \alpha \sqrt{\frac{W_n}{2\pi}} e^{-W_s/2W_n} \left\{ I_0(W_s/2W_n) + \frac{W_s}{W_n} \right. \\ &\quad \left. [I_0(W_s/2W_n) + I_1(W_s/2W_n)] \right\} \end{aligned} \tag{30}$$

which is identical with the result of Section I, noting that $\sigma = \sqrt{W_n}$. We point out that a resistance-capacity coupled amplifier will not pass this component since there is no transmission at zero frequency.

³ Watson, "Theory of Bessel Functions," p. 393. As pointed out by Watson, in a footnote, the difficulty with singularities at the origin could be avoided by expressing Hankel's formula in terms of a contour integral instead of an ordinary integral along the real axis. This procedure would lead directly to the hypergeometric function given in (11).

The amplitude of the typical difference product between the signal and the r th noise component is

$$A_{sn} = \frac{1}{2} a_{100\dots010\dots0} \\ = \frac{\alpha}{\pi} \int dz \frac{J_1(P_0 z) J_0(P_1 z) J_0(P_2 z) \cdots J_1(P_n z) \cdots J_0(P_N z)}{z^2} \quad (31)$$

Using the same process as before, we replace $J_1(P_n z)$ by

$$J_1(P_n z) \doteq \frac{P_n z}{2} e^{-P_n^2 z^2 / 8} \quad (32)$$

and obtain in the limit as N becomes indefinitely large

$$A_{sn} = \frac{\alpha P_n}{\pi} \int_0^\infty \frac{J_1(P_0 z)}{z} e^{-W_n z^2 / 2} dz \\ = \frac{\alpha P_n}{2} \sqrt{\frac{W_s}{\pi W_n}} {}_1F_1\left(\frac{1}{2}; 2; -\frac{W_s}{W_n}\right) \\ = \frac{\alpha P_n}{2} \sqrt{\frac{W_s}{\pi W_n}} e^{-W_s / 2 W_n} \left[I_0\left(\frac{W_s}{2 W_n}\right) + I_1\left(\frac{W_s}{2 W_n}\right) \right] \quad (33)$$

Relations between the ${}_1F_1$ function and Bessel functions are discussed in Appendix II.

The shape of the spectrum of the beats between P_0 and the noise input evidently consists of the superposition of the noise spectra above and below p_0 , so that if we write $w_{sn}(f) \Delta f$ for the mean energy from this source in that part of the filter output lying in the band of width Δf at f ,

$$w_{sn}(f) \Delta f = \frac{\nu \Delta f}{2} [(A_{sn}^+)^2 + (A_{sn}^-)^2] \quad (34)$$

$$A_{sn}^+ = [A_{sn}]_{p_n=p_0+2\pi f} \quad (35)$$

$$A_{sn}^- = [A_{sn}]_{p_n=p_0-2\pi f} \quad (36)$$

$$P_n = \sqrt{\frac{2w(f_n)}{\nu}} \quad (37)$$

$$w_{sn}(f) = \frac{\alpha^2 W_s}{4\pi W_n} e^{-W_s / W_n} \left[I_0\left(\frac{W_s}{2W_n}\right) + I_1\left(\frac{W_s}{2W_n}\right) \right]^2 \\ \times [w(f_0 + f) + w(f_0 - f)] \quad (38)$$

The total noise from this source in the output of a particular filter of transfer admittance $Y(f)$ is obtained by integrating $w_{sn}(f)Y(f)df$ throughout the band of the filter. In the particular case in which the original band of noise is

symmetrical about f_o and occupies the range $f_o - f_a$ to $f_o + f_a$ and an ideal low pass filter cutting off at $f = f_a$ is used in the rectifier output, the total noise output from beats between signal and noise is

$$W_{sn} = 2 \int_0^{f_a} w_{sn}(f) df = \frac{\alpha^2 w_s}{4\pi} e^{-w_s/w_n} [I_0(W_s/2W_n) + I_1(W_s/2W_n)]^2 \quad (39)$$

Next we shall calculate the spectrum of the energy resulting from beats between individual noise components. We write

$$\begin{aligned} A_{nn} &= \frac{1}{2} a_{00} \dots a_{10} \dots a_{10} \dots a_{00} \\ &= \frac{\alpha}{\pi} \int_C dz \frac{J_0(P_0 z) J_0(P_1 z) \dots J_1(P_r z) \dots J_1(P_s z) \dots J_0(P_N z)}{z^2} \\ &= \frac{\alpha P_r P_s}{2\pi} \int_0^\infty J_0(P_0 z) e^{-w_n z^2/2} dz \\ &= \frac{\alpha P_r P_s}{2\sqrt{2\pi W_n}} {}_1F_1 \left\{ \frac{1}{2}; 1; -\frac{W_s}{W_n} \right\} \\ &= \frac{\alpha P_r P_s}{2\sqrt{2\pi W_n}} e^{-W_s/2W_n} I_0(W_s/2W_n) \end{aligned} \quad (40)$$

To find the resulting spectrum $w_{nn}(f)df$ produced at f by the resultant of all such components, we note that we may sum over all components by beating each component of the primary band with the frequency f above it and adding the resultant power values. The result is

$$w_{nn}(f) = \frac{\alpha^2}{4\pi W_n} e^{-w_s/w_n} I_0^2(W_s/2W_n) \int_0^\infty w(\lambda) w(\lambda + f) d\lambda \quad (41)$$

In the particular case of a flat band of energy extending from f_1 to f_2 ,

$$\begin{aligned} \int_0^\infty w(\lambda) w(\lambda + f) d\lambda &= \int_{f_1}^{f_2-f} \frac{W_n^2}{(f_2 - f_1)^2} d\lambda = \frac{f_2 - f_1 - f}{(f_2 - f_1)^2} W_n^2, \\ &0 < f < f_2 - f_1 \end{aligned} \quad (42)$$

$$\begin{aligned} w_{nn}(f) &= \frac{\alpha^2 (f_2 - f_1 - f) W_n}{4\pi (f_2 - f_1)^2} e^{-w_s/w_n} I_0^2(W_s/2W_n), \\ &0 < f < f_2 - f_1 \end{aligned} \quad (43)$$

The total mean power of this type lying in the band 0 to f_b is

$$\begin{aligned} W_{nn}(f_b) &= \int_0^{f_b} w_{nn}(f) df = \frac{\alpha^2 W_n (f_2 - f_1 - f_b/2) f_b}{4\pi (f_2 - f_1)^2} e^{-w_s/w_n} \\ &I_0^2(W_s/2W_n) \end{aligned} \quad (44)$$

provided $f_b < f_2 - f_1$. The spectrum is confined to the region $0 < f < f_2 - f_1$. If f_b is equal to $f_2 - f_1$ so that the output filter passes all the noise of this type, we have

$$W_{nn}(f_2 - f_1) = W_{nn} = \frac{\alpha^2 W_n}{8\pi} e^{-W_s/W_n} I_0^2(W_s/2W_n) \quad (45)$$

This result seems to hold approximately for a considerable range of input spectra. For example, if we assume that the original noise is shaped like an error function about f_o , i.e.,

$$w_n(f) = W_n \sqrt{a/\pi} e^{-a(f-f_o)^2} \quad (46)$$

with f taken from $-\infty$ to $+\infty$ with small error for large f_o ,

$$\int_{-\infty}^{\infty} w(\lambda)w(\lambda + f) d\lambda = W_n^2 \sqrt{a/2\pi} e^{-af^2/2} \quad (47)$$

$$\int_0^{\infty} df \int_{-\infty}^{\infty} w(\lambda)w(\lambda + f) d\lambda = W_n^2/2 \quad (48)$$

which is in agreement with (45).

The output of a half-wave linear rectifier contains fundamental components and all even order modulation products. In general, the amplitudes of the higher order products are small compared with the lower order. In a particular problem some consideration of where the principal products fall in the frequency band is required. The products just considered give a fair approximation for the problem of detection of a radio frequency band of signal and noise followed by audio amplification. Certain other products should also be added to obtain higher accuracy. We have calculated the products of order zero and two; the next ones of importance are the fourth order, since the third order products vanish in a perfectly linear rectifier. The fourth order products in this case which fall in the audio band are of frequency $2p_o - p_r - p_s$, $p_o + p_q - p_r - p_s$, and $p_n + p_q - p_r - p_s$, where the subscripts n, q, r, s refer to the original noise component frequencies. The latter is, however, less important than the sixth order product $3p_o - p_q - p_r - p_s$, which involves only three noise components. Expressions for the contributions from these products are given in Appendix III.

Figure 4 shows computed curves for the noise produced in an audio band by the various components. Curve A is $W_{sn} + W_{nn}$ and includes what are usually regarded as the principal contributors, the difference frequencies between signal and noise, and between individual noise components. Curve B is obtained by adding to Curve A, the contribution from the fourth order products $2p_o - p_r - p_s$ and $p_o + p_q - p_r - p_s$ and the sixth order products $3p_o - p_q - p_r - p_s$. Thus all products which include three or less noise fundamental components are included. The curves are plotted in terms of

fraction of noise power received compared to the limiting noise when the mean signal input power is made indefinitely large compared to the mean input noise power. Some experimental points given by Williams⁴ are shown for comparison. Williams gives the intercept at zero signal power as 35%; the theoretical value deduced here is $\pi/8$ or 39.27%. It will be noted that the inclusion of the higher order products improves the agreement between experimental and theoretical curves, even though the value of the intercept is unaffected by them. It should also be stressed that our analysis applies

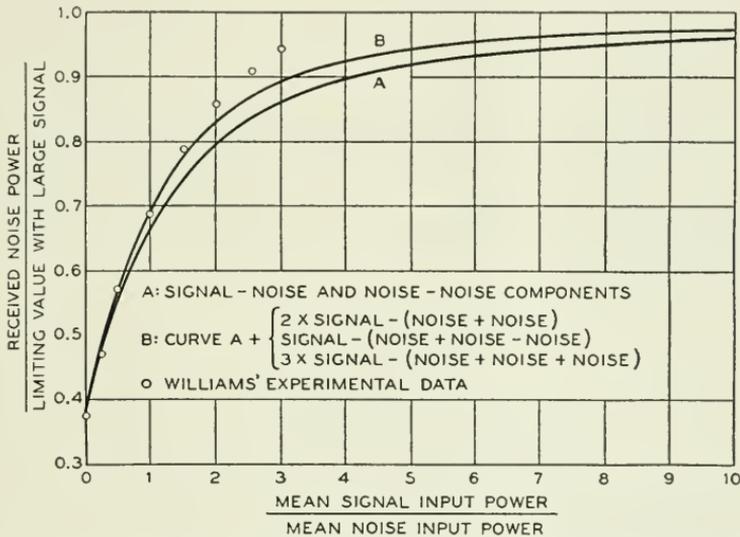


Fig. 4—Calculated noise power in audio band of output of linear rectifier when noise and signal are applied in a relatively narrow high-frequency band. The direct-current component is excluded.

strictly to purely resistive networks. The conventional radio detector circuit (which was used by Williams), in which a condenser is shunted across a resistance in series with a diode, departs from the conditions here assumed because of the reactive element, the condenser. The customary approximation made in treating this circuit is that the condenser has infinite impedance in the audio frequency range and zero impedance at the radio frequencies. This leads to a bias on the detector which depends on the signal. The methods given here may be applied, but the resulting formulas are much more difficult from the standpoint of numerical computation.

A recent paper by Ragazzini⁵ gives an approximate solution based on

⁴ F. C. Williams, "The Response of Rectifiers to Fluctuation Voltages," *Journal I. E. E.*, 1937, Vol. 80, pp. 218-226.

⁵ John Ragazzini, "The Effect of Fluctuation Voltages on the Linear Detector," *Proc. I. R. E.*, June 1942, Vol. 30, p. 277-288.

expanding the envelope of the input wave by the binomial theorem and retaining only the first two terms. The validity depends on the noise amplitude being small compared with the sum of signal and noise, and hence the result should agree with our solution in the neighborhood of $W_n/W_s = 0$, which it does. When W_s/W_n is small, the error is appreciable. Ragazzini's result (Equation 15 of the paper) expressed in our notation is

$$W_{sn} + W_{nn} \doteq \frac{\alpha^2 W_n (1 + \frac{1}{2} W_n/W_s)}{\pi^2 (1 + W_n/W_s)} \quad (49)$$

It will be seen by comparing the limiting values for $W_s/W_n = 0$ with that of $W_s/W_n = \infty$ from (49) that the intercept of the curve of Fig. 4 would be 50% instead of our value of 39.27%.

The results given in the present paper have been compiled from unpublished memoranda and notes by the author extending back as far as 1935. Discussions with colleagues have been of great aid, and in particular acknowledgment is made to Messrs. S. O. Rice and R. Clark Jones for many helpful suggestions.

APPENDIX I

EVALUATION OF INTEGRAL FOR I

Interchanging the order of integration in (9), we have

$$\bar{I} = \frac{\alpha}{\pi \sqrt{2\pi W_n}} \int_0^\pi d\theta \int_0^\infty e^{-(z - P_0 \cos \theta)^2 / 2W_n} z dz \quad (50)$$

By substituting $z = P_0 \cos \theta + u \sqrt{2W_n}$, we may evaluate the second integral in terms of the error function, obtaining

$$\begin{aligned} \bar{I} &= \frac{\alpha}{\pi^{3/2}} \int_0^\pi d\theta \int_{-P \cos \theta / \sqrt{2W_n}}^\infty e^{-u^2} (u \sqrt{2W_n} + P_0 \cos \theta) du \\ &= \frac{\alpha}{\pi} \frac{\sqrt{W_n}}{2\pi} \int_0^\pi e^{-P_0^2 \cos^2 \theta / 2W_n} d\theta \\ &\quad + \frac{\alpha P_0}{2\pi} \int_0^\pi \operatorname{erf} (P_0 \cos \theta / \sqrt{2W_n}) \cos \theta d\theta \\ &= \frac{\alpha}{\pi} \frac{\sqrt{W_n}}{2\pi} e^{-P_0^2 / 4W_n} \int_0^\pi e^{-\cos^2 2\theta / 4W_n} d\theta \end{aligned}$$

$$\begin{aligned}
& + \frac{\alpha P_0}{2\pi} \int_0^\pi \frac{d}{d\theta} \left[\operatorname{erf} \left(\frac{P_0 \cos \theta}{\sqrt{2W_n}} \right) \sin \theta \right] d\theta - \frac{\alpha P_0}{2\pi} \int_0^\pi \sin \theta \\
& \quad \frac{d}{d\theta} \left(\operatorname{erf} \frac{P_0 \cos \theta}{\sqrt{2W_n}} \right) d\theta \\
& = \frac{\alpha}{2\pi} \frac{\sqrt{W_n}}{2\pi} e^{-W_s/2W_n} \int_0^{2\pi} e^{-W_s \cos \Phi/2W_n} d\Phi \\
& + \frac{\alpha W_s}{2\pi \sqrt{2\pi W_n}} \int_0^{2\pi} e^{-W_s \cos \Phi/2W_n} (1 - \cos \Phi) d\Phi \\
& = \alpha \sqrt{\frac{W_n}{2\pi}} e^{-W_s/2W_n} \left(I_0(W_s/2W_n) \right. \\
& \left. + \frac{W_s}{W_n} [I_0(W_s/2W_n) + I_1(W_s/2W_n)] \right) \quad (10)
\end{aligned}$$

In the above we have made use of the relations:

$$\operatorname{erf} z = \frac{2}{\sqrt{\pi}} \int_0^z e^{-z^2} dz \quad (51)$$

$$\frac{d}{dz} \operatorname{erf} z = \frac{2}{\sqrt{\pi}} e^{-z^2} \quad (52)$$

$$\int_0^{2\pi} e^{-z \cos \Phi} \cos m\Phi d\Phi = (-)^m 2\pi I_m(z) \quad (53)$$

APPENDIX II

RELATIONS BETWEEN HYPERGEOMETRIC AND BESSEL FUNCTIONS

The modulation coefficients appearing in the linear rectification of noise are expressible in compact form in terms of the hypergeometric function:

$$\begin{aligned}
{}_1F_1(a; c; -z) & = 1 - \frac{a}{c} \frac{z}{1!} + \frac{a(a+1)}{c(c+1)} \frac{z^2}{2!} - \dots \\
& = \frac{\Gamma(c)}{\Gamma(a)} \sum_{m=0}^{\infty} \frac{\Gamma(a+m)}{\Gamma(c+m)m!} (-z)^m \quad (54)
\end{aligned}$$

The ${}_1F_1$ function is a limiting case of the more familiar Gaussian hypergeometric function ${}_2F_1(a, b; c; z)$, viz.

$${}_1F_1(a; c; z) = \lim_{b \rightarrow \infty} {}_2F_1(a, b; c; z/b) \quad (55)$$

In certain special cases this function may be expressed in terms of exponential and Bessel functions. For example, by a formula given by

Campbell and Foster, *Fourier Integrals for Practical Application*, Bell System Monograph B-584, p. 32 (also Watson, *Theory of Bessel Functions*, p. 191), we may show that

$${}_1F_1\left(\nu + \frac{1}{2}; 2\nu + 1; -z\right) = \frac{2^{2\nu}\Gamma(\nu + 1)e^{-z/2}}{(-z)^\nu} I_\nu(-z/2) \quad (56)$$

or setting $\nu = 0$

$${}_1F_1\left(\frac{1}{2}; 1; -z\right) = e^{-z/2} I_0(z/2) \quad (57)$$

which is one of the functions appearing in our work.

We have also encountered the function ${}_1F_1(1/2; 2; -z)$ which is not directly reducible by the above formula. The reduction may be effected in a number of ways. By making use of the relation obtained from (56) by setting $\nu = 1$,

$${}_1F_1(3/2; 3; -z) = \frac{4}{z} e^{-z/2} I_1(z/2) \quad (58)$$

and noting that

$$\begin{aligned} & {}_1F_1(1/2; 2; -z) - {}_1F_1(1/2; 1; -z) \\ &= \frac{1}{\Gamma(1/2)} \sum_{m=0}^{\infty} \frac{\Gamma(m+1/2)}{m!(m+1)!} (-z)^m - \frac{1}{\Gamma(1/2)} \sum_{m=0}^{\infty} \frac{\Gamma(m+1/2)}{(m!)^2} (-z)^m \\ &= \frac{-1}{\Gamma(1/2)} \sum_{m=0}^{\infty} \frac{\Gamma(m+1/2)m}{m!(m+1)!} (-z)^m \\ &= \frac{z}{\Gamma(1/2)} \sum_{m=0}^{\infty} \frac{\Gamma(m+3/2)}{(m+2)!m!} (-z)^m \\ &= \frac{z}{4} {}_1F_1(3/2; 3; -z), \end{aligned} \quad (59)$$

we find that⁶

$${}_1F_1(1/2; 2; -z) = e^{-z/2} [I_0(z/2) + I_1(z/2)] \quad (60)$$

It may also be verified by integrating the series directly that

$$\int_0^z {}_1F_1(1/2; 1; -z) dz = z {}_1F_1(1/2; 2; -z) \quad (61)$$

Combining this relation with (57) and (60) above, we deduce the indefinite integrals

⁶ The relation (60) was brought to the attention of the author by Mr. R. M. Foster.

$$\left. \begin{aligned} \int e^x I_0(x) dx &= xe^x[I_0(x) - I_1(x)] \\ \int e^{-x} I_0(x) dx &= xe^{-x}[I_0(x) + I_1(x)] \\ \int e^x I_1(x) dx &= e^x[(1-x)I_0(x) + xI_1(x)] \\ \int e^{-x} I_1(x) dx &= e^{-x}[(1+x)I_0(x) + xI_1(x)] \end{aligned} \right\} \quad (62)$$

These integrals may be derived by differentiating the right hand members, and could, therefore, serve as a basis for an alternate derivation of (60).

In addition it was noted in Eq. (11) that the constant term in the modulation spectrum could be expressed in terms of ${}_1F_1(-1/2; 1; -z)$; from the equations given, it follows that we must have the relation:

$${}_1F_1(-1/2; 1; -z) = e^{-z/2} [(1+z)I_0(z/2) + zI_1(z/2)] \quad (63)$$

Another interesting set of formulas which can be obtained as a by-product from (62) by setting $x = iy$ is:

$$\left. \begin{aligned} \int J_0(y) \cos y dy &= y[J_0(y) \cos y + J_1(y) \sin y] \\ \int J_0(y) \sin y dy &= y[J_0(y) \sin y - J_1(y) \cos y] \\ \int J_1(y) \cos y dy &= yJ_1(y) \cos y - J_0(y)(y \sin y - \cos y) \\ \int J_1(y) \sin y dy &= yJ_1(y) \sin y + J_0(y)(y \cos y - \sin y) \end{aligned} \right\} \quad (64)$$

The hypergeometric notation is particularly convenient in determining series expansions for the coefficients to be used for calculation when the variable z is either very small or very large. For small values of z , the form (54) suffices; for large values of z , we may use the general asymptotic expansion formula⁷ for the real part of z positive:

$$\begin{aligned} {}_1F_1(a; c; -z) &= \frac{\Gamma(c)}{\Gamma(c-a)z^a} {}_2F_0(a, 1+a-c; 1/z) \\ &= \frac{\Gamma(c-a)z^a}{\Gamma(c)} \left[1 + \frac{a(1+a-c)}{1!z} \right. \\ &\quad \left. + \frac{a(a+1)(1+a-c)(2+a-c)}{2!z^2} + \dots \right] \end{aligned} \quad (65)$$

The series expansions required here could also be obtained from the appropriate series for Bessel functions. It will be noted, however, that the typical modulation coefficient can be expressed in terms of either a single ${}_1F_1$ function or several Bessel functions, so that manipulations must be performed on the series for the latter to give the final result. The Bessel functions on the other hand are more convenient for numerical computations because of the excellent tables available.

Reduction formulas for certain other hypergeometric functions are needed in evaluating the higher order products. They are:

$${}_1F_1(3/2; 1; -z) = e^{-z/2} [(1-z)I_0(z/2) + I_1(z/2)] \quad (66)$$

$${}_1F_1(3/2; 2; -z) = e^{-z/2} [I_0(z/2) - I_1(z/2)] \quad (67)$$

$${}_1F_1(5/2; 4; -z) = \frac{4}{z} e^{-z/2} \left[\left(\frac{4}{z} + 1 \right) I_1(z/2) - I_0(z/2) \right] \quad (68)$$

Derivation of these is facilitated by the use of the easily demonstrated relations:

$${}_1F_1(a; 1; -z) = \frac{d}{dz} [z {}_1F_1(a; 2; -z)] \quad (69)$$

$$2z {}_1F_1(a; 2; -z) = \frac{d}{dz} [z^2 {}_1F_1(a; 3; -z)] \quad (70)$$

$${}_1F_1(3/2; 3; -z) - {}_1F_1(3/2; 2; -z) = \frac{z}{4} {}_1F_1(5/2; 4; -z) \quad (71)$$

APPENDIX III

HIGHER ORDER PRODUCTS

The methods described in Section II may be applied to calculate the general expression for the general modulation coefficient. The result is for the amplitude of the term $\cos m\phi_0 t \cos p_{n_1} t \cos p_{n_2} t \cdots \cos p_{n_M} t$:

$$a_{m,M} = \frac{(-)^{\frac{m+M}{2}+1} P_{n_1} P_{n_2} \cdots P_{n_M} \Gamma \left(\frac{m+M-1}{2} \right) (W_s)^{m/2}}{\pi (W_n/2)^{(M-1)/2} m!} \times {}_1F_1 \left(\frac{m+M-1}{2}; m+1; \frac{-W_s}{W_n} \right) \quad (72)$$

The coefficient of the term $\cos (m\phi_0 \pm p_{n_1} \pm p_{n_2} \pm \cdots p_{n_M}) t$ is $a_{m,M}$ divided by $2^{M-1} \epsilon_m$. The number of terms of a particular type falling in a particular frequency interval can be calculated by a method previously described by

the author.⁸ Under the assumed conditions that the original noise spectrum is either flat throughout a limited range, or falls off like an error function, and that the audio amplifier passes all the difference components in question, we find the following results:

$$2p_0 - p_r - p_s : \\ W_{2s,nn} = \frac{\alpha^2 W_n}{8\pi} e^{-W_s/W_n} I_1^2(W_s/2W_n) \quad (73)$$

$$p_0 + p_q - p_r - p_s : \\ W_{sn,nn} = \frac{\alpha^2 W_s}{32\pi} e^{-W_s/W_n} [I_0(W_s/2W_n) - I_1(W_s/2W_n)]^2 \quad (74)$$

$$3p_0 - p_q - p_r - p_s : \\ W_{3s,nnn} = \frac{\alpha^2 W_s}{32\pi} e^{-W_s/W_n} [(1 + 4W_n/W_s)I_1(W_s/2W_n) \\ - I_0(W_s/2W_n)]^2 \quad (75)$$

This includes all beats containing not more than three noise fundamentals. The reductions of hypergeometric functions to exponential and Bessel functions given in Appendix II have been used in deriving the above results.

⁸ Bennett, "Cross-Modulation in Multichannel Amplifiers," *Bell Sys. Tech. Jour.*, Oct. 1940, Vol. XIX, pp. 587-610.

Dielectric Constants and Power Factors at Centimeter Wave-Lengths

By CARL R. ENGLUND

The theory underlying the measurement of dielectric constants and power factors, by means of resonant lengths of coaxial transmission line, is developed, apparatus used for such measurements is illustrated and the measurement routine described. A table of typical results is appended together with an "X tan X" table for aid in the calculations.

INTRODUCTION

THERE are two instrumentalities available for measuring dielectric constants and power factors at centimeter wave-lengths. These are, coaxial conductor lines and wave guides. Which one is, for any condition, the more favorable one depends a great deal upon the wave-lengths used. Under the conditions encountered in this work the coaxial line appeared to have the practical superiority, down to something like 10 cms. wave-length, anyway. Below this, the wave guide is very manageable and has several advantageous features.

When this work was begun, the most easily available and practicable vacuum tube which would oscillate around 20 cms. wave-length was the W. E. Co. 368A. This could be pushed down to something below 19 cms. wave-length but was undependable there and as a practical compromise 22.5 cms. wave-length was finally chosen. Later another tube became available and as it could be operated down to at least 9 cms. it was used in the more recent work. Thus, while the bulk of the measurements made were at 22.5 cms. wave-length, a good share of the samples investigated were also measured at approximately 10 cms. wave-length.

Any measurements made at these wave-lengths must be made in the form of transmission line measurements and the dielectric must be physically part of the coaxial line. There are various transmission line quantities definable and measurable, such as series impedance per unit length, shunt admittance per unit length, surge impedance, impedance transformation factor, voltage and current step-up factors, resonance selectivity or "Q", etc. The first two are measurable directly only at long wave-lengths, the last two are properties of space resonant line elements. Of these the "Q" was the most advantageous in the present instance.

"Q" DEFINITION

At low frequencies the resonance selectivity factor of lumped circuits is identified as the "Q" and is defined as $\frac{\omega L}{R}$. It is measured by a detuning process. For a length of transmission line with negligible shunt conductance losses this process gives $\frac{\omega L}{R}$ as for a coil; when this process is applied to complex circuits the physical embodiment of the "Q" becomes difficult to realize and it is preferable to define the "Q" in terms of the detuning process itself. This is equally true for the resonant, centimeter wave, line element and we proceed as follows: For this element some current or voltage amplitude, conveniently measurable, is selected and three values of it are measured as the line tuning is varied. This variation may be either in generator frequency for constant line length or in line length for constant generator frequency.

Thus, for example,

$$\left. \begin{aligned} Q &= \frac{f_0}{f_2 - f_1}, \quad \text{where } f_2 > f_0 > f_1 \\ Q &= \frac{\ell_0}{\ell_2 - \ell_1}, \quad \text{where } \ell_2 > \ell_0 > \ell_1 \end{aligned} \right\} A_2^2 = A_1^2 = \frac{A_0^2}{2} \quad (1)$$

with A_0 as the resonant amplitude. For low-loss lines the two definitions will give the same results in practice. Neither is ideal for second order accuracy since there is a variation of line constants with frequency in the first and a variation in total attenuation in the second.

For practical reasons it is usually preferable to excite and observe the line resonance in terms of the current at one end, this end shorted. The elementary line lengths are then the quarter and the half-wave ones, the former with open circuit far end, the latter with shorted far end. The latter is the more nearly ideal unit. In order to short effectively the input end, the input and output couplings must be made as loose as possible. As these couplings are reduced the observed "Q" will asymptotically approach the line "Q". At the present moment the line variation in length is the most convenient process, the chief trouble being the micrometric measurement of the tiny length changes involved. Thus for 10 cms wave-length and a half-wave coaxial line, a "Q" of 1000 involves a plunger movement of .0019 inches.

THEORY OF MEASUREMENT

It is shown in the appendix that the "Q" of a given resonant line segment can be broken up into parts representing the equivalent "Q's" of the terminal impedances and the line itself. Thus

$$\frac{1}{Q} = \frac{1}{Q_g} + \frac{1}{Q_0} + \frac{1}{Q_\ell} \quad (2)$$

where "Q" is the actually measured quantity, Q_g is the part due to the line itself, Q_0 and Q_ℓ the parts due to the near and far end terminations, respectively.

If we now take a quarter-wave line segment, with near end shorted through a movable plunger and far end open, we may make two "Q" measurements without and with the far end loaded with a dielectric segment, and obtain

$$\left\{ \begin{array}{l} \frac{1}{Q'} = \frac{1}{Q_g} + \frac{1}{Q_0} = \frac{d'}{\lambda/4} \\ \frac{1}{Q} = \frac{1}{Q_g} + \frac{1}{Q_0} + \frac{1}{Q_\ell} = \frac{d}{\lambda/4} \end{array} \right\} \quad (3)$$

and $\frac{1}{Q} - \frac{1}{Q'} = \frac{1}{Q_\ell} = \frac{d - d'}{\lambda/4} \left\{ \begin{array}{l} d' = \ell_2 - \ell_1 \\ d = \ell_2 - \ell_1 \end{array} \right.$

with d' and d equal to the widths of the resonance curves halfway down in power. These two d 's are, of course, directly measurable.

When the line is loaded with a dielectric segment the loaded part of the line can be represented as an impedance Z_ℓ connected to the unloaded remainder of the line. The effect of the loaded segment upon the unloaded

line (See appendix, eq. 4) appears in the form $\frac{\sqrt{\bar{Z}}}{Z_\ell}$ where $\sqrt{\bar{Z}}$ is the surge impedance of the unloaded line, with "Z" and "D" the series impedance and shunt admittance, respectively, for unit length of this line. If we put

$$\tanh \theta = \tanh (a_\ell + ib_\ell) = \frac{\sqrt{\bar{Z}}}{Z_\ell} \quad (4)$$

we have

$$\left\{ \begin{array}{l} \frac{1}{Q_\ell} = \frac{d - d'}{\lambda/4} = \frac{4a_\ell}{\pi} \\ \Delta \ell + t = \frac{\lambda}{2\pi} b_\ell \end{array} \right. \quad (5)$$

where $\Delta \ell$ is the measured plunger movement necessary to retune the line, after adding the dielectric loading, and "t" is the length of the dielectric segment.

Now, the power factor of " Z_ℓ " is the same as that of $\frac{\sqrt{Z}}{Z_\ell}$, as long as $\sqrt{\frac{Z}{D}}$ is substantially a resistance, and since

$$\tanh (a_\ell + i b_\ell) = \frac{\sinh 2a_\ell + i \sin 2b_\ell}{\cosh 2a_\ell + \cos 2b_\ell}, \quad (6)$$

we have

$$\text{power factor } Z_\ell = \text{p.f.} = \frac{\sinh 2a_\ell}{\sin 2b_\ell}. \quad (7)$$

Substituting eq. (5) in (7),

$$\text{p.f.} = \frac{\sinh \frac{2\pi}{\lambda} (d - d')}{\sin \frac{4\pi}{\lambda} (\Delta\ell + t)}, \quad (8)$$

which is the power factor of the loaded line segment in terms only of measurable lengths.

This does not complete the theory, however. We are interested in the power factor of the dielectric itself and it is evident that except for very short dielectric segments, the variation of the standing electrical field along the dielectric segment will result in a calculated power factor smaller than the true one. We also wish to determine the dielectric constant.

The impedance of the dielectric line segment, open circuited at the far end, can be written as

$$Z_\ell = \frac{\sqrt{\frac{L}{\epsilon C}}}{\tanh \left(\alpha + i \frac{2\pi\sqrt{\epsilon}}{\lambda} \right) t} \quad (9)$$

where " α " is the attenuation per unit length and " ϵ " is the dielectric constant.

Hence $\tanh (a_\ell + i b_\ell) = \sqrt{\epsilon} \tanh \left(\alpha + i \frac{2\pi\sqrt{\epsilon}}{\lambda} \right) t$ and

$$\text{p.f.} = \frac{\sinh 2\alpha t}{\sin \frac{4\pi\sqrt{\epsilon}}{\lambda} t} \quad (10)$$

an alternative expression. Now when "t" is very small the functions of the angles become equal to the angles and we write, for the dielectric power factor itself

$$\text{P.F.} = \frac{2\alpha t}{\frac{4\pi\sqrt{\epsilon}t}{\lambda}} \quad (11)$$

Dividing this expression by eq. (10)

$$\text{P.F.} = \text{p.f.} \cdot \frac{\sin \frac{4\pi\sqrt{\epsilon}t}{\lambda}}{\frac{4\pi\sqrt{\epsilon}t}{\lambda}} \cdot \frac{2\alpha t}{\sinh 2\alpha t}$$

and as the last term is always very nearly unity we have, if we put $\frac{4\pi\sqrt{\epsilon}t}{\lambda} = 4X$,

$$\text{P.F.} = \frac{\sinh \frac{2\pi}{\lambda} (d - d')}{\sin \frac{4\pi}{\lambda} (\Delta\ell + t)} \cdot \frac{\sin 4X}{4X} \quad (12)$$

Ordinarily the "sinh" is very closely equal to the angle.

The reactance of the dielectric segment of line is necessarily equal to the reactance of the part of the original line which it displaces, since space resonance occurs in both cases. Hence,

$$\tan \pi \frac{\Delta\ell + t}{\lambda} = \sqrt{\epsilon} \tan \pi \frac{\sqrt{\epsilon}t}{\lambda} \quad (13)$$

which we can rewrite to

$$\frac{\pi t}{\lambda} \cdot \tan \pi \frac{\Delta\ell + t}{\lambda} = \pi \frac{\sqrt{\epsilon}t}{\lambda} \cdot \tan \pi \frac{\sqrt{\epsilon}t}{\lambda}$$

Putting

$$\begin{cases} y = \frac{\pi t}{\lambda} \tan \pi \frac{\Delta\ell + t}{\lambda} \\ X = \frac{\pi\sqrt{\epsilon}t}{\lambda} \end{cases} \quad \text{we have} \quad y = X \tan X. \quad (14)$$

"y" is directly determinable by measurement and this gives X from the X tan X table supplied.¹ The value of $\epsilon = \left[\frac{X}{\frac{\pi t}{\lambda}} \right]^2$ follows and P.F. is immediately calculable. This completes the reduction of the observation.

¹ As no X tan X table to the necessary subdivision was available, one was calculated from the Hayashi tan X tables.

X TAN X

X	0	1	2	3	4	5	6	7	8	9
00	.0000 0000	.0000 0100	.0000 0400	.0000 0900	.0000 1600	.0000 2500	.0000 3600	.0000 4900	.0000 6400	.0000 8100
01	.0001 0000	.0001 2100	.0001 4401	.0001 6901	.0001 9601	.0002 2502	.0002 5602	.0002 8903	.0003 2403	.0003 6104
02	.0004 0005	.0004 4106	.0004 8408	.0005 2909	.0005 7611	.0006 2512	.0006 7613	.0007 2914	.0008 8415	.0009 4116
03	.0009 0027	.0009 6131	.0010 2435	.0010 8940	.0011 5645	.0012 2550	.0012 9656	.0013 6963	.0014 4870	.0016 2177
04	.0016 0085	.0016 8194	.0017 6504	.0018 5104	.0019 3745	.0020 2637	.0021 1749	.0022 1063	.0023 6577	.0024 0292
05	.0025 0209	.0026 0326	.0027 0644	.0028 1163	.0029 1884	.0030 2806	.0031 3928	.0032 6252	.0033 6778	.0034 8504
06	.0036 0433	.0037 2562	.0038 4893	.0039 7426	.0041 0160	.0042 3096	.0043 6234	.0044 9673	.0046 3114	.0047 6857
07	.0049 0802	.0050 4949	.0051 8298	.0053 3849	.0054 8602	.0056 3557	.0057 8716	.0059 4075	.0060 9637	.0062 5402
08	.0064 1369	.0065 7539	.0067 3911	.0069 0486	.0070 7264	.0072 4429	.0074 2185	.0077 1438	.0079 8492	.0081 4332
09	.0081 2194	.0083 0599	.0084 8796	.0086 7402	.0088 6212	.0090 5225	.0092 4442	.0094 3852	.0096 3486	.0098 3315
10	.0100 3347	.0102 3583	.0104 4023	.0106 4668	.0108 5516	.0110 6570	.0112 7827	.0114 9289	.0117 0956	.0119 2828
11	.0121 4904	.0123 7185	.0125 9672	.0128 2363	.0130 5259	.0132 8361	.0136 1658	.0137 5181	.0139 8999	.0142 2823
12	.0144 6952	.0147 1287	.0149 5829	.0152 0576	.0154 5529	.0157 0689	.0159 6055	.0162 1628	.0164 7407	.0167 3493
13	.0169 9585	.0172 5984	.0175 2591	.0177 9404	.0180 6425	.0183 3653	.0186 1098	.0188 8731	.0191 6582	.0194 4640
14	.0197 2907	.0200 1381	.0203 0063	.0205 8954	.0208 8053	.0211 7360	.0214 6876	.0217 6501	.0220 6334	.0223 6279
15	.0226 7028	.0229 7589	.0232 8359	.0235 9339	.0239 0528	.0242 1827	.0245 3535	.0248 5354	.0251 7289	.0254 9619
16	.0258 2071	.0261 4731	.0264 7602	.0268 0683	.0271 3975	.0274 7479	.0278 1193	.0281 5119	.0284 9256	.0288 3605
17	.0291 8166	.0295 2939	.0298 7923	.0302 3120	.0305 8529	.0309 4161	.0312 9985	.0316 6032	.0320 2292	.0323 8765
18	.0327 5452	.0331 2311	.0334 9464	.0338 6791	.0342 4332	.0346 2087	.0350 0056	.0353 8239	.0357 6637	.0361 5250
19	.0366 4077	.0369 3159	.0373 2377	.0377 1849	.0381 1537	.0385 1441	.0389 1561	.0393 1896	.0397 2448	.0401 3216
20	.0405 4201	.0409 5462	.0413 6820	.0417 8455	.0422 0307	.0426 2377	.0430 4664	.0434 7169	.0438 9892	.0443 2832
21	.0447 5991	.0451 9309	.0456 2965	.0460 6780	.0465 0814	.0469 5067	.0473 9540	.0478 4232	.0482 9144	.0487 4275
22	.0491 9627	.0496 5204	.0501 0992	.0505 7006	.0510 3240	.0514 9696	.0519 6373	.0524 3271	.0529 0391	.0533 7733
23	.0538 5297	.0543 3084	.0548 1093	.0552 9325	.0557 7779	.0562 6457	.0567 5358	.0572 4483	.0577 3831	.0582 3404
24	.0587 3201	.0592 3222	.0597 3468	.0602 3939	.0607 4634	.0612 5555	.0617 6702	.0622 8074	.0627 9673	.0633 1497
25	.0638 3548	.0643 5826	.0648 8330	.0654 1061	.0659 4020	.0664 7207	.0670 10621	.0675 4263	.0680 8133	.0686 2232
26	.0691 6560	.0697 1117	.0702 5903	.0708 0918	.0713 6163	.0719 1638	.0724 7343	.0730 3275	.0735 9444	.0741 5382
27	.0747 2470	.0752 9329	.0758 6421	.0764 3744	.0770 1299	.0776 9087	.0781 7107	.0787 5361	.0793 3844	.0799 2567
28	.0805 1521	.0811 0703	.0817 0131	.0822 9787	.0828 9679	.0834 9805	.0840 0166	.0847 0763	.0853 1596	.0859 2665
29	.0866 3971	.0871 5519	.0877 7292	.0883 9308	.0889 1562	.0896 4064	.0902 6783	.0908 9751	.0915 2957	.0921 6403
30	.0928 0088	.0934 4012	.0940 8176	.0947 2579	.0953 7224	.0960 2109	.0966 7234	.0973 2601	.0979 8210	.0986 4060
31	.0993 0153	.0999 6488	.1006 3066	.1012 9886	.1019 6950	.1026 4257	.1033 1809	.1039 9605	.1046 7646	.1053 6930
32	.1060 4461	.1067 3227	.1074 2367	.1081 1841	.1088 1641	.1095 1802	.1102 2310	.1109 3169	.1116 4369	.1123 5921
33	.1130 3321	.1137 4667	.1144 6587	.1151 9184	.1158 9811	.1166 2057	.1173 4564	.1180 7303	.1187 0322	.1195 3521
34	.1202 7054	.1210 0808	.1217 4815	.1224 9074	.1232 3587	.1239 8353	.1247 3373	.1254 8647	.1262 4175	.1269 9959
35	.1277 5997	.1285 2292	.1292 8842	.1300 5649	.1308 2712	.1316 0032	.1323 7610	.1331 5446	.1339 3539	.1347 1891
36	.1355 0503	.1362 9373	.1370 8503	.1378 7893	.1386 7544	.1394 7455	.1402 7628	.1410 8062	.1418 8758	.1426 9716
37	.1435 0937	.1443 2421	.1451 4169	.1459 6189	.1467 8456	.1476 0977	.1484 3802	.1492 6873	.1501 0210	.1509 3813
38	.1517 7683	.1525 1821	.1534 6325	.1543 0988	.1551 5838	.1560 1048	.1568 6527	.1577 2275	.1585 8293	.1594 4582
39	.1603 1142	.1611 7973	.1620 5075	.1629 2450	.1638 0097	.1646 8017	.1655 6211	.1664 4679	.1673 3421	.1682 2437
40	.1691 1729	.1700 1296	.1709 1140	.1718 1259	.1727 1657	.1736 2331	.1745 3283	.1754 4514	.1763 6023	.1772 7811
41	.1781 9879	.1791 2228	.1800 4887	.1809 7767	.1819 0958	.1828 4432	.1837 8188	.1847 2227	.1856 6570	.1866 1156
42	.1876 6047	.1885 1223	.1894 6684	.1904 2430	.1913 8464	.1923 4784	.1933 1391	.1942 8287	.1952 5450	.1962 2942
43	.1972 0704	.1981 8765	.1991 7097	.2001 5730	.2011 4658	.2021 3869	.2031 3377	.2041 3178	.2051 3272	.2061 3661
44	.2071 4343	.2081 6321	.2091 6594	.2101 8162	.2112 0024	.2122 2191	.2132 4651	.2142 7409	.2152 0466	.2163 3822
45	.2173 7478	.2184 1434	.2194 6691	.2205 0249	.2215 6110	.2226 0272	.2236 5738	.2247 1508	.2257 7581	.2268 3960
46	.2279 0643	.2289 7633	.2300 4929	.2311 2532	.2322 0442	.2332 8661	.2343 7189	.2354 6025	.2365 5172	.2376 4629
47	.2387 4397	.2398 4477	.2409 4969	.2420 5574	.2431 6193	.2442 7925	.2453 9573	.2465 1536	.2476 3813	.2487 6409
48	.2498 9320	.2510 1050	.2521 6099	.2532 9966	.2544 4152	.2555 8659	.2567 3487	.2578 8636	.2590 4107	.2601 9902
49	.2613 6019	.2626 2462	.2636 9227	.2648 6319	.2660 3736	.2672 1481	.2683 9607	.2695 7952	.2707 6680	.2719 5737
50	.2731 5125	.2743 4843	.2755 4892	.2767 5273	.2779 5987	.2791 6529	.2803 9416	.2816 0132	.2828 2183	.2840 4670
51	.2862 7296	.2865 0366	.2877 3756	.2889 7495	.2902 1573	.2914 5922	.2927 0752	.2939 5853	.2962 1297	.2964 7084

The above theory applies to the quarter wave line. This is a rather difficult practical one; it is best to add another quarter wave to make a half-wave resonator, shorted at both ends, with the dielectric positioned exactly in the center. From conditions of symmetry we then employ the above equations, taking half of our measured quantities. Or, in terms of the actually measured four lengths which constitute an observation on a half-wave line, $(d-d')$, t , $\Delta\ell$ and λ , we have,

$$\left. \begin{aligned} \text{P.F.} &= \frac{\sinh \frac{\pi}{\lambda} (d - d')}{\sin \frac{2\pi}{\lambda} (\Delta\ell + t)} \cdot \frac{\sin 2X}{2X} \\ \frac{\pi t}{\lambda} \cdot \tan \pi \frac{\Delta\ell + t}{\lambda} &= X \tan X \\ \epsilon &= \left[\frac{X}{\frac{\pi t}{\lambda}} \right]^2 \end{aligned} \right\} \quad (15)$$

which are the expressions used in this work.

In practice the dielectric plug is pushed into the half-wave line and the line is tuned. The line center is then calculated and the plug reset to this. Retuning checks the correct location. Two trials are always sufficient if the plug was nearly centered originally.

There are several shortcomings affecting this theory. The Q of the unloaded line depends partly on metal power loss along the line. When the line is shortened by the dielectric plug, part of this loss disappears and part is transferred to the dielectric plug. Fortunately these losses are small since they are metal losses at a current node, but for long dielectric plugs or plugs of high dielectric constant the need for correction can arise. The necessary calculations have not yet been reduced to a simple form.

Again, the calculation of half-wave results by means of a quarter wave theory is safe only for a high Q situation. It is easy to show, experimentally, that the maximum line shortening results when the dielectric plug is exactly centered in the line but the calculated power factor is not a maximum here, as might be expected. In the meantime, experience shows that results can be duplicated from day to day and at other frequencies and that over a reasonable range of plug thickness no change in dielectric constant and power factor values, greater than the unavoidable errors of measurement, is obtained.

DESCRIPTION OF APPARATUS

The apparatus can be divided into three parts for purposes of description. The high frequency generator consists of a small "relay rack" assembly,

The above theory applies to the quarter wave line. This is a rather difficult practical one; it is best to add another quarter wave to make a half-wave resonator, shorted at both ends, with the dielectric positioned exactly in the center. From conditions of symmetry we then employ the above equations, taking half of our measured quantities. Or, in terms of the actually measured four lengths which constitute an observation on a half-wave line, $(d-d')$, t , $\Delta\ell$ and λ , we have,

$$\left. \begin{aligned} \text{P.F.} &= \frac{\sinh \frac{\pi}{\lambda} (d - d')}{\sin \frac{2\pi}{\lambda} (\Delta\ell + t)} \cdot \frac{\sin 2X}{2X} \\ \frac{\pi t}{\lambda} \cdot \tan \pi \frac{\Delta\ell + t}{\lambda} &= X \tan X \\ \epsilon &= \left[\frac{X}{\frac{\pi t}{\lambda}} \right]^2 \end{aligned} \right\} \quad (15)$$

which are the expressions used in this work.

In practice the dielectric plug is pushed into the half-wave line and the line is tuned. The line center is then calculated and the plug reset to this. Retuning checks the correct location. Two trials are always sufficient if the plug was nearly centered originally.

There are several shortcomings affecting this theory. The Q of the unloaded line depends partly on metal power loss along the line. When the line is shortened by the dielectric plug, part of this loss disappears and part is transferred to the dielectric plug. Fortunately these losses are small since they are metal losses at a current node, but for long dielectric plugs or plugs of high dielectric constant the need for correction can arise. The necessary calculations have not yet been reduced to a simple form.

Again, the calculation of half-wave results by means of a quarter wave theory is safe only for a high Q situation. It is easy to show, experimentally, that the maximum line shortening results when the dielectric plug is exactly centered in the line but the calculated power factor is not a maximum here, as might be expected. In the meantime, experience shows that results can be duplicated from day to day and at other frequencies and that over a reasonable range of plug thickness no change in dielectric constant and power factor values, greater than the unavoidable errors of measurement, is obtained.

DESCRIPTION OF APPARATUS

The apparatus can be divided into three parts for purposes of description. The high frequency generator consists of a small "relay rack" assembly,

including 60-cycle power panel, rectifier panel, meter and control panel and centimeter wave oscillator panel with coaxial conductor output jack. All high-frequency connectors are coaxial conductor units with plug tips.

The measuring unit is shown in the two photographs; Fig. 1, assembled and Fig. 2, disassembled. Two combination input-output heads are shown in Fig. 2. These heads and tubing together with center conductor and plunger are of coin silver. While the highest possible conductivity metal is desirable, pure silver is mechanically too poor for spring fingers and bearing surfaces and the alloy must be used. The good sliding contact properties of silver are preserved but the conductivity is no better than that of copper. Both heads are drilled, for input and output connections, flush with the bottom of the cylindrical cavity terminating the tubing.

Head #1, shown attached in Fig. 1 and detached in lower right-hand corner of Fig. 2, has a silicon crystal, mounted and insulated in a small cylindrical holder which carries a tiny pickup loop, one side of which is grounded to the cylinder. The total length of pickup conductor including loop and crystal "whisker" is about one centimeter and no tuning is necessary. The loop pickup can be adjusted by moving the holder in or out. The d-c circuit is from an insulated pin on the holder through crystal to apparatus body.

The current input connection is through a coaxial plug which is tapped across a fraction of a tunable half-wave line. This fraction consists of a $\frac{1}{8}$ " coaxial conductor terminated in a tiny feed loop; the remainder of the line is an ordinary $\frac{1}{4}$ " coaxial with sliding plunger. The line is used, well off tune, as an input current amplitude control. The coupling with the cavity in head is adjusted by moving the feed loop in or out.

By inverting another half-wave coaxial with feed loop, so as to put the crystal where the feed jack was, it is possible to use an externally mounted crystal as in head #2. For this head the input current amplitude control is obtained by using, as a feeder, a short $\frac{1}{8}$ " coaxial tipped with a tiny loop and a coaxial jack, at opposite ends. This coaxial is mounted in a spring clamped bearing so as to permit a rotation of the plane of the loop. All coaxials, except the measuring unit itself, are 72-ohm ones.

There is no essential difference in operation between these two heads; they are interchangeable. However, head #1 is more convenient in manipulation, during the disassembly required to insert the dielectric sample. (This sample is always positioned in the piece of tubing connecting to the head.)

An ordinary model 301 microammeter, low resistance, served as indicating instrument. By replacing the crystal holder of head #2 with a loop tipped coaxial and plug, a conventional double-detection radio receiver with

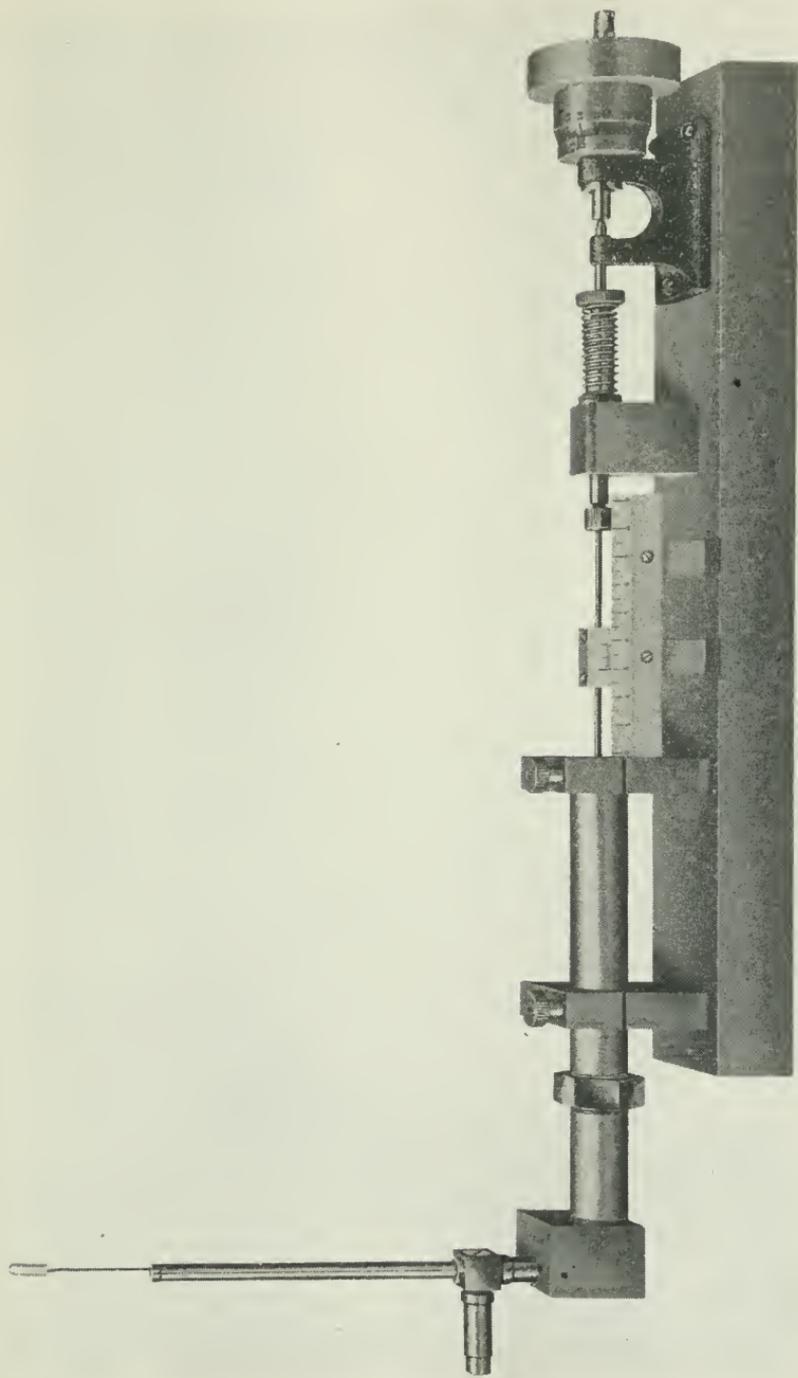


Fig. 1—Measuring unit, assembled

output meter could be used instead. The crystal type detector is by far the most convenient but with the power available wouldn't give workable outputs when bad dielectrics were to be measured. With the amplification available in the double detection set, any dielectric could be measured, while retaining the necessary attenuation between generator-resonator and resonator-receiver to keep these elements electrically independent of each other.

It is necessary to maintain an electrical isolation of this sort to get a high apparatus Q . The equivalent Q of all good dielectrics being high, the measuring apparatus Q must be of the same order to give favorable measuring conditions. And, further, unless the generator-resonator coupling is weak, the act of varying the resonator tune will drag the generator frequency around and will also vary the generator output amplitude.

The crystal plus microammeter required something like 80 millivolts for full scale deflection and this could be obtained with the present apparatus with couplings giving a resonator Q of 1500, while having enough power in reserve to measure any of the good dielectrics. However, most of the dielectrics with power factor greater than .01 were measured with the d.d. receiver. All the 10 cm wave-length measurements were made with this receiver. For the latter measurements a shorter tube was substituted for the tubes shown screwed into the two heads in the disassembly photo.

The crystals were calibrated at 60 cycles by means of a 70-ohm $\sqrt{2}$ attenuation pad.² With full scale deflection this pad was introduced and the new scale deflection read. This $\sqrt{2}$ ratio was, as far as was possible to check, maintained in the kilo megacycle range. For calibration the crystal was tapped across 4 ohms in the attenuator pad output. A 15 mf electrolytic condenser was permanently connected across the meter terminals and, by means of a pair of switches, calibration could be checked in a few seconds, during a measurement run.

The calibration process, using the d.d. set, was to adjust the output to a convenient meter deflection and then calibrate the meter by throwing in 3 db in the IF attenuator.

The resonator itself constitutes an accurate wave meter when corrected for the change in diameter at the moving plunger. The method of operation was then as follows. The plunger vernier, which allowed reading to 0.01 cm., was set at the desired wave-length. The oscillator was then turned on and after it had attained temperature equilibrium, was adjusted if necessary to resonance at this value. This adjustment was infrequently necessary and always slight. The apparatus Q was then determined by traversing the plunger across the resonance setting by means of the micrometer. This

² Exact, not 3 db.

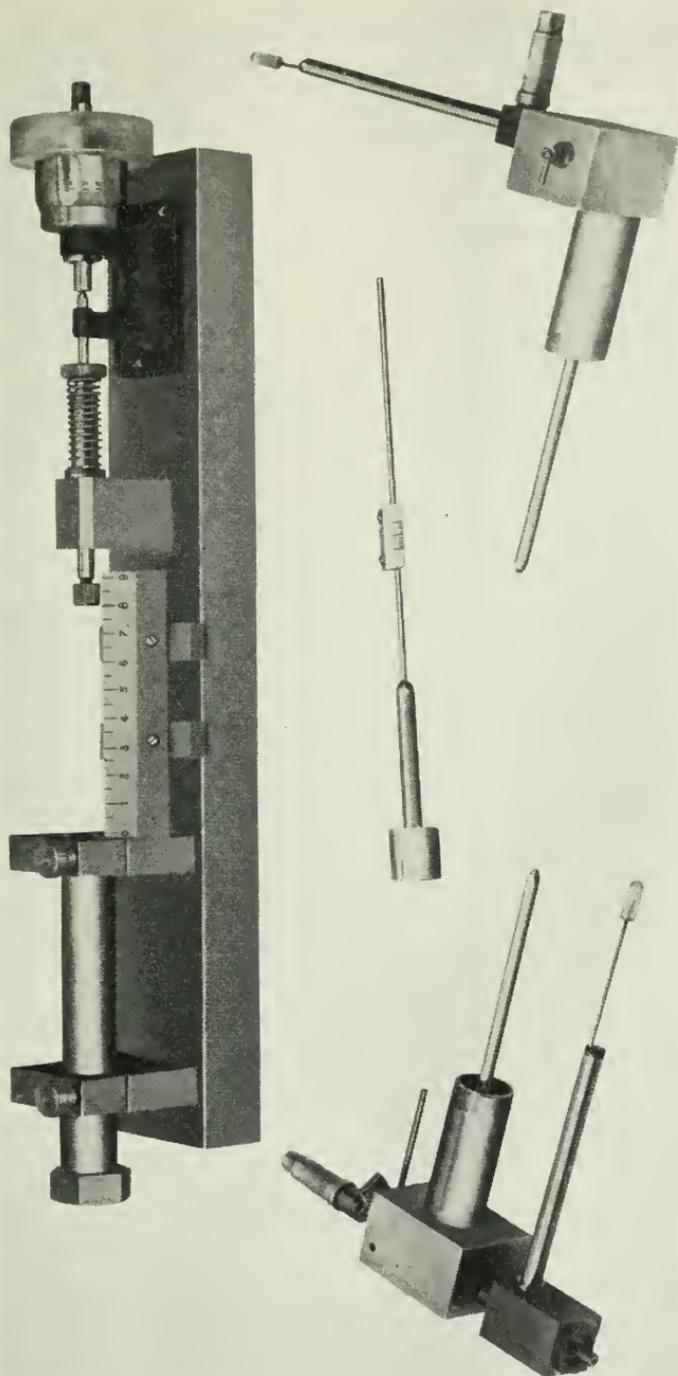


Fig. 2—Measuring unit, disassembled. Two different heads shown

"mike" read to the ten-thousandth of an inch and could be estimated to one-fifth of this. Initially, by means of the amplitude control, the microammeter deflection had been adjusted to the desired scale value at the resonance point. The traverse was observed between the two $\sqrt{2}$ microammeter deflections and was repeated in the opposite direction. When successive round trips showed consistency the value of d' was noted. The dielectric sample, after thickness measurement, was then introduced, centered by cut and try and the Q traverses repeated. This gave d and, after noting $\Delta\ell$, the change in plunger setting for resonance, the measurement was complete.

During the measurement the generator had to be protected from drafts and, usually, it was necessary to traverse rapidly, the power line voltage not being stable. Settings could usually be reproduced to 1 per cent, with adequate care. A sample observation on a good dielectric is the following:

July 28, 1941 Polystyrene plate, all dimensions in cms.

$$\begin{array}{lll} t = 1.28 & d' = .0084 & \lambda = 22.42 \\ \Delta\ell = 1.79 & d = .010 & \text{P.F.} = .00028, \epsilon = 2.49 \end{array}$$

The dielectric samples were machined on a precision lathe, dimensions being held to .001 inch. The nominal dimensions were O.D. .640 inch, I.D. .174 inch. A favorable thickness, from the standpoint of ease of measurement, is $\left| \frac{\lambda}{10\epsilon} \right|$, in cm's. Cleanliness in handling was carefully observed.

After a lapse of several days the interior bearing surfaces of the resonance cavity would have to be cleaned with fine French crocus cloth. The plunger bearing surfaces also had to be smoothed up, fine scratches being polished off. Dirt was immediately noticeable when the plunger contacted it, and when microscopic bits of silver were rolled up under the plunger springs cleaning was necessary. Otherwise no particular treatment or smoothing up of the contacting surfaces was required.

A table of dielectric power factors and constants is a very desirable piece of information. Unfortunately, experience tends to the conclusion that such a table does not exist. The organic plastics in particular, are rather variable from sample to sample and a table of values is merely a table for particular specimens. Where a great number of samples are available "best", "worst" and "most common" values can be established. The accompanying list of observed values must be interpreted in the light of the above statements.

As a large number of measurements of certain special materials had to be made, dielectrics in general were rather neglected and the tabulated values are more or less incidental. It was noted that for the low loss, sub-

TABLE 1

Material	ε		P.F.	
	22.5 cms.	10 cms.	22.5 cms	10 cms.
Ceramic				
BTL F3 Mg. Silicate type...	5.83		.00023	
“Dielectene”		3.39		.0038
Glass, Corning				
G1, lead	4.30		.0049	
G8, lime, annealed	6.38		.0102	
G12, lead	6.08		.0035	
199-1	8.70		.0019	
702EJ, Pyrex	6.35		.0067	
702P	4.70		.0053	
704EO	4.42		.0033	
705BA	3.80		.00118	
707DG	4.69	4.8	.0037	.0036
Glyptal	3.38	3.36	.030	.036
Lucite	2.58	2.56	.0090	.0087
Mycalex				
Red	5.91		.0030	
White	5.74		.0033	
Phenolics				
Cast specimen		4.63		.139
Bakelite sheet $\frac{1}{32}$ "		3.57		.080
Polyethylene				
Worst			{ .00229	
Most common	2.26		{ .00060	
Best			{ .00031	
Polystyrene				
Worst			{ .00090	
Most common	2.45		{ .00070	
Best			{ .00028	
Polyvinylcarbazole	2.87		.0040	
Rubber				
Hard, brown		2.77		.0041
Hard, black		2.69		.0059
Soft, black	3.15		.0058	
Resin	2.32		.0018	
Styralloy				
No. 10	2.49		.0036	
Desig. Unknown	2.49	2.50	.0019	.00105
No. 22	2.40		.0047	
Styramic				
E1689		2.55		.00087
Tenite II		2.95		.031
Vynlite V	2.78	2.61	.0076	.0068
Wax				
Paraffin	2.17		.00019	
Boler	2.17		.00019	
Superla	2.26	2.26	.00019	.00015

stituted paraffin-type, carbon chain dielectrics no difference, greater than experimental error, exists between the 22.5 and 10 cm. measurements.

ACKNOWLEDGEMENT

The measurements by means of the double-detection set were made by my co-worker, Mr. W. E. Eckner, whose valuable assistance I am glad to acknowledge. To Mr. C. F. Mattke, also of the Bell Laboratories, I am indebted for assistance in getting my crude original apparatus into its final finished form.

APPENDIX

The typical ultra high-frequency transmission line can be represented as in Fig. 3

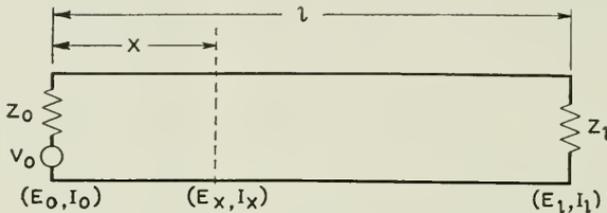


Fig. 3—Equivalent circuit of transmission line

and the equations describing it are

$$\left. \begin{aligned}
 E_x &= V_0 \frac{Z_l \cosh \sqrt{DZ} (\ell - x) + \sqrt{\frac{Z}{D}} \sinh \sqrt{DZ} (\ell - x)}{(Z_0 + Z_l) \cosh \sqrt{DZ} \ell + \left(Z_0 Z_l \sqrt{\frac{D}{Z}} + \sqrt{\frac{Z}{D}} \right) \sinh \sqrt{DZ} \ell} \\
 I_x &= V_0 \frac{\cosh \sqrt{DZ} (\ell - x) + Z_l \sqrt{\frac{D}{Z}} \sinh \sqrt{DZ} (\ell - x)}{(Z_0 + Z_l) \cosh \sqrt{DZ} \ell + \left(Z_0 Z_l \sqrt{\frac{D}{Z}} + \sqrt{\frac{Z}{D}} \right) \sinh \sqrt{DZ} \ell} \\
 Z_x &= \frac{E_x}{I_x} = \sqrt{\frac{Z}{D}} \cdot \frac{Z_l + \sqrt{\frac{Z}{D}} \tanh \sqrt{DZ} (\ell - x)}{\sqrt{\frac{Z}{D}} + Z_l \tanh \sqrt{DZ} (\ell - x)} \\
 E_0 &= V_0 - Z_0 I_0, \quad E_l = Z_l I_l
 \end{aligned} \right\} (1)$$

The line constants are $Z = R + i\omega L$, the series impedance per unit length, and $D = G + i\omega C$, the shunt admittance per unit length. From these we have: surge impedance = $\sqrt{\frac{Z}{D}} = S_0$, propagation constant = \sqrt{DZ} .

For all lines usable as transmission devices the following approximations hold:

$$\begin{aligned} \sqrt{DZ} = \alpha + i\beta \quad \alpha &= \frac{R}{2} \sqrt{\frac{C}{L}} + \frac{G}{2} \sqrt{\frac{L}{C}} \quad \beta = \omega \sqrt{LC} = \frac{\omega}{v} = \frac{2\pi}{\lambda} \\ \sqrt{\frac{Z}{D}} &= \sqrt{\frac{L}{C}}, \quad C = \frac{1}{S_0 V}, \quad L = \frac{S_0}{V}, \quad v = 3 \times 10^{10} \text{ cm/sec.} \end{aligned} \quad (2)$$

for air.

For the case of near-end input and output the second of equations (1) can be rewritten as

$$\begin{aligned} I_0 &= \frac{V_0}{\sqrt{\frac{Z}{D}}} \cdot \frac{\cosh \sqrt{DZ} \ell + \frac{Z_\ell}{\sqrt{\frac{Z}{D}}} \sinh \sqrt{DZ} \ell}{\sqrt{\frac{Z}{D}} \cosh \sqrt{DZ} \ell + \sinh \sqrt{DZ} \ell} \\ &\quad + \frac{Z_0}{\sqrt{\frac{Z}{D}}} \left(\cosh \sqrt{DZ} \ell + \frac{Z_\ell}{\sqrt{\frac{Z}{D}}} \sinh \sqrt{DZ} \ell \right) \end{aligned} \quad (3)$$

If now we assume a quarter-wave line, the condition of resonance implies that $R_\ell \gg \left| \sqrt{\frac{Z}{D}} \right|$ and $R_0 \ll \left| \sqrt{\frac{Z}{D}} \right|$. The condition of reasonable shortening of the line (or lengthening) by the terminal reactance implies that $X_\ell > \left| \sqrt{\frac{Z}{D}} \right|$, $X_0 < \left| \sqrt{\frac{Z}{D}} \right|$. Hence we shall have $|Z_\ell| \gg \left| \sqrt{\frac{Z}{D}} \right|$, $|Z_0| \ll \left| \sqrt{\frac{Z}{D}} \right|$.

If we put $\frac{\sqrt{\frac{Z}{D}}}{Z_\ell} = \tanh \theta$, we get

$$I_0 = \frac{V_0}{\sqrt{\frac{Z}{D}}} \cdot \frac{\tanh (\sqrt{DZ} \ell + \theta)}{1 + \frac{Z_0}{\sqrt{\frac{Z}{D}}} \tanh (\sqrt{DZ} \ell + \theta)} \quad (4)$$

We now make the assumption that " Z_0 " is a pure resistance (which is no limitation on the measurement to be discussed) and put $\theta = a_\ell + ib_\ell$.

Then,

$$|I_0| = \frac{V_0}{\sqrt{\frac{L}{C}}} \cdot \frac{\sqrt{\tanh^2(\alpha\ell + a_\ell) + \tan^2\left(\frac{2\pi\ell}{\lambda} + b_\ell\right)}}{\left[1 + \frac{Z_0}{\sqrt{\frac{L}{C}}} \tanh(\alpha\ell + a_\ell)\right]^2 + \left[\frac{Z_0}{\sqrt{\frac{L}{C}}} + \tanh(\alpha\ell + a_\ell)\right]^2 \tan^2\left(\frac{2\pi\ell}{\lambda} + b_\ell\right)} \quad (5)$$

This expression cycles, as “ ℓ ” is varied, and has its maximum or “tuned” value of

$$\frac{V_0 / \sqrt{\frac{L}{C}}}{\frac{Z_0}{\sqrt{\frac{L}{C}}} + \tanh(\alpha\ell + a_\ell)} \quad \text{for} \quad \tan\left(\frac{2\pi\ell}{\lambda} + b_\ell\right) = \infty$$

$$\text{or} \quad \frac{2\pi\ell}{\lambda} + b_\ell = \frac{(2n + 1)\pi}{2} \quad n = 0, 1, 2, \dots$$

The resonant length is thus $\ell = \frac{\lambda}{4} \left(1 - \frac{2b_\ell}{\pi}\right)$, for $n = 0$. Note that successive resonances differ by a line length of $\frac{\lambda}{2}$; the reactive termination has merely shortened, by the amount of $\Delta\ell = \frac{\lambda b_\ell}{2\pi}$, the first resonant element preceding it. When, therefore, we measure the “ Q ” of this line segment by line-length tuning we use $\ell = \frac{\lambda}{4}$ in the Q process definition.

The Q process now follows. Putting $\ell = \ell_r \pm \delta\ell$ where ℓ_r is the actual observed resonance length, we have

$$\frac{2\pi\ell}{\lambda} + b_\ell = \frac{2\pi\ell_r}{\lambda} + b_\ell \pm \frac{2\pi\delta\ell}{\lambda} = \frac{\pi}{2} \pm \frac{2\pi\delta\ell}{\lambda}$$

$$\text{Then } \tan\left(\frac{2\pi\ell}{\lambda} + b_\ell\right) = \tan\left(\frac{\pi}{2} \pm \frac{2\pi\delta\ell}{\lambda}\right) = \frac{1}{\mp \tan \frac{2\pi\delta\ell}{\lambda}} \quad \text{and}$$

$$|I_0| = \frac{V_0}{\sqrt{\frac{L}{C}}} \sqrt{\frac{1 + \tanh^2(\alpha\ell + a_\ell) \cdot \tan^2 \frac{2\pi\delta\ell}{\lambda}}{\left[\frac{Z_0}{\sqrt{\frac{L}{C}}} + \tanh(\alpha\ell + a_\ell)\right]^2 + \left[1 + \frac{Z_0}{\sqrt{\frac{L}{C}}} \tanh(\alpha\ell + a_\ell)\right]^2 \tan^2 \frac{2\pi\delta\ell}{\lambda}}}$$

Forming the current values $|I_{01}| = |I_{02}| = \left| \frac{I_0(\text{resonant})}{K} \right|$, dividing to eliminate $|I_0|$ and discarding squares and products of small quantities in comparison with unity leaves,

$$\sqrt{K^2 - 1} = \frac{\tan \frac{2\pi\delta\ell}{\lambda}}{\frac{Z_0}{\sqrt{\frac{L}{C}}} + \tanh(\alpha\ell + a_\ell)} \quad (6)$$

$$\text{or } \sqrt{K^2 - 1} \frac{\lambda}{8\delta\ell} = \frac{\frac{\lambda}{8\delta\ell} \cdot \tan \frac{2\pi\delta\ell}{\lambda}}{\frac{Z_0}{\sqrt{\frac{L}{C}}} + \tanh(\alpha\ell + a_\ell)}$$

which becomes our "Q" when $K = \sqrt{2}$.

In most practical situations the "tan" and "tanh" are equal to their

angles. For this condition $Q = \frac{\frac{\pi}{4}}{\frac{Z_0}{\sqrt{\frac{L}{C}}} + \alpha\ell + a_\ell}$. If we now put, by defini-

tion, $Q_a = \frac{\pi}{4\alpha\ell}$ (the Q of the line itself), $Q_0 = \frac{\pi}{4} \cdot \frac{\sqrt{\frac{L}{C}}}{Z_0}$, $Q_\ell = \frac{\pi}{4a_\ell}$, we have

$$\frac{1}{Q} = \frac{1}{Q_a} + \frac{1}{Q_0} + \frac{1}{Q_\ell} \quad (7)$$

the law of Q composition relating the resultant Q to line and terminal Q's.

Abstracts of Technical Articles by Bell System Authors

*A Sampling Inspection Plan for Continuous Production.*¹ H. F. DODGE. This paper presents a plan of sampling inspection for a product consisting of individual units (parts, sub-assemblies, finished articles, etc.) manufactured in quantity by an essentially continuous process.

The plan, applicable only to characteristics subject to non-destructive inspection on a Go-NoGo basis, is intended primarily for use in process inspection of parts or final inspection of finished articles within a manufacturing plant, where it is desired to have assurance that the percentage of defective units in accepted product will be held down to some prescribed low figure. It differs from others which have been published in that it presumes a *continuous flow of consecutive articles or consecutive lots* of articles offered to the inspector for acceptance in the order of their production. It is accordingly of particular interest for products manufactured by conveyor or other straight line continuous processes.

In operation, the plan provides a corrective inspection, serving as a partial screen for defective units. Normally, a chosen percentage or fraction f of the units are inspected, but when a defective unit is disclosed by the inspection it is required that an additional number of units be inspected, the additional number depending on how many more defective units are found. The result of such inspections is to remove some of the defective units, and the poorer the quality submitted to the inspector, as measured in terms of per cent defective, the greater will be the corrective or screening effect. The object of the plan is the same as that incorporated in some of the sampling tables already published, namely, to establish a limiting value of "average outgoing quality" expressed in per cent defective which will not be exceeded no matter what quality is submitted to the inspector. This limiting value of per cent defective is termed the "average outgoing quality limit (AOQL)."

The theoretical solution treats the case of inspecting a continuous flow of individual units and is based on the distribution of *random-order* spacing of defective units in product whose quality is statistically controlled. Part III of the paper extends the application of the method to a continuous flow of individual lots or sub-lots of articles.

*Stability in High-Frequency Oscillators.*² R. A. HEISING. This paper discusses frequency stability with change in plate voltage of high-frequency

¹ *The Annals of Mathematical Statistics*, September 1943.

² *Proc. I.R.E.*, November 1943.

oscillators of around 100 megacycles and shows both theoretically and experimentally that the highest stability found by many is only the result of fortuitous circuit adjustment that may readily lead to the desired result in this frequency range. It is shown that the factor next in importance in producing frequency stability is a low ratio of inductance to capacitance in the frequency-determining circuit. It is also shown that a high Q contributes little directly to stability. A high Q is necessary with low L/C ratios to get oscillations but an improvement in Q alone may give poorer stability. To get the fullest measure of stability with low L/C and high Q calls for slight adjustments in the circuit and possibly the provision of loose coupling to the frequency-determining circuit.

*Modern Spectrochemical Analysis.*³ EDWIN K. JAYCOX. The spectrograph, originally developed by the physicist, has become a most useful tool in the hands of the analytical chemist. Today few large analytical laboratories are without one. The instrument, with its attendant accessories, provides a rapid method for analyzing metals, alloys, minerals, ores, liquids, and gases, particularly for their metallic constituents and in some cases for their anions. Both emission and absorption spectra are important to the analyst. Important applications of the spectrograph to the analytical problems of research and industrial organizations are discussed.

The spectrograph did not come into general use as an analytical tool until the early 1920's, although Kirchhof and Bunsen saw the practicability of the method in 1860, when they published their paper entitled, "Chemical Analysis by Means of Spectral Observations." During the intervening years only a few enthusiasts like Lockyer, Roberts, Hartley, Leonard, Pollack, and de Gramont, kept the art alive. In spite of their persistent efforts to influence chemists to use spectrographic methods, they were quite generally ridiculed and the value of the method was recognized by only a few workers.

In 1922, Meggers, Kiess, and Stimson published their paper "Practical Spectrographic Analysis" and modern spectrochemical analysis was born. Under the stimulus of this paper and the backing of a high caliber scientific organization like the Bureau of Standards, the use of the spectrograph as an analytical tool increased rapidly. This is evidenced from the *Index to the Literature on Spectrochemical Analysis* by Meggers and Scribner. In 1920, for example, only five papers were published concerning spectrochemical analysis, four of which were by de Gramont; whereas in 1930, 33 papers were published and in 1939, 170 papers, indicating an increasing interest in and use of spectrochemical analysis in industrial and research organizations.

³ *Jour. Applied Physics*, December 1943.

*Determination of Small Amounts of Arsenic, Antimony, and Tin in Lead and Lead Alloys.*⁴ C. L. LUKE. A new method for the determination of small amounts of arsenic, antimony, and tin in lead and lead alloys consists of separation of the three metals from the lead by a double co-precipitation with manganese dioxide, reduction of arsenic and antimony to the trivalent state, separation of the arsenic by distillation as chloride, titration of the arsenic and antimony separately by the method of Gyory, and reduction of tin with lead and titration with standard iodine solution.

*Determination of Total Sulfur in Rubber.*⁵ C. L. LUKE. A new volumetric method has been developed for the determination of sulfate sulfur. The sulfate is reduced to sulfide by treatment with hydriodic acid and the hydrogen sulfide is distilled off and titrated iodometrically. The new method has been applied to the determination of total sulfur in natural and synthetic rubber.

*Machine Screws. Fastening Strengths in Various Materials.*⁶ A. C. MILLARD. Although standard machine screws in the numbered sizes have been widely used as fastenings for many years, very little has been published concerning their strength of fastening in various metals and non-metals. Numerous articles have appeared regarding the strength of bolts and machine screws for $\frac{1}{4}$ in. and larger sizes, but very little, if any, published information is available on the strength of machine-screw fastenings in the numbered sizes.

The need for machine-screw fastening-strength information has increased recently due to the use of more compact designs and the shortage of materials. The use of substitute materials has accentuated the lack of machine-fastening-strength information in making fastenings in such materials, as well as in the more commonly used materials. Frequently, it is desirable to know the load-carrying capacity of screw fastenings of various diameters, as well as the length of thread engagement in the weaker materials needed to develop either the full strength of the screw, or the strength of fastening required of the assembly. The purpose of this paper is to make available to designers the results of fastening-strength tests of machine-screw fastenings in a number of materials, which were carried out by the author at the Bell Telephone Laboratories, Inc. The work is by no means complete but is hoped that the data offered will prove to be of some use in its present form.

⁴ *Indus. & Engg. Chemistry*, October 1943.

⁵ *Indus. & Engg. Chemistry*, September 1943.

⁶ *Mech. Engg.* October 1943.

Contributors to this Issue

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.

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.

D. K. GANNETT, B.S. in engineering, University of Minnesota, 1916; E.E. University of Minnesota, 1917. American Telephone and Telegraph Company, Engineering Department, 1917-1919; Department of Development and Research, 1919-1934; Bell Telephone Laboratories, Inc. 1934-. Prior to October 1942 Mr. Gannett, as Toll Transmission Engineer, was concerned with the transmission requirements of toll systems including program circuits. Since then, as Circuit Research Engineer, he has directed a group engaged in research and development on war projects.

IDEN KERNEY, B.S. Harvard University, 1923. American Telephone and Telegraph Company, 1923-1934; Bell Telephone Laboratories, Inc. 1934-. Before the war Mr. Kerney was in charge of the laboratory in which experimental work on program transmission was conducted. Since early in 1942 he has been engaged full time on war projects.

R. A. SYKES, Massachusetts Institute of Technology, B.S. 1929; M.S. 1930. Columbia University, 1931-1933. Bell Telephone Laboratories, Research Department, 1930-. Mr. Sykes has been engaged in the applications of quartz crystals to broad-band carrier systems as filter and oscillator elements. Other work has included the application of coaxial lines as elements of filter networks and more recently the design and development of quartz crystals for radio frequency oscillators.

G. W. WILLARD, B.A., University of Minnesota, 1924; M.A., 1928; Instructor in Physics, University of Kansas, 1927-28; Student and Assistant, University of Chicago, 1928-30. Bell Telephone Laboratories, 1930-. Mr. Willard's work has had to do with special problems in piezo-electric crystals.

Public Lib
City, Mo.

THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS
OF ELECTRICAL COMMUNICATION

Indicial Response of Telephone Receivers . . . *E. E. Mott* 135

Theoretical Analysis of Modes of Vibration for Isotropic
Rectangular Plates Having All Surfaces Free
—*H. J. McSkimin* 151

Principles of Mounting Quartz Plates . . . *R. A. Sykes* 178

The Magnetically Focused Radial Beam Vacuum Tube
—*A. M. Skellett* 190

Abstracts of Technical Articles by Bell System Authors 203

Contributors to this Issue 206

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

THE BELL SYSTEM TECHNICAL JOURNAL

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



EDITORS

R. W. King

J. O. Perrine

EDITORIAL BOARD

F. B. Jewett

M. R. Sullivan

O. B. Blackwell

O. E. Buckley

A. B. Clark

H. S. Osborne

S. Bracken

M. J. Kelly

F. A. Cowan



SUBSCRIPTIONS

Subscriptions are accepted at \$1.50 per year. Single copies are 50 cents each.
The foreign postage is 35 cents per year or 9 cents per copy.



Copyright, 1944
American Telephone and Telegraph Company

Indicial Response of Telephone Receivers

By E. E. MOTT

A method of analyzing telephone receiver characteristics by indicial response is discussed and illustrated by oscillograms. The indicial response of a telephone receiver is the instantaneous response of the receiver to a suddenly applied electromotive force. This type of response is of particular fundamental interest because it furnishes a key to the solution of transient problems such as are involved in the response to speech waves.

Oscillograms of indicial response, together with the more familiar steady-state frequency response characteristics, are shown for different types of receivers. The relationships existing between the two types of measurements are discussed.

From the standpoint of most faithfully reproducing transients, indicial response data indicate that a receiver having a limited range of frequency response should have a frequency response characteristic which droops gradually rather than abruptly near the upper end of the range.

INTRODUCTION

THE use of indicial response analysis as an outgrowth of the Heaviside operational calculus¹ has been extended to a number of different fields. The indicial admittance as defined by J. R. Carson² in his analysis of the submarine cable and other transmission problems has been an effective tool in the study of transients. More recently, a similar type of measurement has been used as an indication of performance of amplifiers³, television equipment⁴, and audio frequency transformers⁵.

In the field of telephone receivers⁶ an analysis by means of impressed square waves has been found useful as a measure of transient response. In the transmission of speech, so much emphasis has been placed upon steady-state frequency response as an indication of performance, that it seems in order to consider the possible advantages of a transient method of analysis, as obtained by measuring the indicial response. Only recently has the technique of such measurement been made feasible by the improvement at low frequencies of amplifiers and related apparatus.

THE INDICIAL RESPONSE

The indicial response of a telephone receiver may be defined as the instantaneous sound pressure generated by the receiver in a closed air chamber due to a suddenly-applied unit voltage. This term differs from Carson's indicial admittance only in that sound pressure rather than current response is used. The sound pressure in an air chamber of pure stiffness is a measure

of the volume displacement, and as such it is proportional to the transfer displacement admittance of the system. When we are interested in the charge rather than in the current, the admittance takes the form of a displacement admittance, related to the ordinary admittance by a factor of the frequency ω . That Carson's original equations apply to such a system with little if any change may be easily demonstrated. The term $A(t)$ may be used to denote any of these forms of indicial admittance or indicial response.

The form of the applied voltage assumed is shown by Fig. 1. This form, defined by Heaviside as the *unit function*, is a function of time equal to zero before, and unity after the time $t = 0$. More properly, however, it may be regarded as an increment in voltage closely analogous to Isaac Newton's concept of infinitesimal elements of rectangular area, the summation of which forms the basis of the integral calculus. The successive application of small increments of voltage likewise forms the basis of the operational calculus, or more particularly, the basis of the Carson extension theorem.

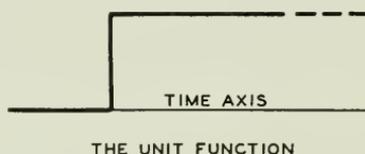


Fig. 1

THE CARSON EXTENSION THEOREM

Having obtained the indicial response, either experimentally or theoretically, we have the key to the more general problem where the applied voltage $e(t)$ may be of any form, such as that of speech waves. Let $e(t)$, Fig. 2, be any arbitrary voltage wave corresponding to speech⁷. Let a series of consecutive increments of voltage, differing in time by $\Delta\tau$ be applied, of such magnitude as to build up the form of the curve $e(t)$. By analyzing each of these components in terms of the indicial admittance $A(t)$, and synthesizing them again, the instantaneous sound pressure may be related to the voltage producing it and the indicial admittance $A(\tau)$ by the Carson extension equation²:

$$p(t) = \frac{d}{dt} \int_0^t A(\tau)e(t - \tau) d\tau$$

• When the above integration is carried out, the term τ disappears and is replaced by t . The above sound pressure $p(t)$ represents the sound pressure generated by the receiver in a closed coupler due to an applied voltage $e(t)$.

$$p(t) = \int_0^{\infty} A(\tau) e(t-\tau) d\tau$$

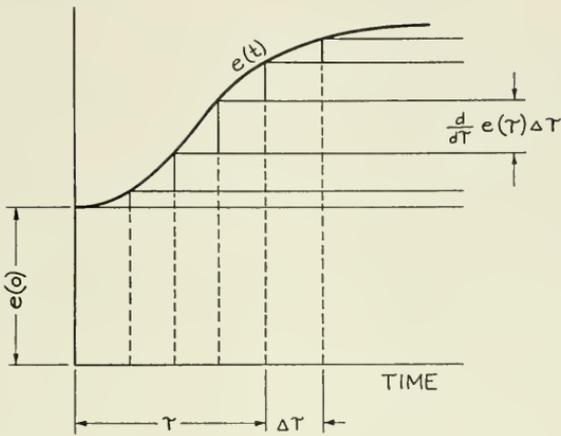


Fig. 2—Method of derivation of Carson's extension formula.

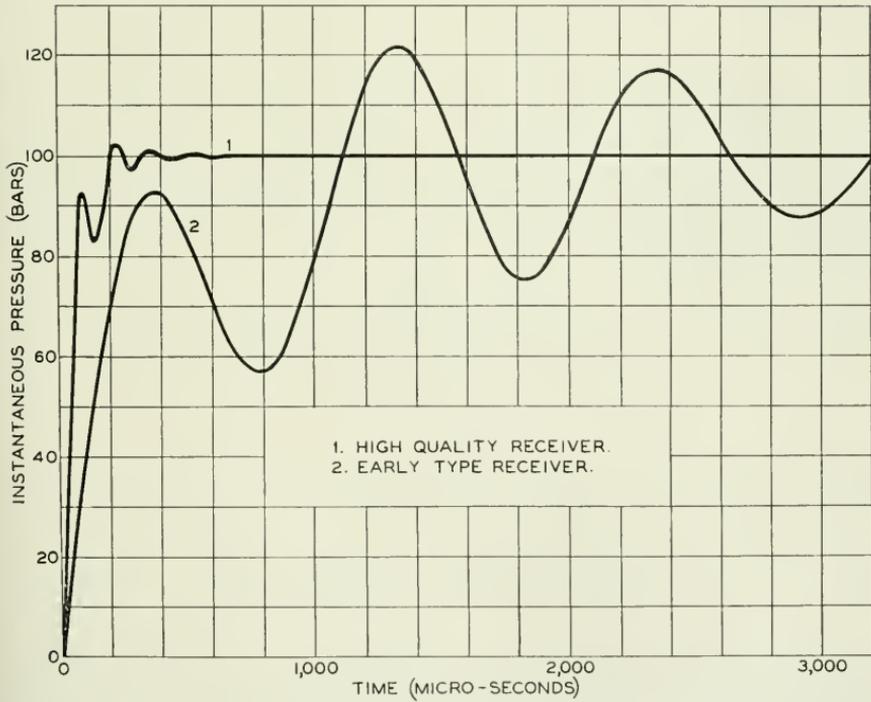


Fig. 3—Indicial admittance of two types of telephone receivers.

From the above, it is evident that the ideal form of receiver response to a suddenly-impressed voltage would be a copy of the unit function shown in Fig. 1, and that any deviation from this form will cause distortion. If the building blocks of the curve $e(t)$ are undistorted, the curve itself will likewise be reproduced free from distortion of wave form. Thus, the more closely the indicial response can be made to approach the form of the unit function, the more closely the receiver sound pressure $p(t)$ will be a copy of any arbitrary speech wave $e(t)$. Curve 1, Fig. 3, shows the indicial response of a receiver having a frequency range of 8000 cps, which comes rather close to this ideal. On the other hand, the further the indicial response departs from this ideal form, the more it will deviate from any impressed transient, such as speech waves. Thus curve 2, Fig. 3, corresponds to a receiver of narrow range, which contains resonant oscillations, and rises much later in time than the other receiver.

CONVERSION FORMULAE

The indicial response is as fundamental in character as frequency response, and may be converted into frequency and phase response if the proper integrations are carried out for any particular system, as follows:

$$\text{Indicial Response } A(t) \Leftrightarrow \left[\begin{array}{l} \text{Frequency Response} \\ + \text{Phase Response} \end{array} \right] A(\omega) = P(\omega) + jQ(\omega)$$

where $A(\omega)$ is the transfer admittance of the system. In order to carry out these conversions, certain integrations must be performed, either mechanically or theoretically. The following are conversions⁷ which may be used to carry out this process:

$$A(t) = \frac{2}{\pi} \int_0^{\infty} \frac{P(\omega)}{\omega} \sin \omega t \, d\omega$$

$$A(t) = P(0) + \frac{2}{\pi} \int_0^{\infty} \frac{Q(\omega)}{\omega} \cos \omega t \, d\omega$$

$$\frac{P(\omega)}{\omega} = \int_0^{\infty} A(t) \sin \omega t \, dt$$

$$\frac{Q(\omega)}{\omega} = \int_0^{\infty} [A(t) - A(0)] \cos \omega t \, dt$$

Where $P(\omega)$ and $Q(\omega)$ are the real and imaginary parts of the frequency response, $A(\omega)$ is expressed in terms of pressure response⁸, while the indicial response $A(t)$ is expressed as an instantaneous sound pressure. The integrations are difficult to carry out, but serve to show how the two systems of

measurement are related, and how they may theoretically be converted one into the other, provided in the case of frequency response the magnitude and phase are both known.

GENERAL APPLICATIONS

The use of indicial response as a tool in telephone receiver studies is particularly adapted to the study of transients. Since all voice and sound transmission, particularly that of orchestral music, may be regarded as essentially a transient problem, it is appropriate that we visualize the effects on the complex wave forms of any distortions which may be present in the transmission apparatus. The indicial response will, in general, depart from the ideal square form, and the amount of this departure may be regarded as indicative of the relative faithfulness of wave form reproduction by apparatus having different frequency characteristics. An examination of these departures should therefore be helpful as a supplementary method of appraising the relative merits of different frequency response characteristics. The effect, for example, of small resonance peaks or dips upon transients is very forcefully shown in the form of the indicial admittance. The departure from squareness of a particular system may often be improved by use of the proper shape of frequency characteristic.

The use of a closed coupler when measuring telephone receivers is particularly adapted for such studies, because the disturbing effects of deficiencies at the low frequencies due to leakage may thus be eliminated. Interpretation by inspection then becomes a matter of observation of the various types of departures at the higher frequencies from the ideal form.

Since listening tests do not always agree with interpretations of physical measurements of steady-state frequency response, it often becomes a matter of interest to obtain different criteria of judgment in which the weight given to the various frequencies may be judged by the relative effects of irregularities in various parts of the frequency spectrum upon the indicial response.

APPARATUS AND METHOD OF TESTING

Various forms of apparatus may be used for receiver testing with square waves. Square-wave generator circuits have been published both for audio⁵ and video³ frequency use, involving vacuum tube circuits which overload at low voltages. For low speeds using low-frequency waves of the order 60 cps, a simple mercury switch operated by an oscillator gives very satisfactory results.

The square-wave voltage is introduced across a small part of the resistance termination as shown in Fig. 4, the whole resistance termination being matched to the magnitude of the receiver impedance at 800 cps. The re-

ceiver is then operating from an idealized resistance source having an impedance which matches that of the receiver approximately, over the range of interest.

The receiver is coupled acoustically to a small-diameter condenser microphone by means of a closed coupler⁸. The condenser microphone has a substantially uniform characteristic up to a frequency of 10 kc. The

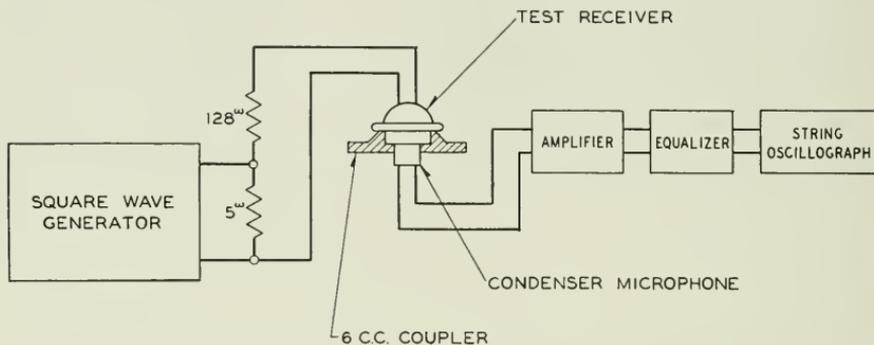
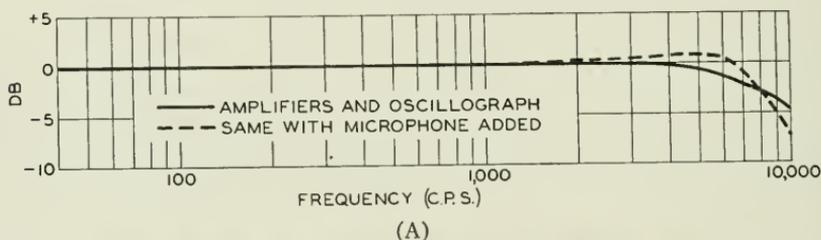


Fig. 4—Circuit diagram of apparatus for indicial response measurements.



(B)

Fig. 5—Frequency response (A) and indicial response (B) of measuring apparatus.

microphone voltage is then amplified to the point where it can be measured by an oscillograph.

Either the cathode-ray oscilloscope or a rapid-recording string oscillograph⁹ may be used, but in the latter case it is necessary to equalize the string oscillograph to a frequency of about 10 kc in order to cover the audio frequency range. The choice of these instruments depends somewhat upon whether a permanent record is desired or whether a visual indication is sufficient.

The amplifier must be compensated at low frequencies in order to maintain a strictly square-wave output. The entire system characteristic is shown in Fig. 5 and covers a range of 1 to 10,000 cps with a substantially uniform frequency response. The indicial response of the system is also shown to be reasonably free from irregularities. Such irregularities as do exist are due largely to the sharp cut-off of the system at 10 kc which was necessitated by the limitations of the string oscillograph.

INDICIAL VS. FREQUENCY RESPONSE

The calculated pairs of curves for telephone receivers in Fig. 6 show the relations between the frequency response and the indicial response. Since the characteristics of receivers measured on a closed coupler of known volume are readily amenable to calculation if the constants of the receiver are known, such a procedure is often useful in predetermining the design of a receiver.

The upper three curves, Fig. 6, are the characteristics of a moving coil receiver calculated for three different frequency ranges, being otherwise similar in shape, the curve being shifted in frequency by an arbitrary factor K . The effect on the indicial admittance is to shift it in time by the same factor without change of shape, if the plot is logarithmic as shown. In general, if the cut-off frequency is divided by the factor K , the corresponding time delay will be increased by the factor K . This is an application of a theorem by Carson² that:

$$\frac{1}{pZ(kp)} = \int_0^{\infty} A(t/k)\epsilon^{-pt} dt$$

where $p = j\omega$ is proportional to frequency, and t is the time, $\frac{1}{Z(kp)}$ is the frequency response, and $A(t/k)$ is the indicial response. In other words, the curve may be shifted in frequency by a simple transformation and the effect on the indicial admittance curve is very similar except that the shift is in a direction opposite to the change in frequency, and is inversely proportional to the change in frequency scale.

The second group of curves, Fig. 6, relates to the effect of damping on an early magnetic type of receiver, showing the freely resonant condition, a moderately damped, and a highly damped receiver. The curves of indicial response show the effects of free resonance to be very detrimental, and the ringing of the diaphragm is sustained over such a long period that any speech waves would have superposed on them a continual train of sine waves. If the rate of decay of these waves is increased, as shown by the damped curves, a noticeable improvement results. By using critical damping as in the highly damped curve, all oscillations can be eliminated, but the time of pickup is degraded and the departure from a square wave is somewhat greater than for the moderately damped condition.

INDICIAL RESPONSE

CALCULATED RECEIVER CHARACTERISTICS

FREQUENCY RESPONSE

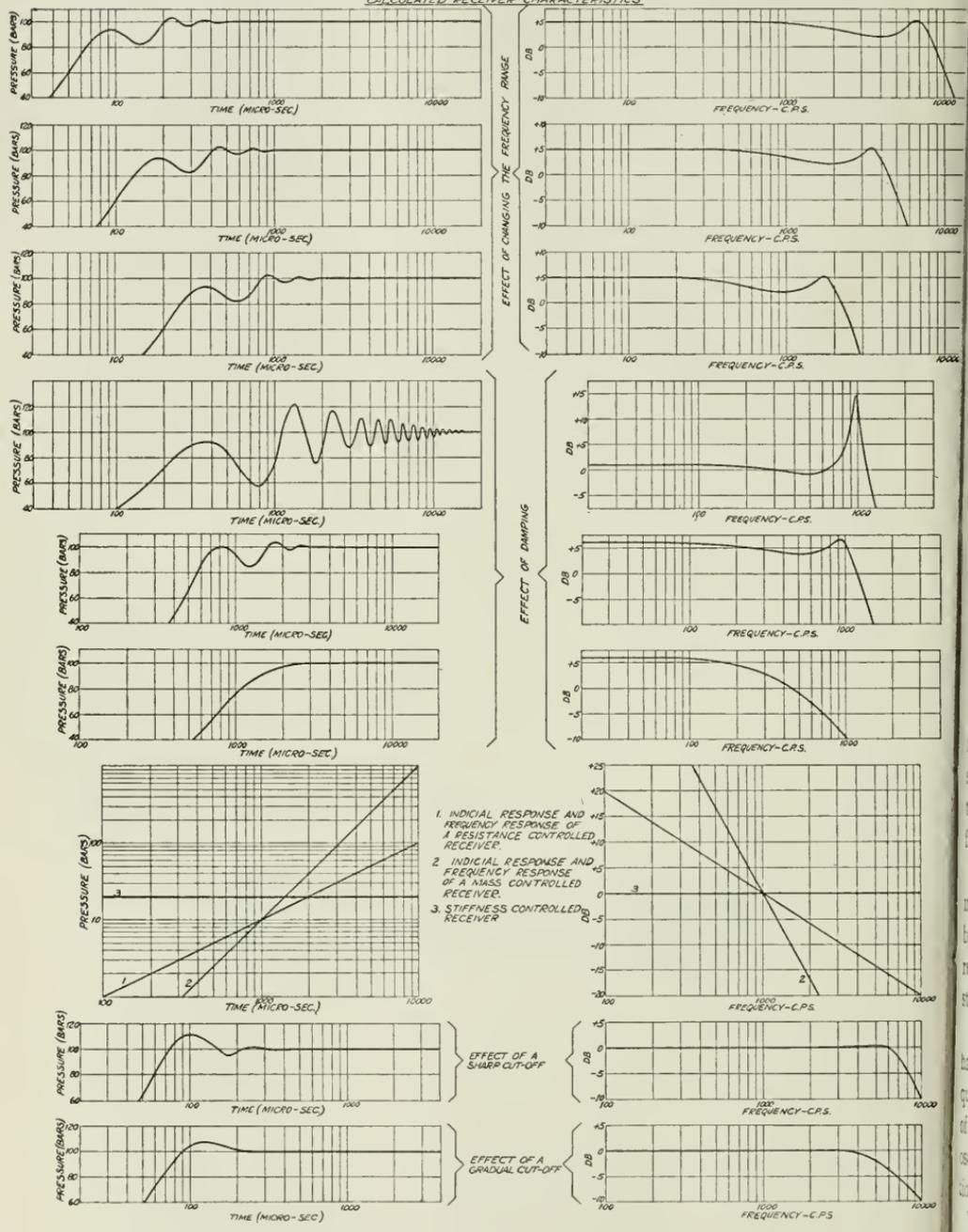


Fig. 6—Calculated indicial response versus calculated frequency response of various types of telephone receivers.

The indicial response shows more emphatically than frequency response, the importance of damping and the oscillations which are to be avoided, or reduced to a minimum. It also shows that the effect of delay is closely related to attenuation of the higher frequencies, and that frequency of cut-off is inversely proportional to the time delay, for a given type of receiver circuit.

There is a noticeable similarity between the appearance of the frequency response and the indicial response curves, and in many cases one curve is approximately the image of the other. As an example of this, the three pairs of linear curves show the similarity of indicial and frequency response for constant velocity, constant acceleration, and constant amplitude devices, as depicted by the three curves denoted by 1, 2, and 3 in which the three moving-coil instruments are assumed to be controlled by (1) a predominance of acoustic resistance behind the diaphragm, (2) a mass controlled system, and (3) a stiffness controlled system. In either case, the fundamental shape of the curves is such that the indicial response is the image of the frequency response in its general character.

The two lower curves, Fig. 6, indicate the effect of a sharp cut-off versus a gradual one. In terms of indicial response, the gradual cut-off appears to be the better of the two, a principle which is widely accepted in television and telegraph transmission.

EXPERIMENTAL MEASUREMENTS

The oscillographic measurements of indicial response, together with corresponding frequency response measurements of telephone receivers, are shown in Figs. 7, 8, and 9. The oscillograms on the left, Fig. 7, show the type of data which constitute indicial response as compared with the more familiar frequency response on the right.

Curve 1, Fig. 7, represents a moving-coil receiver similar to that calculated in Fig. 3, and constitutes the standard of performance which can be obtained by this particular system of measurement. Each division of the oscillogram represents .001 second, a somewhat faster film speed than is usual for the string oscillograph.

Curve 2 shows the characteristics of a magnetic bipolar type of receiver having a frequency range of 3000 cps with a fairly sharp cut-off at this frequency. The acoustic circuits of this receiver serve to damp the resonance of the diaphragm and extend the range from 1600 up to 3000 cps. The oscillogram shows a partially damped but still somewhat oscillatory condition which is due to the receiver.

With all damping circuits removed, we obtain the characteristic of curve 3, a simple diaphragm resonance, which is similar to the earlier type of receivers of the magnetic type. Curve 2 represents a real improvement over

curve 3, both as regards introduction of damping and extending the frequency range.

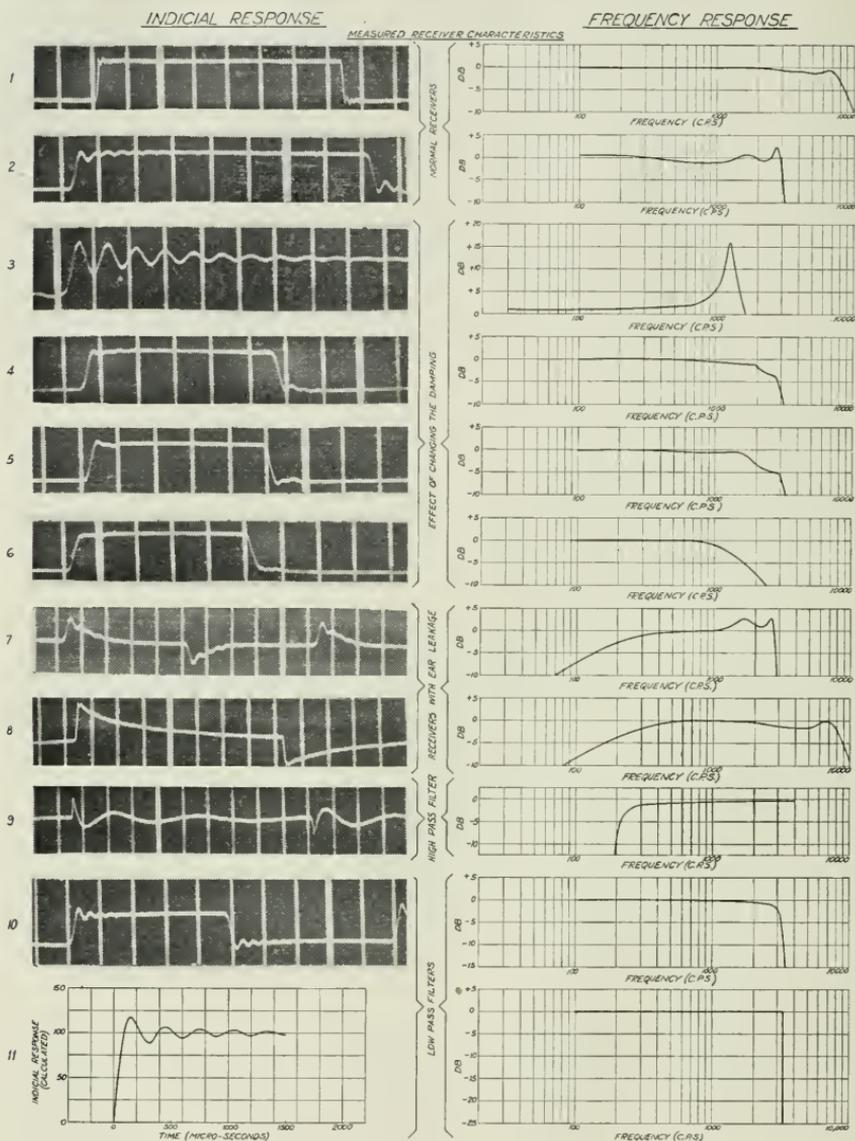


Fig. 7—Measured indicial response versus measured frequency response of various types of telephone receivers and electrical filters.

The effects of further increases in damping are shown by curves 4, 5, and 6. Such changes in the shape of the curve are brought about by relatively simple

changes of the constants of the acoustic circuits. The oscillograms indicate a marked improvement as regards oscillations, which is to be expected with increased damping. The time delay is eventually degraded with further increases of damping, however, and the optimum damping is a matter of compromise.

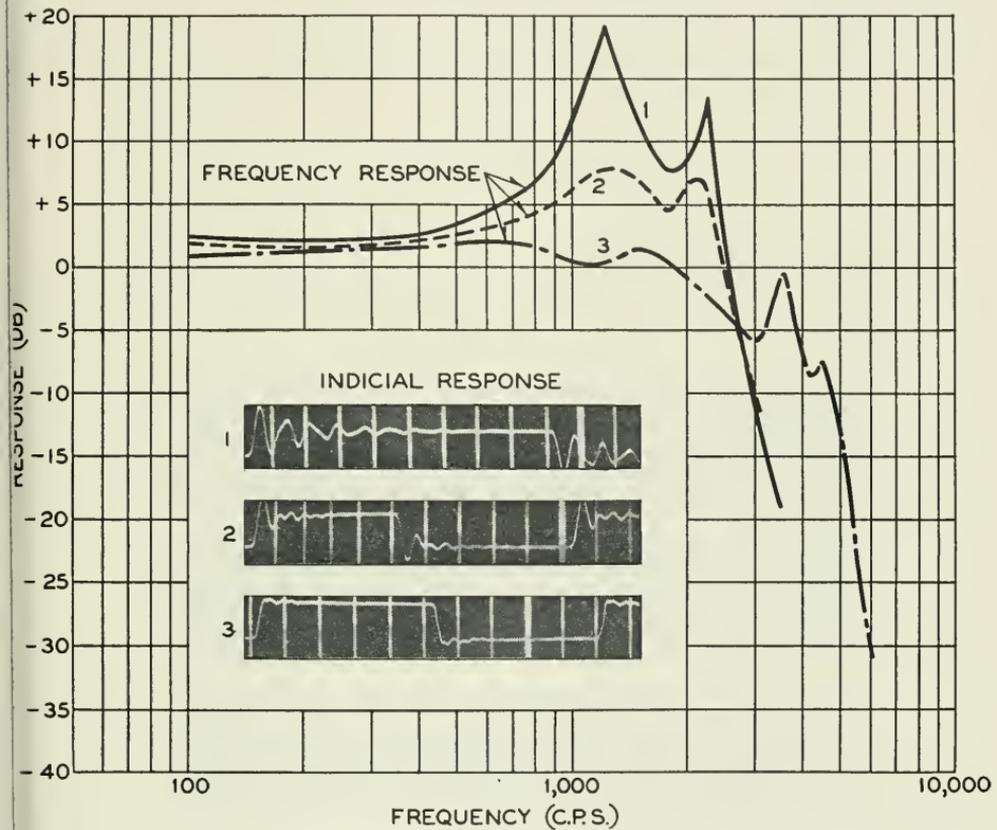


Fig. 8—Three types of hearing aid receivers—frequency response and indicial response.

The effects of a low-frequency cut-off characteristic are shown by curves 7, 8, and 9, Fig. 7. The absence of a *d-c* component makes these curves very difficult of interpretation.

Curve 7, taken with the same receiver as curve 2, except with coupler leakage, shows a loss at low frequencies which is typical of cases where the receiver cap does not make a perfect seal with the ear. The effect on the indicial response is that of a large pulse followed by a few oscillations at the frequency of the leak circuit.

Curve 8 is a similar condition except taken on a high-quality receiver

circuit. This also shows a similar effect. The initial pulse contains most of the receiver characteristic, while the curve which follows is mainly dependent on the leakage constants.

Curve 9 is taken on a high-pass filter of the characteristic shown. It may be proved that this curve is the inverted image of the corresponding low-pass filter characteristic, of which a similar curve is shown as curve 10.

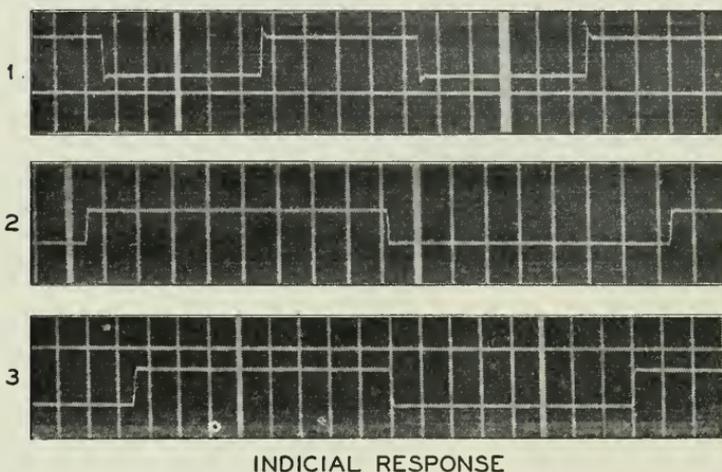
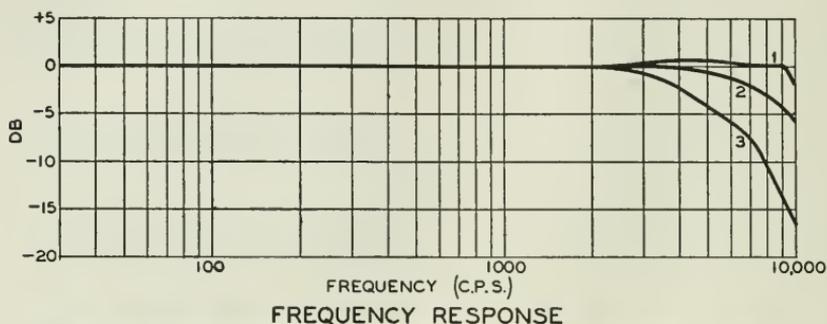


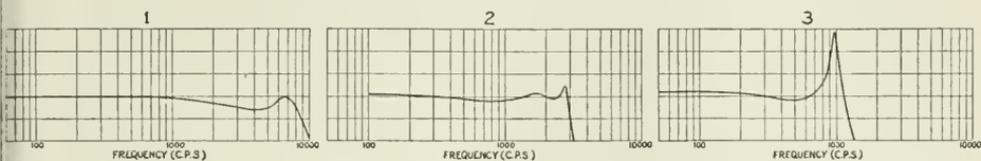
Fig. 9—String oscillograph characteristics—frequency response and indicial response with different amounts of damping.

The curves 7, 8, and 9 show that when the low frequencies are absent, the indicial response becomes too difficult to interpret. We must restrict our measurements to systems which are ideal at the low frequencies in order to interpret the indicial admittance by inspection.

Curves 10 and 11, Fig. 7, are low-pass filter characteristics, the former being a measured curve of a typical filter, while the latter is a calculated curve for an ideal filter. The two curves check reasonably well and indicate the effect of a very sharp cutoff as compared to those of the receivers shown

above. This indicates the oscillatory nature of any system having a sharp cutoff at the upper frequencies.

FREQUENCY RESPONSE OF TELEPHONE RECEIVERS



SQUARE WAVE RESPONSE

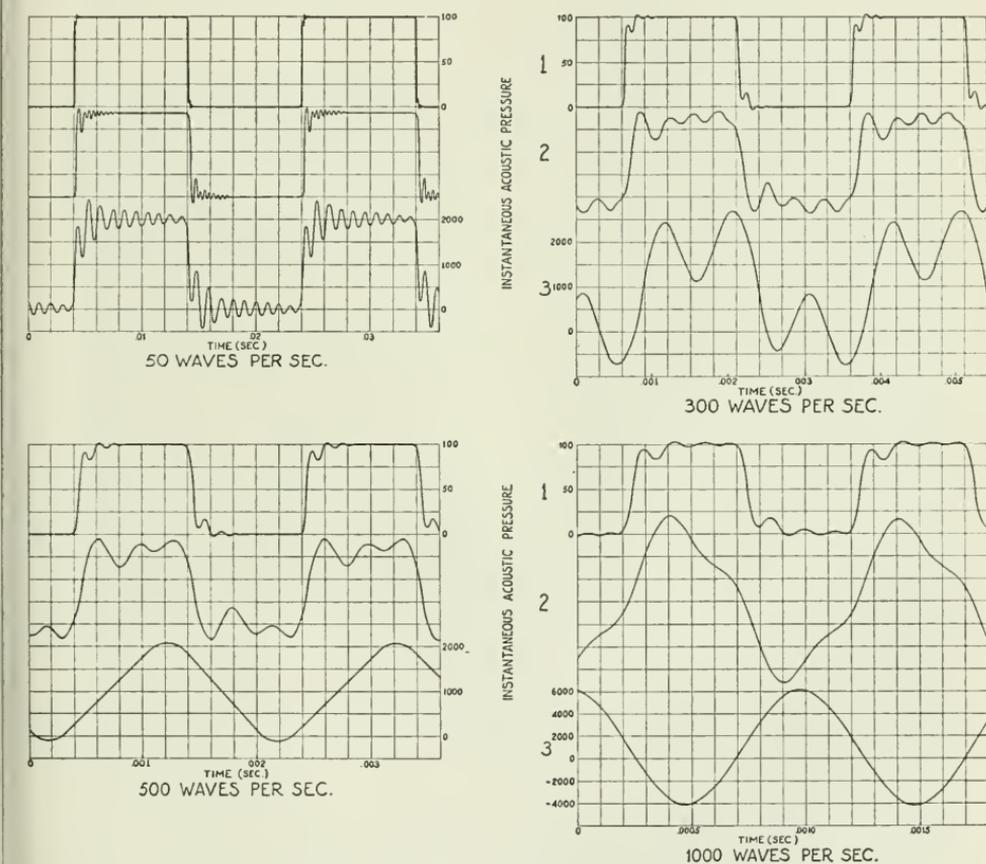


Fig. 10—Transient response to square waves of three different types of telephone receivers denoted by Nos. 1, 2 and 3, whose frequency response characteristics are shown above. Note the change in each type of pattern as the frequency of the square waves is increased.

Figure 8 shows a group of curves of the frequency response and indicial response of a group of receivers used as hearing aids. Curve 1 shows a very efficient but resonant receiver. Curve 2 is somewhat damped but still contains oscillations. Curve 3 is comparatively much better than either of the others from an indicial response viewpoint, and has a drooping frequency response characteristic, and demonstrates the advantages of this form of curve.

Figure 9 shows the effect of adding damping to the system of the string oscillograph when subjected to an ideal square wave. Curve 1, which has a virtually flat characteristic from 1 to 10,000 cps, is characterized by a sharp oscillatory peak in the indicial response. Curve 2 contains some oscillations, while curve 3 is substantially free from oscillations. The trend of these curves also shows the more faithful reproduction of transients obtained with a drooping frequency response.

Figure 10 shows the response to square waves of three receivers having different frequency response characteristics. The low-frequency waves of 50 cps are similar to the indicial response of the three receivers whose frequency characteristics are shown at the top, Fig. 10. As the frequency of these waves is increased to 300 cps, a noticeable departure from the square form is apparent in receiver No. 3. Receiver No. 2 shows a slight departure, while No. 1 is virtually a perfect reproduction.

As the frequency of the square waves is increased to 500 cps, the receiver No. 1 still shows very little departure from the original form. Receiver No. 2 maintains a fair approximation, while receiver No. 3 has lost all resemblance to the square form.

At a frequency of 1000 cps, only the first receiver maintains an approximately square form. Receivers Nos. 2 and 3 have both lost their identity and have become practically pure sinusoids. For all higher frequencies of the square waves, these two receivers will exhibit practically pure sinusoidal forms, due to the relatively sloping character of the frequency response at these frequencies, and the absence of harmonics. The same will be true of receiver No. 1 beyond a frequency of 3000 cps.

It will be realized, of course, that the patterns were obtained with square waves repeated at frequencies of 50, 300, 500 and 1000 cycles per second. While some speech waves approximate square waves in character such waves, when they occur, are repetitive only at the lower range of these frequencies. The above patterns were therefore obtained under conditions much more severe than are involved in the reproduction of speech waves and are included primarily for the purpose of illustrating the sensitivity of this form of analysis when applied to repeated square waves.

CONCLUSIONS

To summarize these data, it seems evident that square wave analysis may be applied in some fields of acoustics for both theoretical and practical applications.

In theory, the indicial response forms a somewhat different approach to the problem of obtaining the optimum characteristics of telephone receivers at the upper end of the frequency range. The greatest value of the square wave analysis lies in the fact that it gives us an entirely different conception of the behavior of an ideal sound system in terms of the unit function. The frequency response characteristic is ordinarily interpreted on the theory that any transient, such as an interval of conversation, may be represented by a Fourier series of sinusoidal frequencies of constant intensity lasting over the entire interval. If these equivalent component frequencies are to be reproduced in their true proportions, the ideal sound system must have mathematically uniform response for all single frequencies. On the other hand, the indicial response characteristic is judged from the Carson extension theorem, which shows that the more closely this characteristic approaches the unit function, the more perfect will be the reproduction of any given transient. Thus, the unit function and the sinusoid may be used as mutually complementary tools of analysis to show different aspects of the same type of problem.

In sound systems which are not ideal, due to inherent physical limitations, we tend to apply the Fourier Theorem out to a certain frequency, just as if it were an ideal system out to this frequency, and then beyond this frequency we do not attempt to sustain the higher frequencies. For most faithful reproduction of transients, it would seem that such practices might be altered somewhat to advantage by allowing the frequency response to drop off more gradually wherever it seems feasible to do so. The exact shape of the ideal curve under these circumstances is a matter of compromise between excessive delay on the one hand and excessive oscillations on the other. In practice, however, a fairly good picture is soon formed when curves such as the last in Figs. 6, 8, and 9 are found to approach the ideal more closely than those of other forms. Such listening tests as have been made tend to confirm these views, but cannot be regarded as being more than an indication.

Square wave analysis is somewhat limited in its practical applications to cases which may be interpreted by inspection. Systems having only a single cutoff frequency, or in the case of an additional low-end cutoff, ratios of the upper and lower cutoff frequencies f_2/f_1 of 100 or more, seem necessary to interpret the results by inspection.

The use of indicial response is not necessarily limited to any particular coupler or method of response measurement, since frequency response and indicial response are so closely related that one is a function of the other. The choice of a closed coupler measurement does, however, permit some interpretation of the results to be made by inspection, whereas other types of measurement may require laborious mathematical means to obtain an interpretation. Other types of vibration instruments, such as recorders, vibration pickups, crystal phonograph reproducers and carbon transmitters, which sustain their response down to zero frequency, should lend themselves to such methods of analysis.

In conclusion, the writer wishes to acknowledge the assistance of Mr. T. J. Pope in connection with the oscillographic work of this paper, and to express his sincere appreciation.

BIBLIOGRAPHY

1. Oliver Heaviside, "Electromagnetic Theory."
2. J. R. Carson:
 - a. "Transient Oscillations of Electrical Networks and Transmission Systems," *Trans. AIEE*, 1919, p. 445.
 - b. "Electric Circuit Theory and the Operational Calculus," McGraw-Hill.
- 3a. Gilbert Swift, "Amplifier Testing by Means of Square Waves," *Communications*, Vol. 19, No. 2, Feb. 1939.
- 3b. Bedford and Frehendahle, "Transient Response of Multi-Stage Video Frequency Amplifiers," *Proc. I. R. E.*, Vol. 25, No. 4, April 1939.
4. H. E. Kallman, "Portable Equipment for Observing Transient Response of Television Apparatus," *I. R. E. Proc.*, Vol. 28, No. 8, August 1940.
5. L. B. Arguimbau, "Network Testing with Square Waves," *General Radio Experimenter*, Vol. XIV, No. 7, Dec. 1939.
6. W. C. Jones, "Instruments for the New Telephone Sets," *B. S. T. J.* Vol. XVII, No. 3, p. 338, July 1938.
7. V. Bush, "Operational Circuit Theory," Wiley and Sons, p. 176.
8. F. F. Romanow, "Methods for Measuring the Performance of Hearing Aids," *Acous. Soc. Am. Jour.*, Vol. 13, p. 294, Jan., 1942.
9. A. M. Curtis, "A Oscillograph for Ten Thousand Cycles," *B. S. T. J.*, Vol. XII, No. 1, January 1933.

CHAPTER VII

Theoretical Analysis of Modes of Vibration for Isotropic Rectangular Plates Having All Surfaces Free

By H. J. McSKIMIN

7.1. INTRODUCTION

The comparatively recent advent of crystal controlled oscillators and of wave filters employing piezoelectric elements has resulted in an extensive study of the ways in which plates made of elastic materials such as quartz or rochelle salt can vibrate. Of special interest have been the resonant frequencies associated with these modes of motion. As will be indicated in subsequent paragraphs, the general solution to the problem of greatest interest is quite complex, and has not been forthcoming, (i.e., as applied to rectangular plates completely unrestrained at all boundary surfaces). For this reason numerous approximate solutions have been developed which yield useful information in spite of their limitations. Several of these solutions will be discussed in the following sections. The three general types of modes (i.e., the extensional, shear, and flexural) will be analyzed in some detail. Also, as a preliminary step the formulation of the general problem along classical lines will be developed.

For the most part, the solutions obtained here are limited to those for an isotropic body. However, such solutions provide considerable guidance for the modes of motion existing in an aeolotropic body such as quartz.

7.2. METHOD OF ANALYSIS

In order to set up the desired mathematical statement of our problem it will be necessary to consider first of all two very fundamental relationships. The first of these is the well known law of Newton which states that a force f acting on a mass m produces an acceleration a in accordance with the formula

$$f = m \cdot a$$

The second relationship which we shall need is Hooke's law relating the strains in a body to the stresses. If forces are applied to the ends of a long slender rod made of an elastic material such as steel (Fig. 7.1) a certain amount of stretching takes place. If the forces are not too great, a linear

relationship between the applied stress and ensuing strain is found to exist. Expressed as an equation

$$\frac{X_x}{x_x} = E \quad \text{in which } X_x \text{ is the force per unit area,}$$

x_x is the strain per unit length, and E is a constant known as Young's Modulus. (Refer to Section 7.7 for further definition of terms).

In an analogous manner, shearing stresses applied to an elastic solid as shown by Fig. 7.2 produce a shearing strain such that

$$\frac{X_y}{x_y} = A, \quad \text{the shear modulus.}$$

In general there will be contributions to a particular strain from any of the stresses which may happen to exist. For example, when an isotropic

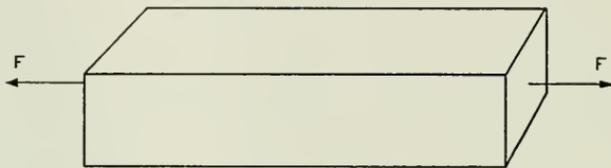


Fig. 7.1—Bar under tensional stress

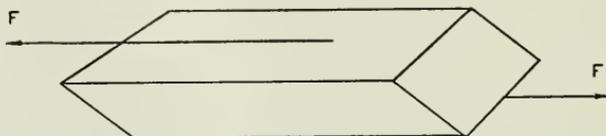


Fig. 7.2—Bar under shearing stress

bar is stretched, there will be a contraction along the width which has been produced by a stress along the length. A statement of these relationships (known as Hooke's Law) is given by the equations of Section 7.8.

It is now of interest to consider the conditions of equilibrium for a very small cube cut out of the elastic medium which in general is stressed and in motion. Reference to Fig. 7.3 will help to visualize the stresses which may exist on the faces of this cube. Since these stresses vary continuously within the medium, a summation of the forces acting on the cube along each of the major axes can be made with the use of differential calculus. From Newton's Law previously cited, it is apparent that any unbalance of these forces will result in an acceleration inversely proportional to the mass of our small cube. Three equations may then be derived, one for each major direction.¹ If only simple harmonic motion is considered (i.e. all displace-

¹ Refer to "Theory of Elasticity" by S. Timoshenko or to any standard text on elasticity.

ments are proportional to $\sin \omega t$ where $\omega = 2\pi$ times frequency) the following simplified equations result.

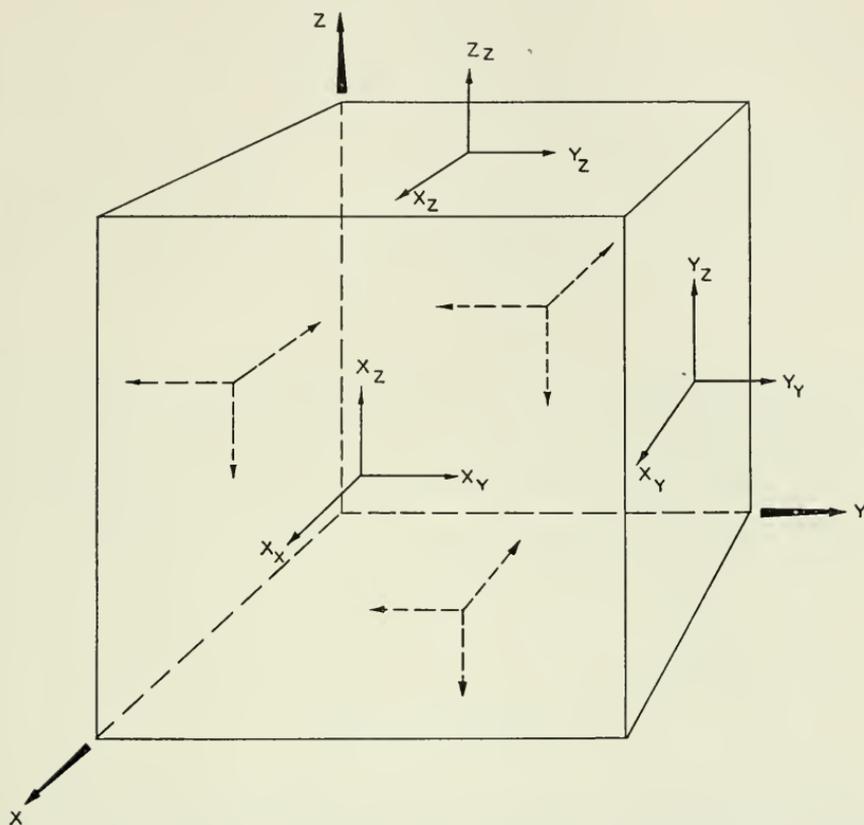


Fig. 7.3—Stresses acting on small cube

$$\left. \begin{aligned} \frac{\partial X_x}{\partial x} + \frac{\partial X_y}{\partial y} + \frac{\partial X_z}{\partial z} &= -\rho\omega^2 u \\ \frac{\partial Y_y}{\partial y} + \frac{\partial X_y}{\partial x} + \frac{\partial Y_z}{\partial z} &= -\rho\omega^2 v \\ \frac{\partial Z_z}{\partial z} + \frac{\partial X_z}{\partial x} + \frac{\partial Y_z}{\partial y} &= -\rho\omega^2 w \end{aligned} \right\} \quad (7.1)$$

Since stresses are related to strains in a very definite manner, the above equations may be converted into a more useful form involving only displacements. For *isotropic media*, the following results.

$$\left. \begin{aligned} A\nabla^2 u + B \frac{\partial \epsilon}{\partial x} &= -\rho\omega^2 u \\ A\nabla^2 v + B \frac{\partial \epsilon}{\partial y} &= -\rho\omega^2 v \\ A\nabla^2 w + B \frac{\partial \epsilon}{\partial z} &= -\rho\omega^2 w \end{aligned} \right\} \quad (7.2)$$

In this grouping,

$$\begin{aligned} \nabla^2 &= \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \\ \epsilon &= \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} \end{aligned}$$

and A and B are given in terms of the fundamental elastic constants λ and μ with $A = \mu$, $B = \lambda + \mu$.

An even more elegant statement of the equilibrium conditions attributable to Love² follows immediately from equations 7.2, since by differentiating each one in turn with respect to x , y , and z respectively, and then adding results, one obtains the wave equation

$$(\nabla^2 + h^2)\epsilon = 0 \quad (7.3)$$

where

$$h^2 = \frac{\rho\omega^2}{A + B} = \frac{\rho\omega^2}{\lambda + 2\mu}$$

Whatever our solution may be, then, it must satisfy equation (7.3). If such an expression for ϵ is found, the displacements formed in the following way will satisfy equations 7.2 as can be shown by direct substitution.

$$u = -\frac{1}{h^2} \frac{\partial \epsilon}{\partial x} \quad v = -\frac{1}{h^2} \frac{\partial \epsilon}{\partial y} \quad w = -\frac{1}{h^2} \frac{\partial \epsilon}{\partial z} \quad (7.4)$$

In addition to equations 7.2, another set of requirements will be necessary when any particular problem is considered. They are known as the boundary conditions, and in general are easily deduced from a knowledge of how the plate or bar is held.

For a rectangular plate free on all surfaces, the boundary condition is simply that all surface tractions vanish. This requires certain stresses to become zero at the boundary as can be seen from the following expressions for the x , y , and z components of traction in terms of unit stresses.

² A. E. Love, "A Treatise on the Mathematical Theory of Elasticity."

$$\left. \begin{aligned} \bar{X} &= X_x \ell + X_y m + X_z n \\ \bar{Y} &= Y_x m + Y_z n + X_y \ell \\ \bar{Z} &= Z_x n + X_z \ell + Y_z m \end{aligned} \right\} = 0 \text{ for free surfaces} \quad (7.5)$$

(ℓ , m , and n are direction cosines of the normal to the surface at the point in question).

The general problem is now seen to be one of finding solutions for the displacements u , v , and w such that both the equilibrium and boundary conditions are satisfied. In the following section several interesting solutions will be considered for rectangular plates having all surfaces free, this being the case of greatest interest in so far as this paper will be concerned.

7.3. EXTENSIONAL VIBRATIONS

One of the most useful modes of vibration of practical interest is the extensional, in which particle motion takes place in essentially one direction so as to alternately stretch and compress the elastic medium. Piezoelectric

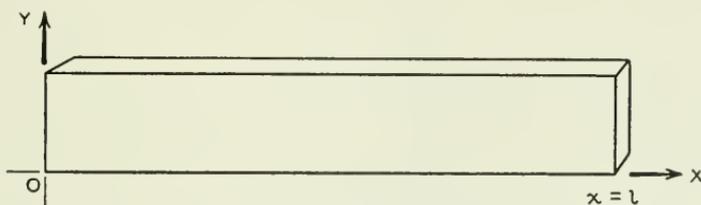


Fig. 7.4—Longitudinal bar

plates vibrating in this manner, and of the shapes shown in figures 7.4 and 7.5 have been used extensively in wave filter and oscillator circuits. The approximate resonant frequencies corresponding to this type of motion are easily obtained by a consideration of equations 7.1 and 7.2. For the longitudinal bar of Fig. 7.4 the only stress that need be considered is the X_x extensional, all other stresses being so small that they can be neglected. The equilibrium equation then becomes

$$\frac{\partial X_x}{\partial x} = -\rho \omega^2 u \quad (7.6)$$

or, since

$$\begin{aligned} \frac{\partial u}{\partial x} &= \frac{1}{E} X_x \\ \frac{\partial^2 u}{\partial x^2} &= -\frac{\rho}{E} \omega^2 u \end{aligned} \quad (7.7)$$

It is easily seen that $u = \cos kx$ is a solution to this equation if $k = \omega \sqrt{\frac{\rho}{E}}$.

If now the boundary condition that the stress X_x must become zero at the ends of the bar (i.e., $x = 0$, $x = \ell$ — refer to Eq. (7.5)), is fulfilled, the solution will be complete. At $x = 0$, $X_x = E \frac{\partial u}{\partial x}$ will always equal zero.

Furthermore, if $k = \frac{\pi}{\ell}$ or any whole number multiple of $\frac{\pi}{\ell}$ the extensional stress will likewise reduce to zero at $x = \ell$. The desired solution will then be as follows, f being the resonant frequencies.

$$\left. \begin{aligned} u &= \cos \omega \sqrt{\frac{\rho}{E}} x \\ \omega &= 2\pi f = \frac{m\pi}{\ell} \sqrt{\frac{E}{\rho}} \\ m &= 1, 2, 3, \text{ etc.} \end{aligned} \right\} \quad (7.8)$$

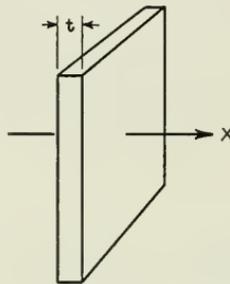


Fig. 7.5—Thin plate

The plate of Fig. 7.5 will now be considered. Here it can no longer be assumed that the X_x stress is the only one of importance. Instead, the displacements v and w will be considered zero and the displacement u a function of x only. This means that the shear stresses X_y , X_z , Y_z vanish, so that the equilibrium equations 7.2 reduce to

$$A \frac{\partial^2 u}{\partial x^2} + B \frac{\partial^2 u}{\partial x^2} = -\rho \omega^2 u \quad (7.9)$$

or

$$\frac{\partial^2 u}{\partial x^2} = \frac{-\rho \omega^2 u}{A + B} \quad (7.10)$$

This is seen to be of the same form as equation (7.7) previously discussed, and will again have the solution $u = \cos kx$ with $k = \omega \sqrt{\frac{\rho}{A + B}}$. The

boundary condition on the X_x stress will be met if $k = \frac{m\pi}{t}$ so that the following solutions result.³

$$\begin{aligned}
 u &= \cos \omega \sqrt{\frac{\rho(1 + \sigma)(1 - 2\sigma)}{E(1 - \sigma)}} x = \cos \omega \sqrt{\frac{\rho}{\lambda + 2\mu}} a \\
 \omega &= 2\pi f = \frac{m\pi}{t} \sqrt{\frac{E}{\rho} \cdot \frac{(1 - \sigma)}{(1 + \sigma)(1 - 2\sigma)}} = \frac{m\pi}{t} \sqrt{\frac{\lambda + 2\mu}{\rho}} \quad (7.11) \\
 m &= 1, 2, 3, \text{ etc.}
 \end{aligned}$$

It is seen that this formula for resonant frequencies is the same as given by equations 7.8, with E replaced by $\frac{E \cdot (1 - \sigma)}{(1 - 2\sigma^2 - \sigma)}$, so that the frequency constant $f \cdot t$ will be somewhat higher than $f \cdot \ell$ for a long slender bar.

It is recognized that the solutions derived above hold true only for the limiting cases of a long slender bar, and a very thin plate respectively. It is therefore of interest to trace the resonant frequencies corresponding to these extensional modes of vibration as departure is made from the limiting cases mentioned above.

An experimental plot of the resonant frequencies of a thin plate of length ℓ and width w reveals that the frequency of the longitudinal mode first discussed is gradually lowered as the width of the plate is increased. There is also another frequency corresponding to an extensional vibration along the width which for a very narrow plate corresponds to the second type of extensional mode considered in the foregoing paragraphs, except that the frequency constant will be slightly different because coplanar stresses are involved.

As seen from Fig. 7.6, the resonant frequency curves do not cross, but exhibit coupling effects. This is understandable from the fact that motion in one direction is mechanically coupled to motion in the other as indicated by Poisson's ratio σ .

In order to derive expressions for the u and v displacements associated with the extensional mode along the length, taking into account the above coupling effect, the following analysis proves interesting.

Consider the infinite isotropic strip of width b as shown by figure 7.7. As will be demonstrated presently, solutions can be found such that the equilibrium equations and the boundary conditions are precisely satisfied. Furthermore it will be found possible to cut a section out of this strip in

³ If the length and width of the plate are very large in comparison to the thickness, the boundary conditions for the Y_y and Z_z stresses may be neglected without causing appreciable error. The quantity $A + B$ has been evaluated in terms of E and σ for purposes of comparison.

such a way that the boundary conditions for the cut edges are very nearly satisfied. The plate formed in this way may then be considered as vibrating at the required frequency f , which will then be the resonant frequency desired.

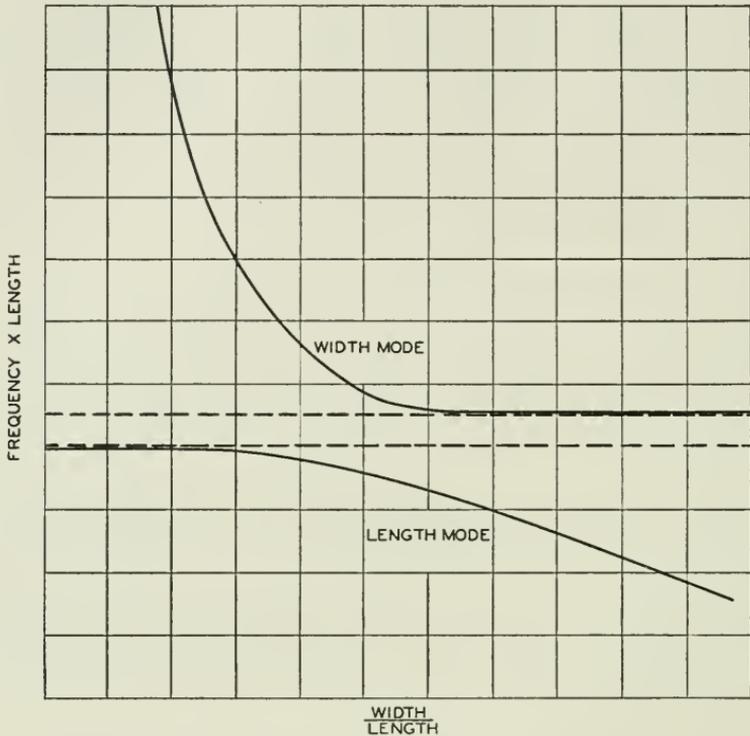


Fig. 7.6—Extensional modes with mechanical coupling

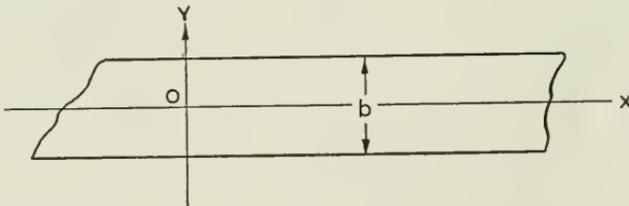


Fig. 7.7—Infinite strip

Let displacements be arbitrarily chosen in the following way:

$$\left. \begin{aligned} u &= U \cos kx \cos \ell_y \\ v &= V \sin kx \sin \ell_y \end{aligned} \right\} \quad (7.12)$$

As shown in Section 7.9, two solutions of this type will satisfy the equilibrium equations precisely. One corresponds to $\epsilon = 0$ in the wave equation 7.3, while for the other $\epsilon \neq 0$. Superposition of these two solutions and proper evaluation of parameters make it possible to satisfy the boundary conditions at the edge of the strip; namely, that at $y = \pm \frac{b}{2}$, $Y_y = 0$ and $X_y = 0$. (Refer to equations 7.5). The following transcendental equation is obtained

$$\frac{\cot \ell_1 \frac{b}{2}}{\cot \ell_2 \frac{b}{2}} = \frac{-2\ell_1 \ell_2 k^2 (1 - \sigma)}{(\ell_2^2 - k^2)(\ell_1^2 + \sigma k^2)} \tag{7.13}$$

in which

$$\left. \begin{aligned} \ell_1^2 &= \frac{A}{A+B} \theta^2 - k^2 \\ \ell_2^2 &= \theta^2 - k^2 \\ \theta^2 &= \frac{\rho \omega^2}{A} \end{aligned} \right\} \tag{7.14}$$

This equation may be solved graphically to yield values of frequency corresponding to given values of k . For our discussion of the length extensional mode of vibration, the first root only will be considered.

Fig. 7.8 shows a plot of $\theta \cdot b$ against $b \cdot k$ assuming that Poisson's ratio is .33.⁴ If $k = 1$, and $b = 1$, for example, $\theta = \sqrt{\frac{\rho}{A}} \omega = 1.62$.

The equations for the displacements when determined as explained in Section 7.9 become:

$$\begin{aligned} u &= U_1 [\cosh .344 y + .402 \cos 1.278 y] \cos x \\ v &= U_1 [.344 \sinh .344 y + .315 \sin 1.278 y] \sin x \end{aligned} \tag{7.15}$$

All three stresses X_x , Y_y , and X_y may be calculated from the above equations. If the length of our plate is made equal to $m\pi$, where m is an integer, the extensional stress X_x will equal zero regardless of y at the boundaries $x = 0$ and $x = \ell$ since $X_x \propto \sin x = 0$ when $x = m\pi$. Also it can be shown by calculation that X_y is so small in comparison to the extensional stresses as to be entirely negligible; hence our solution is complete.

If $\ell = \pi$, the plate will be vibrating in its fundamental longitudinal mode. The distortion which results is shown by Fig. 7.9. It is seen that most of

⁴ Plotted in this way, the same curve results regardless of the value of b chosen for the purpose of solving Eq. 7.13.

the motion is along the x axis, though there is a certain amount of lateral contraction as the plate elongates.

The second harmonic will have the same resonant frequency if $\ell = 2\pi$, the third if $\ell = 3\pi$, etc.

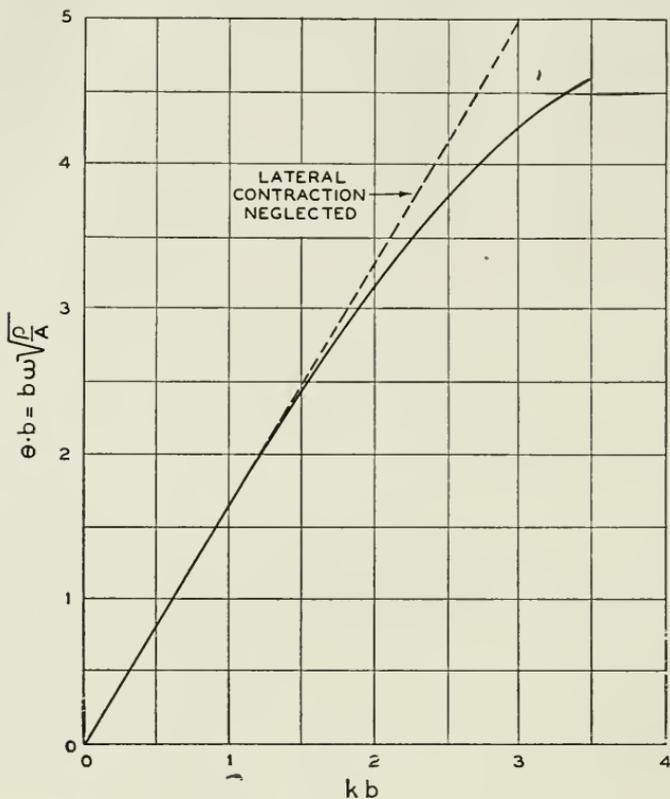


Fig. 7.8— $\theta \cdot b$ versus $k \cdot b$ for plate longitudinal modes ($\sigma = .33$, $k = \frac{m\pi}{\text{length}}$)

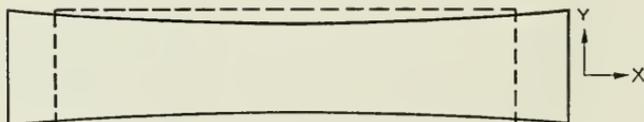


Fig. 7.9—Distortion of plate vibrating in first longitudinal mode

In addition to harmonic modes along the length just considered there will be those for which the motion breaks up along the width. In general, the distortion of the plate may be quite complex with simultaneous variations along both dimensions. Similarly, for plates such as shown in Fig. 7.5

there will be many extensional modes which have resonant frequencies somewhat above those given by Eq. 7.11. Analysis of the motion shows that for these modes the displacement along the thickness varies periodically (or “breaks up”) along the major dimensions of the plate. There again the distortion pattern of the plate may become very complex.

7.4. SHEAR VIBRATIONS

The second class of vibrations which will now be considered is the shear. This type of mode is of special importance because of the fact that piezo-electric plates vibrating in shear are widely used for frequency control of oscillators. For example, the AT quartz plate which is so much in demand utilizes a fundamental thickness shear mode in which particle motion is principally at right angles to the thickness. The distortion of the plate will be similar to that shown in Fig. 7.2.

A simple, yet very useful formula for the resonant frequencies associated with the above type of displacement has been derived on the assumption

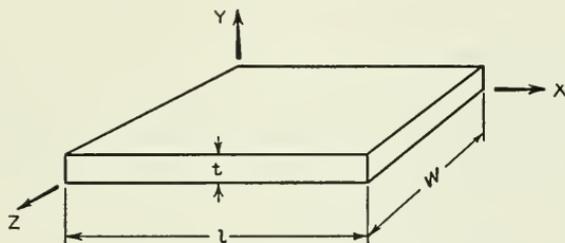


Fig. 7.10—Orientation of thin plate

that the length and width of the plate are very large in comparison to the thickness. For the xy shear mode, the displacement u is assumed to be $u = U \cos ky$, all other displacements being equal to zero. The only stress that need be considered then, is the X_y shear which is proportional to $\sin ky$. Boundary conditions on this stress at the major surfaces of the plate are easily satisfied by choosing k such that $X_y = 0$ at $y = 0$ and $y = t$. (Refer to Fig. 7.10.) This will be the case if $k = \frac{m\pi}{t}$, where m is any integer, and t is the thickness of the plate. By using the simplified equilibrium equation as reduced from equations 7.1, a formula for the resonant frequencies is obtained in much the same manner as for extensional thickness modes.

$$\omega = 2\pi f = \frac{m\pi}{t} \sqrt{\frac{A}{\rho}} \quad m = 1, 2, 3, \text{ etc.} \quad (7.16)$$

In this formula the shear modulus A appears instead of Young’s modulus as in the case of longitudinal modes. Harmonic modes are given by values of m greater than unity.

In addition to the resonant frequencies predicted by the foregoing analysis, there will be others corresponding to shear vibrations in which the principal shear stress varies periodically along the length and width of the plate. A formula which yields the approximate frequencies for these modes is developed in Section 7.9. It is shown that if the length and width are large in comparison to the thickness, the following expression may be used:

$$\omega = 2\pi f = \pi \sqrt{\frac{1}{\rho}} \sqrt{c_{11} \frac{n^2}{\ell^2} + \frac{c_{66} m^2}{l^2} + \frac{c_{55} p^2}{w^2}} \quad (7.17)$$

In this formula which has been derived for xy shears the c constants are the standard elastic constants for aeotropic media. For isotropic plates such as have been considered up to this point

$$c_{11} = \frac{E(1 - \sigma)}{1 - 2\sigma^2 - \sigma} = \lambda + 2\mu$$

and

$$c_{55} = c_{66} = A, \text{ the shear modulus} \quad (7.18)$$

Various combinations of the integers m , n , and p may be chosen, with the restriction that neither m nor n can equal zero. It is seen that if ℓ and w are very large the formula reduces to that of Eq. 7.16 which was derived on precisely that basis. Also, it is seen that the more complex modes all lie somewhat above the fundamental shear obtained by setting $m = n = 1$ and $p = 0$.

Plate shear modes are also of considerable interest, particularly the one of lowest order. For a plate having a large ratio of length to width a formula similar to that given by equation 7.17 (but for two dimensions only) may be developed. If the plate is nearly square, however, this formula no longer yields sufficiently accurate values for the resonant frequencies. Coupling to other modes of motion⁵ complicates the problem so much that only experimental results have been of much practical consequence. Fig. 7.11 shows in an exaggerated way the distortion of a nearly square plate vibrating in the first shear mode.

7.5. FLEXURAL VIBRATIONS

7.51. Plate Flexures

One of the most studied types of vibrations has been the flexural. Perhaps this is true because it is the most apparent and comes within the realm of experience of nearly everyone. The phenomena of vibrating reeds, xylophone bars, door bell chimes, tuning forks, etc. are quite well known.

⁵ It is found experimentally that odd order shears are strongly coupled to even order flexures; similarly, even order shears and odd order flexures are coupled.

Beam theory has been used quite extensively to derive the equations which yield the resonant frequencies and displacements for bars vibrating in flexure. To obtain reasonably accurate results for ratios of width to length approaching unity, however, the effects of lateral contraction, rotary inertia, and shearing forces must be considered. This leads to a rather complicated solution which is much more accurate than that derived by the use of simple beam theory only, though it is still approximate in nature.

For two dimensional plates free on all edges a method of analysis may be used which is similar to that described under extensional modes. While it is somewhat involved it yields direct expressions for the two displacements u and v , so that all stresses may be calculated, and the extent to which boundary conditions are satisfied determined.⁶

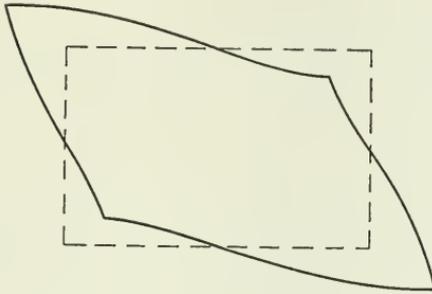


Fig. 7.11—Distortion of plate in first shear mode

Solutions for u and v are assumed to be of the form

$$\begin{aligned} u &= U \sin \ell y \cos kx \\ v &= V \cos \ell y \sin kx \end{aligned} \tag{7.19}$$

For the infinite strip previously considered a transcendental equation is obtained which is the same as equation 7.13 with the exception that the left-hand expression is inverted.

$$\frac{\tan \ell_1 \frac{b}{2}}{\tan \ell_2 \frac{b}{2}} = \frac{-2\ell_1 \ell_2 k^2 (1 - \sigma)}{(\ell_2^2 - k^2)(\ell_1^2 + \sigma k^2)} \tag{7.20}$$

(Refer to Eq. 7.14 also.)

⁶ This is an extension of Doerfler's analysis used to obtain harmonic flexure frequencies for plates—"Bent and Transverse Oscillations of Piezo-Electrically Excited Quartz Plates"—*Zeitschrift Für Physik*, v. 63, July 7, 1930, p. 30. Also refer to "The Distribution of Stress and Strain for Rectangular Isotropic Plates Vibrating in Normal Modes of Flexures"—New York Univ. Thesis by Author, June 1940.

The lowest order solution to this equation is found to correspond to flexure vibrations in the infinite strip. A calculation of stresses, however, reveals that boundary conditions cannot be satisfied properly even for the case of a long narrow plate. It can be shown, however, that another solution may be derived for the same value of frequency by letting k become imaginary. This simply means that the u and v displacements become hyperbolic functions of x instead of sinusoidal. The two complete solutions for the infinite strip may then be superimposed and parameters adjusted so that for definite values of length corresponding to fundamental and harmonic modes the proper stresses reduce essentially to zero on the ends of the plate. For plates having a ratio of width to length less than .5, this method gives very accurate expressions for displacements and stresses. If only the resonant frequency is required, ratios up to unity and beyond (for the fundamental mode) may be considered.

An example has been worked out to provide a complete picture of the displacements for a bar of width = 1, $k = 1$ and $\sigma = .33$. Use of equation 7.20 yields the quantity $\theta^2 = \frac{\rho\omega^2}{A} = .166$ from which the resonant frequency may be obtained. Using this value of θ^2 , one finds that $k^2 = -.800$ also satisfies equation 7.20. By making the total length of the bar equal to 4.50 the X_x extensional stress and the X_y shear stress may be made essentially zero on the ends of the plate regardless of y .⁷

The following expressions for u and v are obtained:

$$\begin{aligned} u &= (\sinh .9132 y - 1.02 \sinh .9718y) \sin x \\ &\quad - .160 (\sin .9828y - .9568 \sin .9250y) \sinh .8944x \\ v &= (-1.094 \cosh .9132y + .9915 \cosh .9718) \cos x \\ &\quad - .160 (.9095 \cos .9828y - .990 \cos .9250y) \cosh .8944x \end{aligned} \tag{7.21}$$

Fig. 7.12 shows the distortion of the plate as calculated from the above expressions. It is seen that there will be two points at which there is no motion in either the x or y directions. These nodal points can be used in holding the plate, since it may be clamped firmly there without altering the displacements or resonant frequency. For the example shown, these nodes are positioned a distance of $.211\ell$ from the ends of the plate as compared to $.224\ell$ for a long thin bar.

⁷ A graphical solution to determine ℓ is most convenient in which parameters are adjusted so that $X_x = 0$ at $x = \pm \frac{\ell}{2}$ and $y = \pm \frac{b}{2}$; $X_y = 0$ at $x = \pm \frac{\ell}{2}$ and $y = 0$. These stresses will remain essentially zero for all values of y if the ratio of $\frac{w}{\ell}$ is not too great.

Figures 7.13 and 7.14 show the distribution of the principle stresses as a function of position along the length. It is seen that for the particular

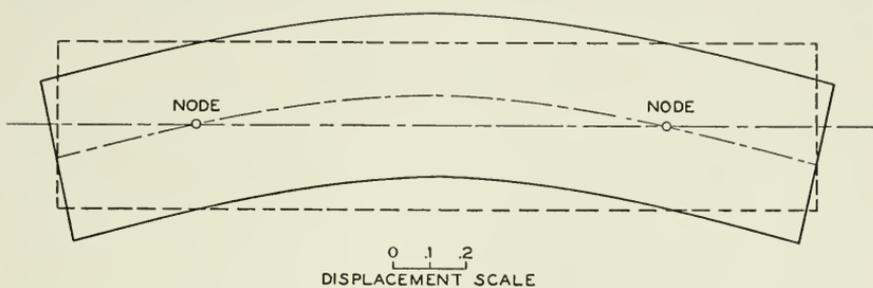


Fig. 7.12—Distortion of bar vibrating in first free-free flexure mode

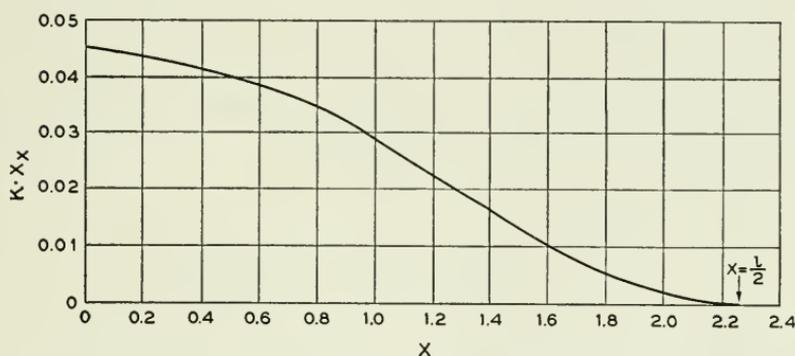


Fig. 7.13—Distribution of longitudinal stress for free-free bar vibrating in first flexure mode

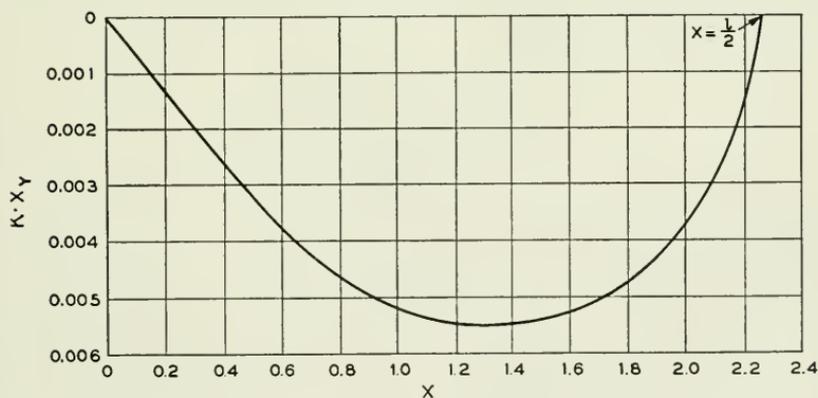


Fig. 7.14—Distribution of shear stress for free-free bar vibrating in first flexure mode

example cited, the maximum shear stress is only about one-tenth the maximum X_x extensional stress. Both of these stresses reduce to zero at

the ends of the plate as they should in order to satisfy the boundary conditions. As the ratio of $\frac{w}{l}$ is increased the shear stress becomes of greater importance.

7.52. Thickness Flexures

The final analysis to be considered in this paper is for thickness flexures along the width or length of a thin plate. These modes are of particular interest in connection with the dimensioning of quartz plates for which it is desirable to utilize the fundamental thickness shear mode. (AT plate, for example.) It is found experimentally that even ordered thickness flexures are coupled to this shear to such a degree that at certain ratios of dimensions the operation of the plate as an oscillator or filter component is impaired.

The two-dimensional solution derived in the preceding paragraphs can be used to predict certain harmonic thickness flexures; however, in order to obtain a complete picture it is necessary to extend the theory to three dimensions. This has been done by the author with the following transcendental equation as a result (refer to Section 7.93).

$$\frac{\tan \ell_1 \frac{b}{2}}{\tan \ell_2 \frac{b}{2}} = \frac{-2\ell_1 \ell_2 A \alpha^2}{[\sigma B(\ell_1^2 + \alpha^2) + A\ell_1^2][\ell_2^2 - \alpha^2]} \quad (7.22)$$

Solutions to this equation are exact in nature for a plate of thickness b and of infinite extent in both the x and z directions. The quantity α^2 is equal to the sum of the squares of k and m which appear in the expressions for displacements as follows:

$$\left. \begin{aligned} u &= U f_1(y) \sin kx \cos mz \\ v &= V f_2(y) \cos kx \cos mz \\ w &= W f_3(y) \cos kx \sin mz \end{aligned} \right\} \quad (7.23)$$

Also in equation (7.22)

$$\left. \begin{aligned} \ell_1^2 &= \theta^2 \frac{A}{A+B} - \alpha^2 \\ \ell_2^2 &= \theta^2 - \alpha^2 \\ \theta^2 &= \frac{\rho \omega^2}{A} \end{aligned} \right\} \quad (7.24)$$

The lowest order solution to equation (7.22) with α^2 positive again corresponds to flexure vibrations, as in the two dimensional case. Fig. 7.15 shows a plot of $\theta \cdot b$ against $\alpha \cdot b$ calculated for $\sigma = .3$.

For reasonably high order flexures it may be reasoned that the true displacements will be very nearly the same as those for the doubly infinite plate as derived by the above method since the correction necessary to fulfill the boundary conditions will only apply very close to the edges of the plate. It will then be sufficient to choose values for k and m such that $k = p \frac{\pi}{\ell}$ and $m = \frac{q\pi}{w}$ where p and q are integers. The values of α^2 obtained in this way determine the corresponding resonant frequencies.

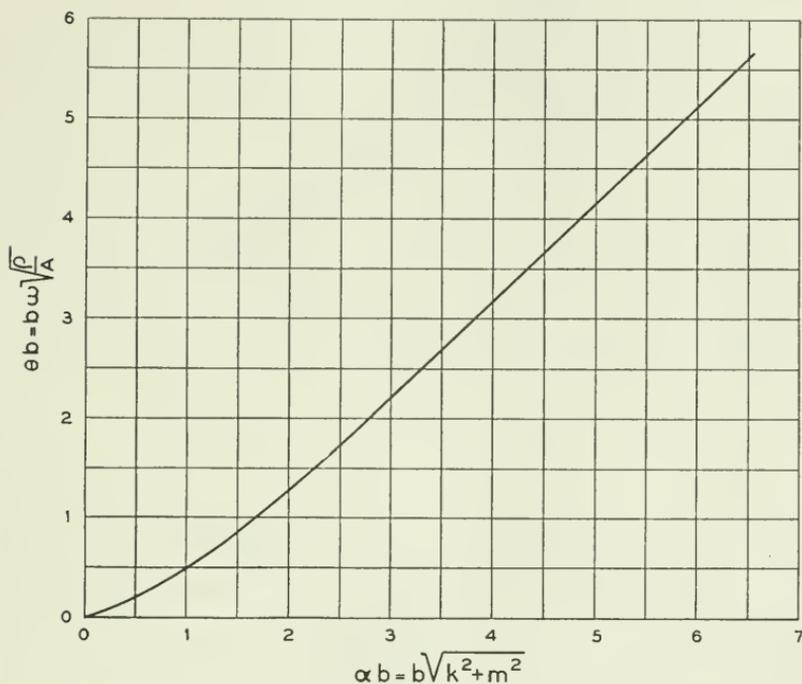


Fig. 7.15— $\theta \cdot b$ versus $\alpha \cdot b$ for thickness flexures

If it is desired to solve for the ordinal xy flexures, for example, m should be set equal to zero. The displacements in this case will be independent of the z dimension. When q is assigned values other than zero however, the resulting modes may be considered as xy flexures which vary or break up along the third major dimension. If q is small the resonant frequencies will lie only slightly higher than that of the corresponding ordinal flexure for which $q = 0$.

Fig. 7.16 shows a few of the resonant frequencies as calculated for values of shear modulus and density corresponding to AT quartz. The effects of coupling to the fundamental thickness shear are shown by dotted lines for the 14th xy flexure. As might be expected there is similar coupling between

the 14th flexure which breaks up once along the z dimension and the shear which breaks up once along z —etc. A few of these flexures which break up along z are shown for the 16th ordinal flexure.

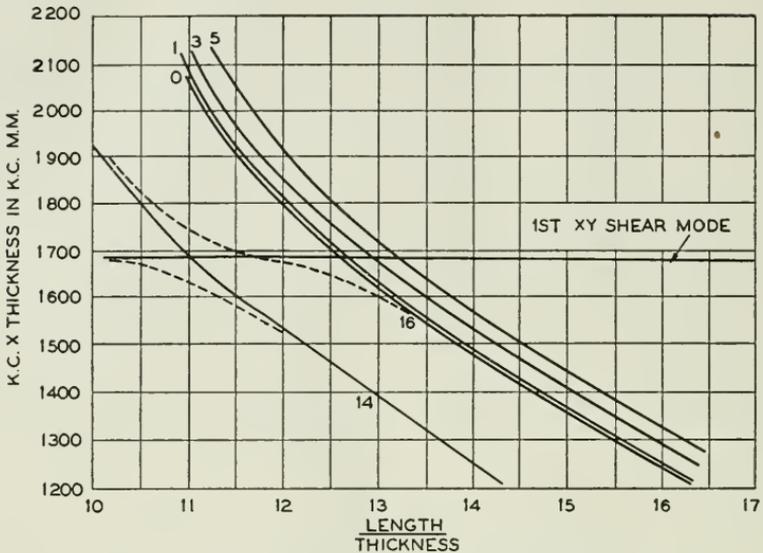


Fig. 7.16— XY thickness flexure modes for square plate

7.6. SUMMARY

Three main classes or families of vibrational modes are found to exist in rectangular elastic plates free on all surfaces; namely, the extensional, the shear, and the flexural. In general, the associated displacements are functions of all three dimensions and may vary in such a manner as to make the distortion of such plates quite complex.

For certain limiting cases, approximate solutions for the resonant frequencies and displacements (from which strains and stresses may be calculated) can be derived. Though there are a number of methods that can be used for specific problems, it has been found very convenient to utilize the classical formulation. For this reason the basis of this method has been discussed briefly. In essence it requires that displacements and stresses occurring within the elastic solid satisfy conditions of equilibrium as derived from Newton's Law. At the boundaries, certain other relations must be satisfied in order that conditions of clamping might be fulfilled. For plates entirely unrestrained the latter requires that all forces (tractions) acting through the free surfaces must vanish.

For thin rectangular plates (such as quartz crystal oscillator plates) the modes of greatest practical consequence are plate modes, for which all

stresses are essentially coplanar and independent of the thickness, and thickness modes, for which all dimensions must be considered except in limiting cases.

Because of their great utility, simplified formulae have been derived for the resonant frequencies associated with long, narrow bars vibrating longitudinally, thin plates with extensional motion along the thickness dimension, and thin plates vibrating with shearing motion at right angles to the thickness.

Exact solutions for the infinite strip have been derived, and used in obtaining the displacements and resonant frequencies for flexural and longitudinal modes. Such solutions take account of the fact that the width of the plate may become appreciable. While limiting cases of plate shear may be analyzed, solutions for ratios of $\frac{w}{l}$ approaching unity have not proved very satisfactory. This is attributable to the fact that coupling to flexural modes is severe.

Thickness flexural modes which exhibit displacement variations along both length and width dimensions of the plate have been analyzed by extending the "infinite strip" theory to three dimensions. Solutions obtained are fairly accurate if the harmonic order of the flexure is sufficiently great.

7.7. NOMENCLATURE

ρ = density

E = Young's modulus

σ = Poisson's ratio

A = Shear modulus = $\frac{E}{2(1 + \sigma)} = \mu$

$B = \frac{E}{2(1 + \sigma)(1 - 2\sigma)} = \lambda + \mu$ for 3 dimensions
 $= \frac{E}{2(1 - \sigma)}$ for plane stress

ω = angular velocity = $2\pi f$

$\theta^2 = \frac{\rho\omega^2}{A}$

u, v, w = displacements in x, y and z directions

$\epsilon = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z}$

$$\nabla^2 = \text{Laplacian} = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$$

$\left. \begin{array}{l} x_x, y_y, z_z \\ x_y, x_z, y_z \end{array} \right\}$ unit strain components

$\left. \begin{array}{l} X_x, Y_y, Z_z \\ X_y, X_z, Y_z \end{array} \right\}$ unit stresses

7.8. STRESS-STRAIN EQUATIONS FOR ISOTROPIC MEDIA

$$x_x = \frac{1}{E} (X_x - \sigma Y_y - \sigma Z_z)$$

$$y_y = \frac{1}{E} (Y_y - \sigma X_x - \sigma Z_z)$$

$$z_z = \frac{1}{E} (Z_z - \sigma X_x - \sigma Y_y)$$

$$\text{shear strain} = \frac{1}{A} \times \text{shear stress}$$

$$X_x = 2 \left(\sigma B \epsilon + A \frac{\partial u}{\partial x} \right)$$

$$Y_y = 2 \left(\sigma B \epsilon + A \frac{\partial v}{\partial y} \right)$$

$$Z_z = 2 \left(\sigma B \epsilon + A \frac{\partial w}{\partial z} \right)$$

$$X_y = A \left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \right)$$

$$X_z = A \left(\frac{\partial u}{\partial z} + \frac{\partial w}{\partial x} \right)$$

$$Y_z = A \left(\frac{\partial v}{\partial z} + \frac{\partial w}{\partial y} \right)$$

For plane stress in xy plane

$$x_x = \frac{1}{E} (X_x - \sigma Y_y)$$

$$y_y = \frac{1}{E} (Y_y - \sigma X_x)$$

$$x_y = \frac{1}{A} X_y$$

$$X_x = \frac{E}{1 - \sigma^2} (x_x + \sigma y_y)$$

$$Y_y = \frac{E}{1 - \sigma^2} (y_y + \sigma x_x)$$

$$X_y = A x_y$$

7.9. MATHEMATICAL DERIVATIONS

7.91. *Longitudinal Vibrations in Two-Dimensional Plates*

As explained in the text, solutions for the infinite strip of Fig. 7.7 are first derived. Let

$$\left. \begin{aligned} u &= U \cos kx \cos \ell_y \\ v &= V \sin kx \sin \ell_y \end{aligned} \right\} \quad (7.12)$$

where U and V are constant. From these expressions $\epsilon = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}$ can

be obtained and substituted into the equilibrium equations 7.2. Two expressions as follows result after dividing through by the common term $\cos kx \cos \ell_y$.

$$A(k^2 + \ell^2) - \frac{Bk}{U} (-kU + \ell V) = \rho\omega^2 \quad (7.25)$$

$$A(k^2 + \ell^2) + \frac{B\ell}{V} (-kU + \ell V) = \rho\omega^2$$

Subtracting the second from the first of these equations, it is seen that

$$\left(\frac{Bk}{U} + \frac{B\ell}{V} \right) (-kU + \ell V) = 0 \quad (7.26)$$

Either or both of these factors equal to zero will satisfy 7.26, so that two values of $\frac{V}{U}$ are obtained. By substituting back into equations (7.25), conditions on ω^2 are found. The two solutions will be

$$\frac{V_1}{U_1} = -\frac{\ell_1}{k} \text{ with } (A + B)(k^2 + \ell_1^2) = \rho\omega^2 \quad (7.27)$$

$$\frac{V_2}{U_2} = \frac{k}{\ell_2} \text{ with } A(k^2 + \ell_2^2) = \rho\omega^2$$

By superimposing the two solutions the u and v displacements now become

$$\begin{aligned} u &= [U_1 \cos \ell_1 y + U_2 \cos \ell_2 y] \cos kx \\ v &= \left[-\frac{\ell_1}{k} U_1 \sin \ell_1 y + \frac{k}{\ell_2} U_2 \sin \ell_2 y \right] \sin kx \end{aligned} \quad (7.28)$$

Using the relationships of Section 7.8, one may now calculate all stresses. The argument k has purposely been kept the same for both of the superimposed solutions in order that boundary conditions at $y = \pm \frac{b}{2}$ might be satisfied regardless of σ .

For Y_y to equal zero at the edges of the strip

$$\frac{\partial v}{\partial y} + \sigma \frac{\partial u}{\partial x} \Big|_{y=\pm \frac{b}{2}} = 0 \quad (7.29)$$

This gives rise to the equation:

$$U_1(\ell_1^2 + \sigma k^2) \cos \ell_1 \frac{b}{2} - U_2[k^2(1 - \sigma)] \cos \ell_2 \frac{b}{2} = 0 \quad (7.30)$$

Similarly, if $X_y = 0$ at $y = \pm \frac{b}{2}$

$$\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \Big|_{y=\pm \frac{b}{2}} = 0 \quad (7.31)$$

Another relation is obtained from (7.31):

$$-2\ell_1 \ell_2 U_1 \sin \ell_1 \frac{b}{2} + U_2(k^2 - \ell_2^2) \sin \ell_2 \frac{b}{2} = 0 \quad (7.32)$$

The two equations (7.30) and (7.32) will be satisfied if the determinant of the coefficients of U_1 and U_2 vanishes. The following transcendental equation will then be obtained; values of ℓ_1^2 and ℓ_2^2 being those required by Eq. 7.27.

$$\frac{\cot \ell_1 \frac{b}{2}}{\cot \ell_2 \frac{b}{2}} = \frac{-2\ell_1 \ell_2 k^2(1 - \sigma)}{(\ell_2^2 - k^2)(\ell_1^2 + \sigma k^2)} \quad (7.13)$$

By using either of equations (7.30) and (7.32), one may derive the relation between U_2 and U_1 provided a solution to (7.13) is found.

$$U_2 = U_1 \frac{-2\ell_1 \ell_2 \sin \ell_1 \frac{b}{2}}{(\ell_2^2 - k^2) \sin \ell_2 \frac{b}{2}} \quad (7.33)$$

To solve equation (7.13) assume a value for k^2 and plot graphically the right and left hand expressions as functions of $\theta^2 = \frac{\rho\omega^2}{A}$. Roots are indicated

by the crossover points. Values of θ^2 corresponding to different values of k^2 may also be found in this way and a curve plotted for θ^2 versus k^2 , (or for $\theta \cdot b$ versus $k \cdot b$).

7.92. Thickness Shear Vibrations

To obtain a formula for the approximate resonant frequencies of thickness shear for a plate having large ratios of $\frac{W}{T}$ and $\frac{L}{T}$, one may consider the following displacements:

$$\begin{aligned} u &= U \sin kx \cos \ell y \cos rz \\ v &= 0 \\ w &= 0 \end{aligned} \tag{7.34}$$

If there are no cross couplings between shear stresses or between shear and extensional stresses one may write:⁸

$$\begin{aligned} X_x &= c_{11} \frac{\partial u}{\partial x} + c_{12} \frac{\partial v}{\partial y} + c_{13} \frac{\partial w}{\partial z} = c_{11} kU \cos kx \cos \ell y \cos rz \\ X_y &= c_{66} \left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \right) = -c_{66} \ell U \sin kx \sin \ell y \cos rz \\ X_z &= c_{55} \left(\frac{\partial u}{\partial z} + \frac{\partial w}{\partial x} \right) = -c_{55} rU \sin kx \cos \ell y \sin rz \end{aligned} \tag{7.35}$$

Substituting into the first of the equilibrium equations (7.1) and dividing through by common factors

$$c_{11} k^2 + c_{66} \ell^2 + c_{55} r^2 = \rho \omega^2 \tag{7.36}$$

The other two of equations (7.1) may be neglected if k and r are quite small so that it will only be necessary to consider equation (7.36) which can be solved for ω^2 . It will be noticed that the X_y shear stress will predominate under these restrictions on k and r . Letting $k = \frac{n\pi}{L}$, $\ell = \frac{m\pi}{T}$, and $r = \frac{p\pi}{W}$ (n, m and p are integers) in order to satisfy the boundary conditions for this stress and also for X_z , one obtains the following formula. (This choice of k, ℓ , and r is also required if the shear stress is to vary in essentially the same manner as is experimentally observed.)

$$\omega = 2\pi f = \pi \sqrt{\frac{1}{\rho}} \sqrt{\frac{c_{11}n^2}{L^2} + \frac{c_{66}m^2}{T^2} + \frac{c_{55}p^2}{W^2}} \tag{7.17}$$

in which L and W must be much larger than T .

⁸ Refer to equation (7.18) for values of c constants for isotropic case.

The boundary condition for the extensional stresses will not be met; however, they will be quite small in comparison to X_v if k is small, and may be neglected.

7.93. Thickness Flexures

Consider a three dimensional plate having a thickness b lying along the y direction. The following displacements are found to be of a form that can be made to satisfy the equilibrium equations 7.2.

$$\left. \begin{aligned} u &= U \sin kx \sin \ell y \cos mz \\ v &= V \cos kx \cos \ell y \cos mz \\ w &= W \cos kx \sin \ell y \sin mz \end{aligned} \right\} \quad (7.37)$$

Performing the operations indicated and substituting into the equilibrium equations give the following result:

$$\left. \begin{aligned} A(k^2 + \ell^2 + m^2) + \frac{Bk}{U} (kU - \ell V + mW) &= \rho\omega^2 \\ A(k^2 + \ell^2 + m^2) - \frac{B\ell}{V} (kU - \ell V + mW) &= \rho\omega^2 \\ A(k^2 + \ell^2 + m^2) + \frac{Bm}{W} (kU - \ell V + mW) &= \rho\omega^2 \end{aligned} \right\} \quad (7.38)$$

Subtract the second and third equations of (7.38) from the first:

$$\text{then} \quad \frac{Bk}{U} + \frac{B\ell}{V} = 0 \quad \text{and} \quad \frac{Bk}{U} - \frac{Bm}{W} = 0 \quad (7.39)$$

$$\text{or} \quad \frac{V}{U} = -\frac{\ell}{k} \quad \text{and} \quad \frac{W}{U} = \frac{m}{k}$$

Putting these values back into 7.38, it is seen that the following relationship must be satisfied.

$$(A + B) (k^2 + \ell^2 + m^2) = \rho\omega^2 \quad (7.40)$$

Letting $\frac{V}{U} = -\frac{\ell}{k}$ as in (7.39), another value for $\frac{W}{U}$ may be obtained.

The first and second equations of (7.38) will be satisfied for any ratio of $\frac{W}{U}$, so the 3rd equation is used.

$$A(k^2 + \ell^2 + m^2) + B \left(k^2 + \ell^2 + mk \frac{W}{U} \right) = \rho\omega^2 \quad (7.41)$$

Solving for $\frac{W}{U}$, using (7.41) and first of equations (7.38)

$$\frac{W}{U} = \frac{m}{k} \quad \text{and} \quad \frac{W}{U} = \frac{-k^2 - \ell^2}{mk} \tag{7.42}$$

The first ratio is the same as (7.39). For the second solution to the equilibrium equations, then the following relationships exist.

$$\frac{V}{U} = \frac{-\ell}{k} \quad \text{and} \quad \frac{W}{U} = \frac{-k - \ell^2}{mk} \tag{7.43}$$

When the above are substituted back into (7.38), it is found that⁹

$$A(k^2 + \ell^2 + m^2) = \rho\omega^2 \tag{7.44}$$

In a similar way, using $\frac{W}{U} = \frac{m}{k}$ the following are obtained:

$$\frac{W}{U} = \frac{m}{k}; \quad \frac{V}{U} = \frac{k^2 + m^2}{k\ell} \tag{7.45}$$

with $A(k^2 + \ell^2 + m^2) = \rho\omega^2$ as before. This is the second solution for $\epsilon = 0$.

The three different solutions may now be combined or superimposed to give

$$\begin{aligned} u &= [U_1 \sin \ell_1 y + U_2 \sin \ell_2 y + U_3 \sin \ell_3 y] \sin kx \cos mz \\ v &= \left[-U_1 \frac{\ell_1}{k} \cos \ell_1 y - U_2 \frac{\ell_2}{k} \cos \ell_2 y \right. \\ &\quad \left. + \frac{U_3(k^2 + m^2)}{\ell_3 k} \cos \ell_3 y \right] \cos kx \cos mz \tag{7.46} \\ w &= \left[U_1 \frac{m}{k} \sin \ell_1 y - U_2 \frac{(k^2 + \ell_2^2)}{mk} \sin \ell_2 y \right. \\ &\quad \left. + U_3 \frac{m}{k} \sin \ell_3 y \right] \cos kx \sin mz \end{aligned}$$

In the above equations $\ell_2^2 = \ell_3^2$ because of the double requirement of 7.44.¹⁰

It is now possible to calculate the stresses existing at any point. It is desired to choose U_1 , U_2 , and U_3 in such a manner that the boundary conditions at the two major surfaces of the plate are satisfied. By using the relations given in Section 7.8, the extensional stress Y_y , and the two shear stresses X_y and Y_z are calculated with the use of 7.46. They are then

⁹ It should also be noticed that $\epsilon = 0$ for this solution.

¹⁰ $k_1 = k_2 = k_3 = k$
 $m_1 = m_2 = m_3 = m$

set to zero at the faces of the plate; i.e. at $y = \pm \frac{b}{2}$. Three equations, after simplifying, result.

For $Y_y = 0$ at $y = \pm \frac{b}{2}$

$$U_1 \left[\sigma B(k^2 + \ell_1^2 + m^2) \sin \ell_1 \frac{b}{2} + A\ell_1^2 \sin \ell_1 \frac{b}{2} \right] \\ + U_2 \left[A\ell_2^2 \sin \ell_2 \frac{b}{2} \right] - U_3 \left[A(k^2 + m^2) \sin \ell_3 \frac{b}{2} \right] = 0 \quad (7.47)$$

For $X_y = 0$ at $y = \pm \frac{b}{2}$

$$U_1 \left[2\ell_1 \cos \ell_1 \frac{b}{2} \right] + U_2 \left[2\ell_2 \cos \ell_2 \frac{b}{2} \right] \\ + U_3 \left[\ell_3 - \frac{(k^2 + m^2)}{\ell_3} \right] \left[\cos \ell_3 \frac{b}{2} \right] = 0 \quad (7.48)$$

For $Y_z = 0$ at $y = \pm \frac{b}{2}$

$$U_1 \left[2\ell_1 m \cos \ell_1 \frac{b}{2} \right] + U_2 \left[\left(\ell_2 m - \frac{\ell_2}{m} (\ell_2^2 + k^2) \right) \cos \ell_2 \frac{b}{2} \right] \\ + U_3 \left[\left(\ell_3 m - \frac{m}{\ell_3} (k^2 + m^2) \right) \cos \ell_3 \frac{b}{2} \right] = 0 \quad (7.49)$$

In order for these three equations to be satisfied simultaneously a necessary condition is that the third order determinant formed by the coefficients of the U 's vanish. That is,

$$\begin{vmatrix} \left[(\sigma B(k^2 + \ell_1^2 + m^2) + A\ell_1^2) \sin \ell_1 \frac{b}{2} \right] & \left[A\ell_2^2 \sin \ell_2 \frac{b}{2} \right] & - \left[A(k^2 + m^2) \sin \ell_3 \frac{b}{2} \right] \\ \left[2\ell_1 \cos \ell_1 \frac{b}{2} \right] & \left[2\ell_2 \cos \ell_2 \frac{b}{2} \right] & \left[\left(\ell_3 - \frac{(k^2 + m^2)}{\ell_3} \right) \cos \ell_3 \frac{b}{2} \right] \\ \left[2\ell_1 m \cos \ell_1 \frac{b}{2} \right] & \left[\left(\ell_2 m - \frac{\ell_2}{m} (\ell_2^2 + k^2) \right) \cos \ell_2 \frac{b}{2} \right] & \left[\left(\ell_3 m - \frac{m}{\ell_3} (k^2 + m^2) \right) \cos \ell_3 \frac{b}{2} \right] \end{vmatrix} = 0 \quad (7.50)$$

By dividing row 1 by $\cos \ell_1 \frac{b}{2}$, row 2 by $\cos \ell_2 \frac{b}{2}$; row 3 by $\cos \ell_3 \frac{b}{2}$ and by subtracting the elements of row 3 from those of row 2, considerable simplification results.

The full expansion gives the following equation, after further simplifying operations are performed.

$$\left[(\sigma B(k^2 + \ell_1^2 + m^2) + A\ell_1^2) \tan \ell_1 \frac{b}{2} \right] \cdot [\ell_3^2 - k^2 - m^2] + 2\ell_1 \ell_3 \left[A(k^2 + m^2) \tan \ell_3 \frac{b}{2} \right] = 0 \tag{7.51}$$

It should be noticed that ℓ_2 has dropped out entirely. Actually $\ell_2 = \ell_3$ as previously explained. Also the expression

$$\left[\ell_2 m + \frac{\ell_2}{m} (\ell_2^2 + k^2) \right]$$

must not be zero, for in simplifying equation (7.51) it was used as a divisor.

Equation (7.51) may be rewritten to give

$$\frac{\tan \ell_1 \frac{b}{2}}{\tan \ell_3 \frac{b}{2}} = \frac{-2\ell_1 \ell_3 A(k^2 + m^2)}{[\sigma B(k^2 + \ell_1^2 + m^2) + A\ell_1^2][\ell_3^2 - k^2 - m^2]} \tag{7.52}$$

In the above

$$\left. \begin{aligned} (A + B)(k^2 + \ell_1^2 + m^2) &= \rho\omega^2 \\ A(k^2 + \ell_3^2 + m^2) &= \rho\omega^2 \end{aligned} \right\} \tag{7.53}$$

By letting $\theta^2 = \frac{\rho\omega^2}{A}$ and $k^2 + m^2 = \alpha^2$ equations (7.52) and (7.53) above become

$$\frac{\tan \ell_1 \frac{b}{2}}{\tan \ell_3 \frac{b}{2}} = \frac{-2\ell_1 \ell_3 A\alpha^2}{[\sigma B(\ell_1^2 + \alpha^2) + A\ell_1^2] \cdot [\ell_3^2 - \alpha^2]} \tag{7.22}$$

with

$$\begin{aligned} \ell_1^2 &= \theta^2 \cdot \frac{A}{A + B} - \alpha^2 \\ \ell_2^2 &= \ell_3^2 = \theta^2 - \alpha^2 \end{aligned}$$

Equation (7.22) represents the general solution for normal thickness vibrations in an isotropic plate of finite thickness extending to infinity in both major directions. The analogy for plates of finite dimensions is considered in the text.

CHAPTER VIII

Principles of Mounting Quartz Plates

By R. A. SYKES

INTRODUCTION

IT IS the object of this chapter to show some of the fundamental considerations involved that govern the design of mountings or holders of quartz crystals. This discussion is restricted to the three common types, namely, rod or clamp type, wire type and airgap type. The development of these three types of mountings for applications in telephone transmission and radio systems has led to many and varied forms. Commercial designs of units for telephone uses employing these principles are described in detail in a later chapter.

In chapter VI regarding the vibrations of crystals we have assumed in all cases that the crystal is free to vibrate. In order that this condition shall be fulfilled it is necessary that any mounting which supports the crystal shall not restrict its vibration or at most the effect shall be made as negligible as possible.

8.1 CLAMP TYPE SUPPORTS

Of the known types of vibration it is noticed in all cases that there have been nodal points. These points by definition are points of zero motion and in all cases that we have studied appear to be single isolated points or lines of very small size in comparison with the total crystal area. The obvious type of mounting is then one which simply clamps the crystal with a very small area at these points or nodes. The early type of mountings for low-frequency crystals were all based on this principle and the area of the clamp was determined experimentally by reducing it until, with sufficient pressure to hold the crystal, a good Q was obtained. The first mountings consisted simply of two pressure points located as nearly as possible to the nodal point. It was apparent at first that this type of mounting allowed the crystal to rotate about the mounting axis and very shortly the plating or electrode open-circuited. With the development of the “-18 degree X-cut” crystal it was found that the nodal region of a longitudinally vibrating crystal was a nodal line and permitted the use of a knife-edged type of mounting instead of the single point. This type of pressure mounting was used with this crystal for quite a number of years in the crystal filters for carrier systems and is shown in Fig. 8.1. This consists mainly of four pressure edges whose

dimensions along the length of the crystal are small and width sufficiently large to insure a rigid clamp. Pressure was applied by a phosphor bronze spring in the center of the two top pressure points. This gave a satisfactory mounting and also allowed the use of a divided plating necessary for the balanced type crystal filters. This type of mounting was used in crystals of relatively low frequency, for example, 60 to 150 kc. of the “-18 degree X-cut” type.

With the use of higher-frequency crystals of different types of vibration than that described above, it has been found that this method of mounting has not been very satisfactory. In order to reduce the size of the mounting in proportion to the decreased crystal area it would be a delicate mechanical job and quite costly. This type of mounting could not be used for crystals which did not employ this type of vibration, for example the face shear type

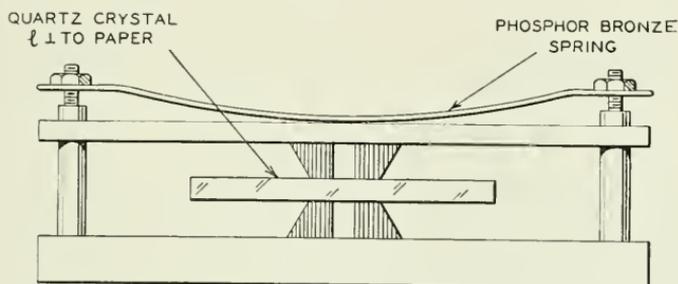


Fig. 8.1—Pressure mounting for extensional crystals.

such as the *CT* and *DT*, since there is only one spot near the center which would permit clamping at all.

To permit a crystal to vibrate freely the object used to support the crystal and maintain contact to the plated surfaces must have a very low mechanical impedance. At the same time it should possess sufficient rigidity that the complete assembly may be shocked without changing characteristics of the crystal as an oscillator. For example, if a rod or bar is held against a crystal at any point we would expect that the crystal in an oscillating condition would tend to generate motion in the bar and as this bar is placed closer to the nodal point we would expect the motion to be less. It can be seen that there are two objects to be accomplished in mounting a crystal: First, that the support must be placed as close as possible to a nodal point; and second, that the support shall have a very low mechanical impedance. This mechanical impedance needs to be low only at or near the operating frequency of the crystal. One type of support which would meet this requirement is that of a rod in flexure a discussion of which is given in Chapter VI. In this case, however, we may clamp one end of the bar and allow the other end to

be free to vibrate. This free end would then be in contact with the surface of the crystal. If the bar were clamped and were of a length such that its frequency of resonance equalled that of the crystal or approximately so, it would require very little energy from the crystal to drive it, and any energy received from the crystal would be reflected from the clamped end of the bar and thereby kept within the vibrating system. This type of support is shown in Fig. 8.2, where l = length of the rod and d its diameter. The slightly rounded end is to allow the rod to seat firmly on the crystal surface. An enlarged view of Fig. 8.2 is shown in Fig. 8.3 and shows how the rod would vibrate. Figure 8.3A shows the type of motion for the first mode of a clamp-

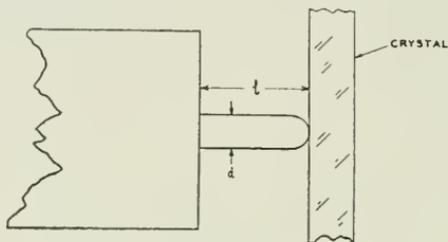


Fig. 8.2—Cantilever type mounting.

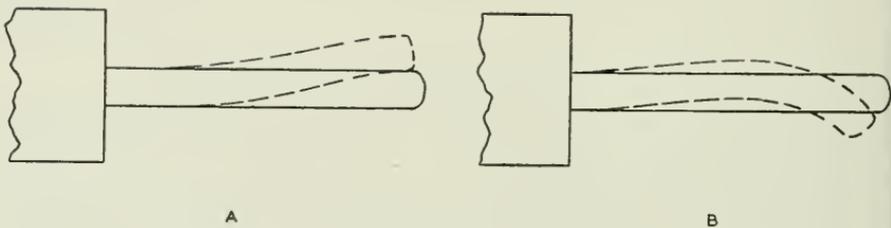


Fig. 8.3—Type of motion in cantilever support mountings.

free bar. Figure 8.3B shows the type of motion of the same bar vibrating in its second mode. This would indicate that for a given length of bar we could use it at several different frequencies by simply using higher orders of vibration. By using a clamp type mounting where the clamping rods are designed as shown in Fig. 8.2, we may now have a mounting which at the crystal frequency will allow the crystal to vibrate unrestricted but at the same time provide a very secure clamp thus preventing the crystal from moving about in its holder. To prevent rotation of the crystal about the axis of the clamped points, more than two can be used provided they are of the proper design. The frequency of a clamp-free rod in flexure is given by equation (8.1) where m now has values different than in the case of free-free flexure.

$$f = \frac{m^2 dv}{8\pi \ell^2} \quad (8.1)$$

where v = velocity in cm./sec.

d = diameter in cm.

ℓ = length in cm.

$m = 1.875$ for first mode

= $(n-1/2)\pi$ for 2nd, 3rd, etc.

From this we can compute the length necessary for a given rod at a given frequency and use this for the design of the clamping rods. This length is given in equation (8.2) for the case of a 100-kc crystal using phosphor bronze rods 1 millimeter in diameter

$$\begin{aligned} \ell &= 1.875 \sqrt{\frac{.1 \times 36 \times 10^5}{8\pi \times 10^5}} \\ &= .225 \text{ cm} \end{aligned} \quad (8.2)$$

This corresponds to the case of Fig. 8.3A. For the case of Fig. 8.3B, the length is given by

$$\ell = .567 \text{ cm}$$

Using this same diameter rod, if we should go to a considerably higher frequency, for example 5 megacycles, the value of ℓ would be extremely small even for the case of Fig. 8.3A and would be somewhat smaller than the diameter of the rod. As mentioned before in Chapter VI, the simple formulae that apply in the case of flexure are only for the case of a long thin rod. When the length becomes equal to or less than this diameter, it is very probable that the support member should be designed as though it were vibrating in shear. These follow well-known rules and are only mentioned here in case designs for high-frequency crystals are contemplated using this method.

The design of rod-supported crystals following this procedure has not been carried on to a large extent in these laboratories because, at present, the wire-supported crystal appears to have many advantages. A great deal more of the work in regard to resonating supports has been done for the case of the soldered lead type¹.

8.2 WIRE TYPE SUPPORTS

The theory of resonating supports involving soldered leads on crystals is very similar to that just discussed for the case of rods. There are two additional elements that we have here that are not present in the case of the rod, these elements being the actual solder connections that fasten the wire

¹ The presence of standing waves on the lead wires of CT crystals was found experimentally by Mr. I. E. Fair.

to the crystal and the coupling between the crystal and wire vibrating systems. Considerable work has been done in regard to the amount of solder necessary and the most desirable shape for the solder cone. The complete assembly of a wire support for a crystal is shown in Fig. 8.4. The shape of the solder cone shown in Fig. 8.4 has proved to be the most desirable and has been termed as "bell-shaped." This type of cone formation allows the wire to be twisted in handling and still not break away the top of the cone and form an appreciable crater. For the purposes of analysis we may then assume that the cone becomes part of the crystal and moves with it so that when computing the length of a wire vibrating in flexure, this length should be determined from the top of the cone. The amount of solder used in the cone since it is part of the crystal must be kept at a minimum in order that the constants of the crystal equivalent circuit will not be modified too much by it. One established fact of the effect of the solder in the cone on the

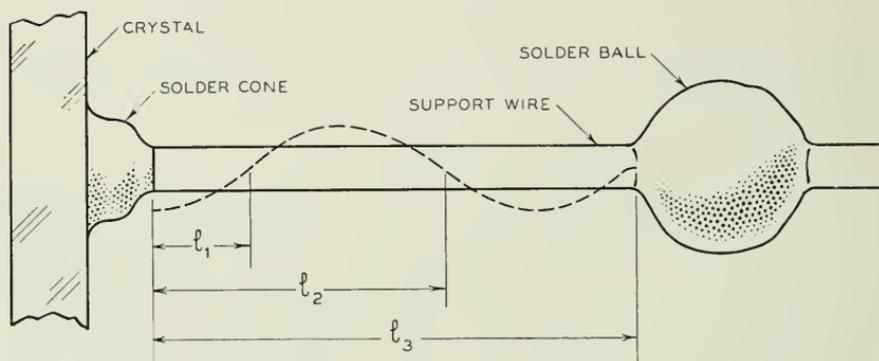


Fig. 8.4—Soldered lead type mounting.

equivalent circuit is to raise the resistance in the equivalent circuit for the crystal and this resistance increases considerably with an increase in temperature. The amount of solder permissible in the cone would then be determined by the maximum temperature at which the crystal is to be operated and the minimum Q allowable. The type of motion that the crystal would generate in the support wire when oscillating is that shown in Fig. 8.4 by the dotted line. The solder ball shown to the right of the figure acts as the clamp for the wire. This solder ball may be placed at any point along the wire corresponding to a node. The diameter of this ball need only be sufficient to act as a clamp. In general, this will be in proportion to the wire diameter. For example, at 200 kc it was necessary to use a solder ball 60 mils in diameter on a 6.0-mil diameter phosphor bronze wire. The spacing between the solder ball and the head of the cone may be readily computed from equation (8.1). In practice, it has been found that in most all cases this distance is slightly greater than that given by the formula due to the

fact that the free end is restricted to zero slope and for a given crystal and support wire it should be determined experimentally using the values obtained from equation (8.1) as a guide in the design. The diameter of the solder ball that acts as a clamp may also be determined experimentally by

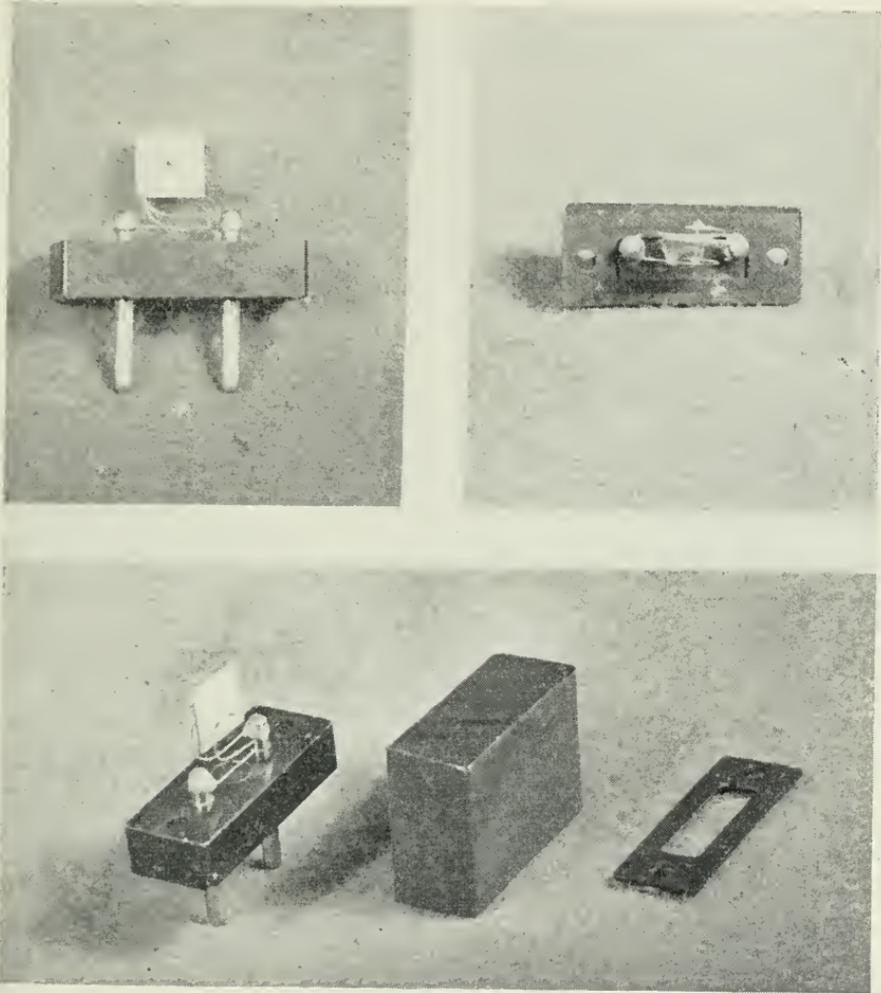


Fig. 8.5—FT-241 crystal mounting.

increasing its size until the standing waves on the wire to the right of the ball are sufficiently reduced. A practical application of this type of support is shown in Fig. 8.5. The top view shows the small wires soldered to the crystal as well as the solder balls that are spaced at points corresponding to the second node on the lead wire from the crystal. These solder balls act

as mechanical termination for the lead wires and also as connection to larger size spring wires forming the rest of the shock-proof mounting.

Another type of wire support that has found considerable practical use and is superior to the straight lead and solder cone type of connection is that of the headed wire. This is shown in Fig. 8.6. A headed wire is similar to that of common pin and may be connected to the crystal by sweating the head to the crystal as shown. This has certain advantages over the solder cone in that the head of the wire being a machined part is always constant and the distance d , as shown in Fig. 8.6, is the same for all mountings. The amount of solder necessary to sweat the head to the crystal is considerably less than in the case of the cone and hence this type of mounting will have less dissipation at the higher temperatures. One other factor not mentioned above is that the coupling between the vibrating system of the wire and the vibrating system of the crystal is considerably reduced by the use of

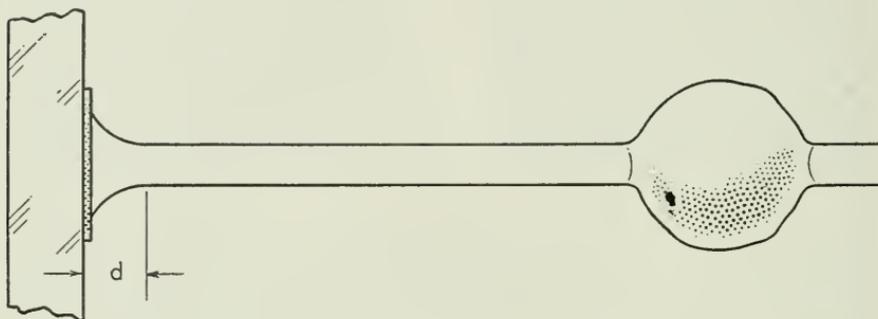


Fig. 8.6—Headed wire type mounting.

the headed wire. This is an important factor in reducing what may be termed a double system of standing waves on the wire. One standing wave system would result from reflections from the clamped end of the wire, while the other would result from reflections between the clamped wires coupled through the crystal. This may be reduced by a reduction of coupling between the crystal and wire vibrating systems.

Measurements have been made on the effect of clamping the wire-supported crystal at various points, on the activity and frequency of several different crystals used in oscillators and filters. Figure 8.7 shows the effect of clamping a 500-kc CT type crystal such as now used in the FT-241 holder. Figure 8.8 shows the same condition for a 370-kc CT crystal. It will be noted that in these two cases with the decrease in frequency of the crystal that the coupling between the wire and crystal has decreased, as shown by a smaller change in frequency and also, that for the lower frequency crystal the change in activity is modified only when the clamp is very close to a loop of motion on the wire. The mountings of these crystals were of the type

shown in Fig. 8.4 where the amount of solder in the cone equals that of a solder pellet 20 mils in diameter and 12 mils high.

Figure 8.9 shows the change in frequency as a result of clamping one wire of a four-wire mounting of a GT-cut crystal designed for use as a filter ele-

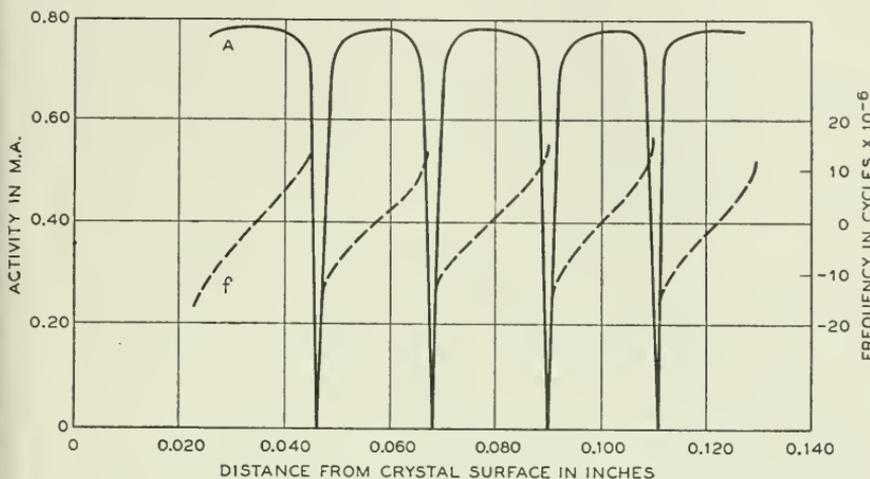


Fig. 8.7—Effect on frequency and activity of clamping one lead of 500 kc. CT-cut crystal.

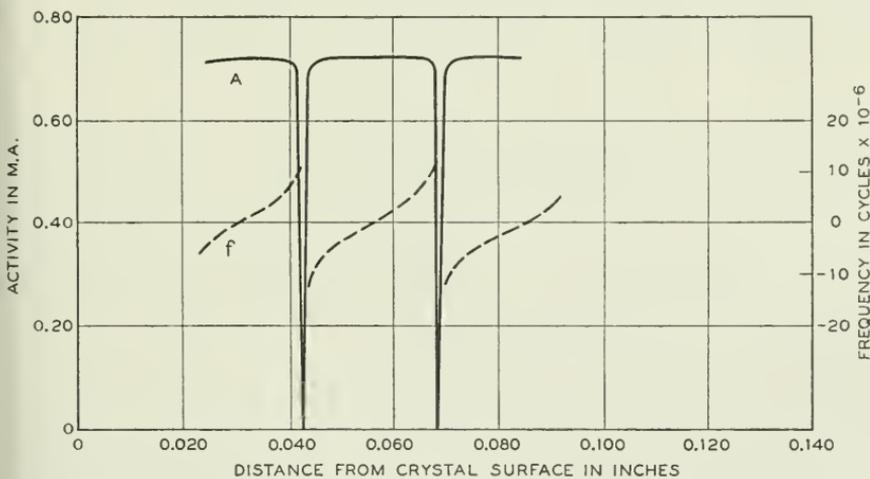


Fig. 8.8—Effect on frequency and activity of clamping one lead of 370 kc. CT-cut crystal.

ment at 164 kilocycles. This change is shown for the lower resonance at 143 kilocycles since this mode would be more affected by clamping. The large deviations in frequency correspond to clamping at the loops of the wire as shown in Figs. 8.8 and 8.9 but the small sudden changes in frequency are a result of a second system of standing waves as previously described. This

second system of standing waves results from too much coupling between the crystal and the two oppositely disposed lead wires. It may be reduced by first placing the wires closer to the nodal point and second, using a smaller amount of solder in the cone to attach the lead wire to the crystal. Measurements on this same type of crystal when the above conditions were fulfilled showed practically no effects of secondary standing waves. It is important to keep the energy transmitted to the lead wires low since a soldered connection near a loop of motion resulting from secondary standing waves on the wire will act as a clamp and will materially decrease the resulting Q of the crystal. This is probably the best reason for the use of the headed wire type of lead wherever practical.

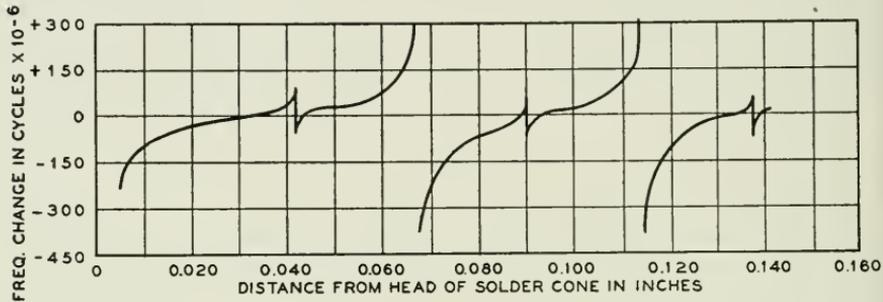


Fig. 8.9—Effect on the frequency of lower resonances of clamping one lead of 164 kc. GT-cut crystal.

8.3 AIR-GAP TYPE SUPPORTS

A third form of mounting for quartz crystals is that of the airgap type shown in Fig. 8.10 where the crystal plate is held between two flat electrodes. Two forms of the airgap type of mounting are shown. In Fig. 8.10A the crystal is free to vibrate between two flat electrodes held together to produce a definite airgap of thickness t . In Fig. 8.10B small lands are left on the corners of the electrodes to produce a uniform airgap on each side of the crystal as well as to clamp the crystal plate.

This type of mounting has found its greatest use for oscillator crystals of the AT and BT type. The factor that determines the choice of mount is the ratio of length to thickness of the crystal. For example, when the length is less than 20 times the thickness, clamping the corners of AT and BT type crystals will decrease the activity in proportion to the clamping pressure. This is apparent from a study of the type of motion for these crystals described in Chapter VI. This then indicates that AT and BT type crystals for broadcast frequencies should employ a mounting with the crystal unrestricted as shown in Fig. 8.10A while the higher radio frequency crystals may be clamped as shown in Fig. 8.10B. The clamping pressure will be

dependent upon the area of the crystal, its frequency and the amount of activity required. One advantage of the clamped type support lies in the fact that many of the unwanted modes of motion are restricted or damped to the extent that they will not cause serious dips in the activity character-

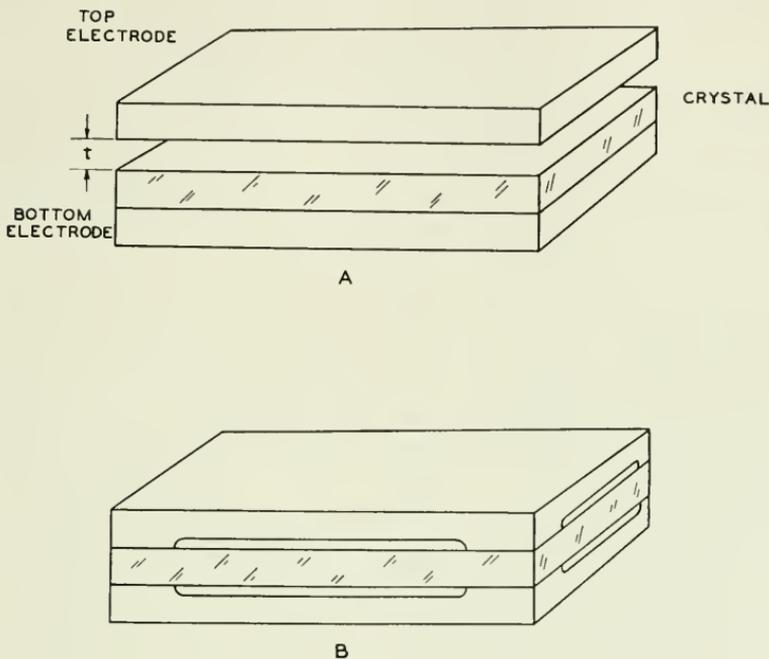


Fig. 8.10—Air gap type mounting.
 A—Crystal free.
 B—Crystal clamped at corners.

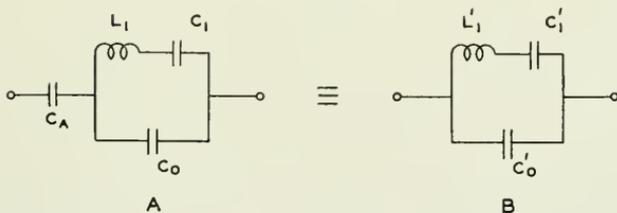


Fig. 8.11—Equivalent circuit of a quartz crystal in an air gap type mounting.

istic over a wide temperature range. This explains in part the necessity for accurate control of the length and width dimensions for crystals of low radio frequencies using the type of mounting shown in Fig. 8.10A.

The effect of the airgap on the constants of the crystal equivalent circuit may be determined from Fig. 8.11. In Fig. 8.11A is shown the usual crystal equivalent circuit in series with a capacity C_A which represents the capacity

of the airgap. This may be reduced to the circuit of Fig. 8.11B where the constants are given by

$$C'_0 = \frac{C_A}{C_A + C_0} C_0$$

$$C'_1 = \frac{C_A^2}{(C_A + C_0)(C_1 + C_A + C_0)} C_1$$

$$L'_1 = \left[\frac{C_A + C_0}{C_A} \right]^2 L_1$$

The circuit of Fig. 8.11B is the same form as that of the original crystal and therefore we may assume that the effect of the airgap is to produce a similar

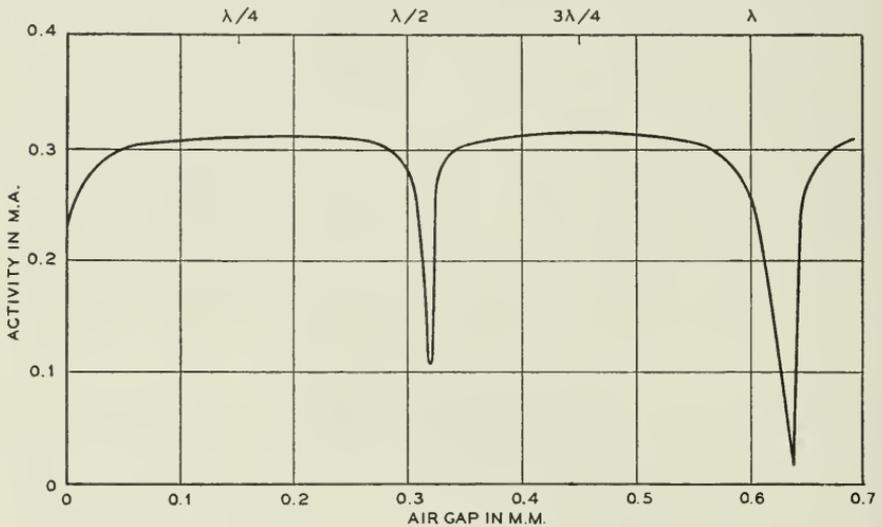


Fig. 8.12—Effect on frequency of the air gap thickness on a 550 kc. AT-cut crystal.

crystal of reduced capacity and reduced effective piezoelectric coupling. In the case of oscillatory crystals the effect of the airgap is to reduce the activity and decrease the range of frequency adjustment with parallel capacity. For filter applications the effect of the airgap is to produce narrower transmission bands and higher characteristic impedance. One other effect of the airgap results from the propagation of acoustic waves from the crystal.

It is known that most any type of crystal in a vibrating condition will produce acoustic waves in air and if an object capable of reflecting these waves is the proper distance away, these acoustic waves may be reflected back to the crystal surface. The reflections from distances corresponding to even quarter wave-lengths will cause considerable damping while the reflections from distances corresponding to odd quarter wave-lengths will

cause very little. The wave-length of a sound wave in air may be readily computed, and since we are interested in multiples of one-quarter wave-length, it is desirable to determine these for a given frequency. This can be computed readily from equation 8.3,

$$\frac{\lambda}{4} = \frac{v}{4f} \quad 8.3$$

where v is the velocity of sound in air at room temperature and pressure and equals 33,000 centimeters per second. For example, a quarter of a wave-length at 5 megacycles is given by

$$\frac{\lambda}{4} = \frac{33,000}{4 \times 5 \times 10^6} = .00165 \text{ cm}$$

which indicates that if l of Fig. 8.10 were made equal to this or odd multiples, there would be very little effect of the electrode on the crystal and if l corresponded to even multiples of a quarter wave-length, we would expect considerable damping. Some measurements of this effect have been made with a low frequency AT -cut quartz crystal and are shown in Fig. 8.12. The sound wave generated by an AT -cut probably results from flexure waves generated by the high-frequency shear wave. It will be noted that when the airgap is equal to even multiples of a quarter wave-length, the activity is considerably reduced. Further, it will be noticed that airgaps in the order of $1/8$ of the wave-length may be used and produce very little effect. Since a large airgap reduces the piezoelectric coupling it is desirable to keep this about $1/8$ of a wave-length as a maximum unless, in special cases, a reduction in piezoelectric coupling may be tolerated.

The Magnetically Focused Radial Beam Vacuum Tube

By A. M. SKELLETT

A new type of vacuum tube is described in which a flat radial beam of electrons in a cylindrical structure may be made to rotate about the axis. Features of the tube are its absence of an internal focusing structure and resultant simplicity of design, its small size, its low voltages, and its high beam currents. The focusing of the beams and their directional control are accomplished by the magnetic fields in small polyphase motor stators. A time division multiplex signaling system for 30 channels using these tubes is briefly described.

IT HAS long been recognized that the substitution of electron beams for mechanical moving parts would offer decided advantages in many applications in the field of communications. The high voltages required for the usual cathode-ray type of tube and the very low currents obtainable therefrom prevent their use in most such proposals; their complicated guns and their large sizes are also undesirable features. The kind of tube described herein has no focusing structure, is small in size, requires only low voltages, utilizes the cathode power efficiently, and produces beam currents of the same order of magnitude as the space currents of ordinary vacuum tubes.

Figure 1 shows the elementary tube structure. It consists, in the simplest case, of a cylindrical cathode of the sort in common use in vacuum tubes, surrounded by a cylindrical anode structure. When this structure is made positive with respect to the cathode and there is no magnetic field in the tube, the electrons flow to the anode structure in all directions around the axis. When a uniform magnetic field is applied with its direction at right angles to the axis, the electrons are focused into two diametrically opposite beams as shown. The beams are parallel to the lines of force of the magnetic field so that if the field is rotated the beams move around with it. Thus the magnetic field serves both to focus the electrons and to direct the resulting beams to different elements of the anode structure.

If ordinary commercial cathodes are used with anode structures an inch or two in diameter, 100 volts or less on the anode will draw the full space current for which the cathode was designed. The application of the magnetic field will then focus from 85 to 90 per cent of this electron current into the two beams, the remaining 10 or 15 per cent being lost at the cathode due to an increase in the space charge which the magnetic field produces. Some of the smaller tubes produce beam currents of more than 5 milliamperes with only 50 volts on the anode structure, and in some of the tubes with larger cathodes beam currents of 50 milliamperes or more are easily obtainable. The magnetic field strengths range from 50 to 300 gauss.

For some applications it is desirable to eliminate one of the two beams and this may be accomplished by substituting a uniform electrical field in the tube for the cylindrical one described above. The uniform field may be obtained by applying to the anode elements a series of potentials that vary according to the sine of the angle taken around the axis. The line joining the maximum potentials (+ and -) is maintained parallel to the magnetic field so that on one side of the cathode the potentials are all negative and the

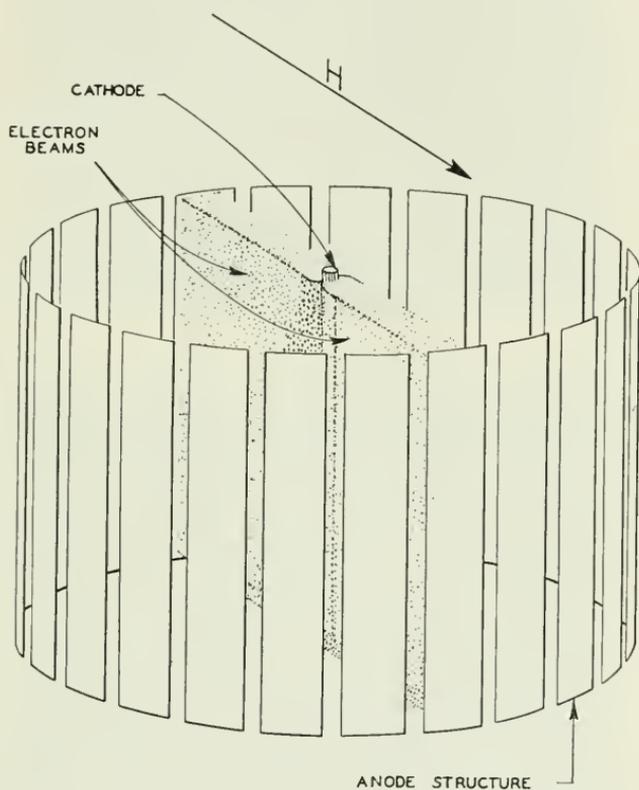


Fig. 1.—Elementary tube structure showing focused beams.

beam on that side is suppressed. The remaining beam will have somewhat less current than the corresponding one in the cylindrical field but the magnetic field-strength required for focus is reduced.

CYLINDRICAL ELECTRICAL FIELD

For the case of the cylindrical electric field the focus is obtained by applying a magnetic field that is strong enough to reduce the radius of curvature of the spiral electron trajectories to a small value. There is not obtained an electron optical image of the cathode in the usual sense that for

each point on the cathode there is a corresponding point on the image. The sharpness of the image may be increased by increasing the strength of the magnetic field and the field required for any degree of focus is not sharply critical.

Figure 2 shows a series of drawings of the various electron images that were obtained as the magnetic field-strength was increased in a tube having

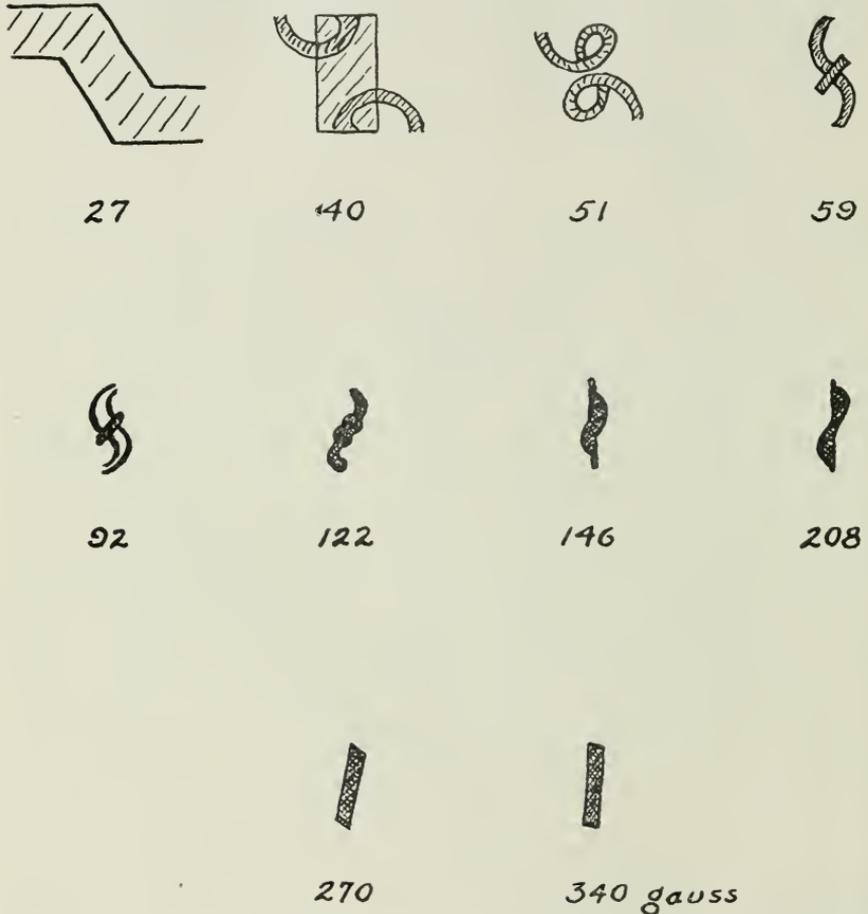


Fig. 2.—Drawings of the patterns obtained with a fluorescent coating on the inside of the anode when the magnetic field strength is increased from zero to the focus values.

a fluorescent coating on the inside cylinder. The cathode and anode diameters were 0.0625 and 2.5 inches, respectively, and the axial length was 2 inches. The anode was held at 150 volts. Only one-half inch of the cathode length, located centrally along the axis, was coated to emit electrons. The image at 340 gauss appeared to be one-half inch long. In attempting to interpret these patterns it should be remembered that on the two sides of the cathode at right angles to the plane of the beam the electrons follow



A



B

Fig. 3.—Electron trajectories made visible with a small amount of gas. A.—Magnetic field lined up with active spots on the cathode. B.—Magnetic field at 45° with respect to the active spots.

cycloid-like paths along the cathode, moving up on one side and down on the other.

The photographs of Fig. 3 showing the trajectories were obtained by

introducing argon at a pressure of about a micron into the tube. The electrons are emitted from only two spots of active material located at the opposite ends of a diameter on the cathode sleeve. In Fig. 3a the line joining the spots is lined up with the magnetic field and in 3b this line is at an angle of about 45° with respect to the field. This arrangement does not reproduce exactly the space charge conditions in the tube as actually used but does serve to give a picture of the electron paths in a qualitative sort of way.

As shown by the patterns of Fig. 2 above a minimum strength of magnetic field the shape of the focus does not change greatly. An approximate equation may be derived for the beam width in terms of the magnetic field above this minimum value that is useful for predicting the performance of new designs. The electrons that leave the cathode at right angles to the beam require the strongest magnetic field to keep them in focus. Now because of the cylindrical structure the electric field is concentrated near the cathode and we will assume that after leaving the vicinity of the cathode the velocity does not change appreciably. Setting v equal to the component of this velocity at right angles to the magnetic field we have that the radius r of the spiral path is given by the relation

$$r = \frac{mv}{eH}$$

where H is the magnetic field-strength and m and e are the mass and charge of an electron.

We also write

$$v = \sqrt{\frac{2eKV}{m}}$$

where K is the fraction of the anode voltage corresponding to v .

The width of the focus A is approximately equal to the cathode diameter D plus twice the maximum radius of curvature of the spiral paths

$$A \cong D + \frac{6.7\sqrt{KV}}{H}$$

where A and D are in centimeters and V is in practical volts. By substitution in this formula we have found that the empirical constant K is about 0.7 for the tubes that have been made to date. A minimum value for H is obtained, again approximately, by setting the last term in the equation equal to D .

UNIFORM ELECTRIC FIELD

As mentioned above the uniform field is obtained by imposing potentials around the anode periphery varying as the sine of the angle. The cathode is

at a point of zero potential. In this case a real electron optical image of the cathode is obtained.

Neglecting the distortion of the field in the vicinity of the cathode, the force equation for the electrons is

$$m \frac{d^2 x}{dt^2} = e \frac{V}{R}$$

where V is the maximum anode potential, R is the radius of the anode structure and x is measured in the direction of the fields. Since the acceleration is uniform the transit time t , neglecting space charge effects, may be obtained from the expression

$$\frac{1}{2} \left(\frac{d^2 x}{dt^2} \right) t^2 = R$$

Combining these equations we get

$$t = \frac{R}{\sqrt{\frac{Ve}{2m}}}$$

The condition for focus is that the electrons make one revolution around the lines of force in time t . The angular velocity of the electrons is given by the well-known expression

$$\omega = \frac{He}{m}$$

Setting $\omega t = 2\pi$ we get

$$H = \frac{2\pi}{R} \sqrt{\frac{m}{2e}} V$$

or in practical units

$$H = \frac{10.6\sqrt{V}}{R}$$

Since the effect of the magnetic field on the space charge has not been evaluated, we can only estimate the order of magnitude of the increase of transit time due to the space charge. On the assumption that this increase introduces a factor of $3/2^*$ the above expression with space charge is

$$H = \frac{7.1\sqrt{V}}{R}$$

This formula has been found to check well experimentally.

* The factor of $3/2$ is the ratio of the transit times in a plane parallel diode with and without space charge. See for example Millman and Seely, "Electronics," Chapt. 7, p. 231.

These last two formulae are for the first focus. Focii will also be obtained for values of H equal to nH where n is an integer and equal to the number of electronic revolutions. Actually as the field is increased beyond that necessary for the first focus the beam does not get very badly out of focus because the radius of curvature of the spiral path is small and for still higher fields the beam remains in approximate focus for all values of H .

In applications where the beam is rotated by means of a rotating magnetic field this electrostatic field is made to turn by separating the anode structure into four or six elements (or groups thereof) and applying either two- or three-phase alternating potentials to them.

MAGNETIC FIELD SUPPLY

The stator of a two-pole polyphase alternating-current motor furnishes an excellent magnetic field for use with these tubes. The tube is inserted in place of the armature and when the polyphase currents are applied the beams are formed and rotate at the cyclic frequency. For applications where the beams are not rotated continuously, a two-phase stator may be used in which the currents through the two windings are adjusted to be proportional to the sine and cosine of the desired direction angle of the beam. Permanent magnets of the horseshoe design have also been found to be suitable.

The power consumed by a stator depends on its size and the strength of the field it produces and on the cyclic frequency if it is used to rotate the beam. At low frequencies, e.g., 20 or 60 cycles, the power consumed is primarily that due to the copper loss. At higher frequencies the losses in the core material become important. For some of the smaller tubes operating at a low frequency, the power consumed by the stator is less than three watts. This stator has the regular motor windings which do not completely fill the slots.

Since a polyphase source of power is not always readily available, it is sometimes advantageous to split single-phase power in the stator itself to produce the rotating field. This may be done by inserting a condenser in series with each winding so that the current through one phase winding lags by 45° and that through the other leads by an equal angle. Polyphase potentials for producing a rotating electrostatic field in the tube may then be taken from the windings of the stator if desired.

TUBE DESIGN

The particular design of tube depends on its application. The simple design shown in Fig. 1 has been found adequate for some purposes but more elaborate designs which increase the versatility of the tube are also needed.

Figure 4 shows a tube with 30 anodes that incorporates various auxiliary elements. This tube is 2.25 inches in diameter. Figure 5 shows the internal

arrangement of the elements. Closely surrounding the cathode is a control grid that may be used for modulating the current density of the electron beams. Farther out is a cylindrical element with 30 windows that is maintained positive and which by virtue of its similarity in position to the third element of a tetrode is called a screen. Immediately behind each window there is a pair of paraxial wires which because of its similarity in function to the fourth element of a pentode is called a suppressor grid. In back of each suppressor grid there is an anode. In this particular tube there are pro-

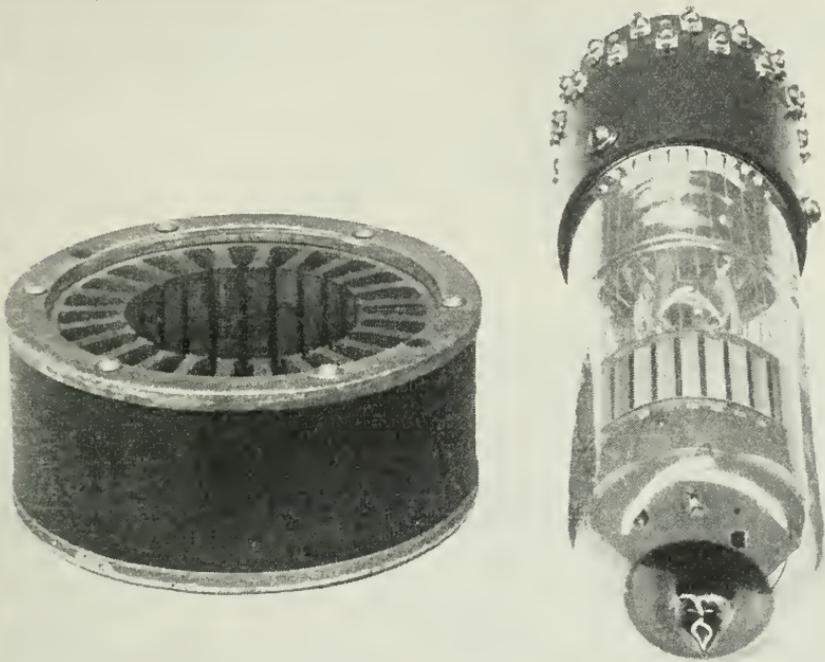


Fig. 4.—Radial beam tube with 30 anodes and unwound stator used with it.

jections like gear teeth on the back of the screen element to prevent electrons, destined for one anode, from reaching an adjacent one.

The control grid that is close to the cathode is biased negatively and controls the electron current in the same way that it would if the magnetic field were not present. The space current vs. grid potential curve is nearly identical for the two cases: with and without the magnetic field. The slight difference is due to the fact that the presence of the magnetic field increases the space charge near the cathode. Thus the tube may be used for amplification in the usual way when the electrons are focused. The presence of this grid has no appreciable effect on the focusing of the electrons.

Since the screen element is in one piece there will be present two beams out to it. One of these may be suppressed after it has passed through the screen by the suppressor grids or by the anodes in the manner described below.

These suppressor grids are generally operated at cathode potential or at a potential that is negative with respect to the cathode. They may be used for three purposes: to suppress secondaries from the anodes, to modulate the beam current to their particular anode, and to suppress one of the two beams. For the first of these functions they are biased at cathode potential. For the second they are biased negatively and have a modulation curve similar to that of the suppressor grid in a pentode. Curve A of Fig. 6 shows the variation of beam current to one anode when the potential of the suppressor grid in front of it is varied. This curve is for a grid similar to the two paraxial

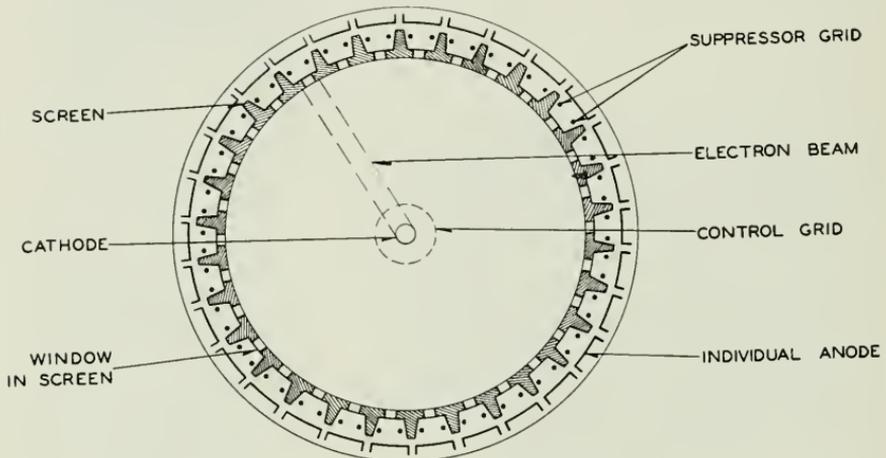


Fig. 5.—Arrangement of elements in the tube shown in Figure 4. Only the operating beam is shown.

wires in the tube shown in Fig. 5. For some applications a higher suppressor-anode transconductance or a lower cut-off is desirable and these may be obtained by welding lateral wires across this grid window to make the grid action more effective. Curve B of Fig. 6 was taken with the same size window across which laterals were welded. The table below gives the data for this suppressor grid with and without the lateral cross wires.

	Without Laterals	With Laterals
Transconductance (mho).....	100	250
Anode Resistance (ohms).....	30,000	64,000
Amplification Factor.....	3.5	16.0
Cut-Off Voltage.....	-80	-20

It is apparent from these data that amplification of the signals applied to the individual suppressors may be readily obtained.

If the screen element is split to give a uniform electrostatic field to suppress one beam, the beam current is only about half that of one beam of the cylindrical field case. This is because with the uniform electrostatic field the potential gradient at the cathode decreases with azimuthal angle away from the beam axis. If the unwanted beam is rejected by the suppressor grids, however, the beam current for the cylindrical case is obtained since the screen in this latter case supplies a cylindrical electrostatic field at the cathode and the unwanted beam is rejected between the screen and suppressor grids.

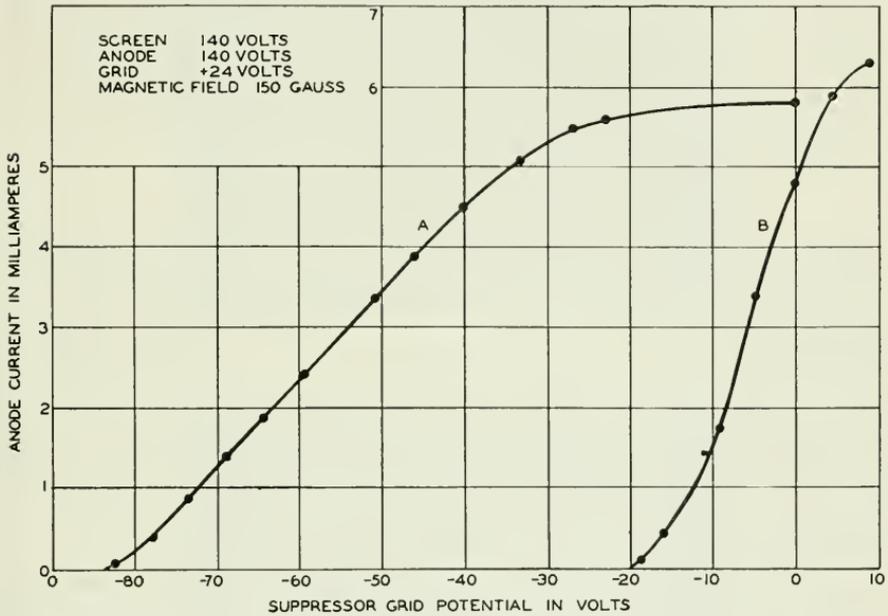


Fig. 6.—Suppressor grid characteristics. A.—Without lateral wires. B.—With lateral wires.

For this case the screen is maintained at the same positive potential required for the two-beam condition and the suppressors are so biased that they are beyond cut-off on one side of the tube and at or near cathode potential on the other side. If the beam is rotated the suppressors are connected to the polyphase supply in groups in the same way that the screen elements would be connected except that the d-c. bias above and below which the a-c. potentials swing is made negative at a value near cut-off for the suppressors.

When one beam is suppressed either by splitting the screen or by grouping the suppressors, the currents to the different anodes are not all exactly the same. For instance, maximum current will be received by an anode back

of the center of one of the screen elements or one of the suppressor groups and a minimum current will be received by an anode back of the junction of two such elements or groups. If two-phase supply is used (4 elements or groups) the ratio of maximum to minimum anode current will be 0.707 and for three-phase supply this ratio will be 0.866. There will be 4 or 6 maxima, respectively, around the tube. This variation may be effectively eliminated by varying the individual anode load impedances or in other ways.

The anode characteristics are similar to those of a pentode if suppressor grids are used and to that of a tetrode if these grids are not used.

There is still another method of effectively eliminating one beam. This consists in using an odd number of anodes so that when one beam is focused on an anode the opposite one falls on the screen in between two anode positions. With this type of tube the effective rotational frequency is twice the cyclic frequency of the rotating field, that is, all of the anodes are contacted twice (once for each beam) per revolution of the field.

APPLICATIONS

The many possible combinations of the tube elements just described permit a variety of applications. One of the simplest and most obvious is that of an electronic commutator which has the advantages over the corresponding mechanical device of speed and freedom from contact trouble. There is, however, a practical limitation to the speed of this electronic commutator that is set primarily by the alternating-current losses in the stator. This is estimated to be in the neighborhood of 10,000 cycles per second for ordinary stator and tube designs. The highest cyclic speed for a stator that has been used to date was 600 cycles per second which with utilization of both beams gave an effective cyclic frequency of 1200 cps.

One of the earliest systems of multiplex telegraphy was based on time division using mechanical rotating commutators. A small portion of the time of one cycle of the moving brush was allotted to each channel. The usefulness of this system is limited because of the faults of the mechanical commutators. The substitution of these electronic commutators eliminates these difficulties and puts the time division system on a more practical basis. It has an advantage over the frequency division multiplex system (carrier system) in that the elaborate filters of the latter are not required.

A 30-channel multiplex system for signaling using two of the 30 anode tubes described above has been successfully tested over short distances in the metropolitan area in New York City. The tube at the transmitter had all of the anodes tied together and the signal from them was sent over the line. The 30 input channels terminated on the suppressor grids of this tube. At the receiver, the input was fed to the negative grid surrounding the cathode and each of the anodes was connected in series with a small neon lamp for

an indicator. A signal on any one or signals on any group of the 30 input channels would actuate the corresponding lamp or lamps at the receiver. No amplification other than that provided by the receiver tube was needed.

A single beam was used in each tube, the other one being rendered ineffective in the transmitter by means of two-phase potentials applied to the

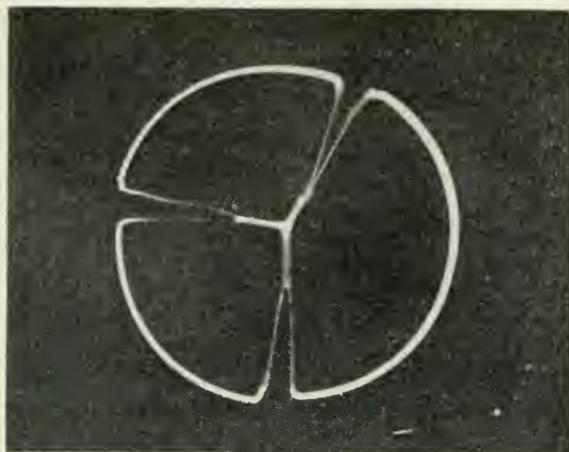


Fig. 7.—Circular trace oscillograph of transmitted signal when 3 out of 30 channels are in operation.

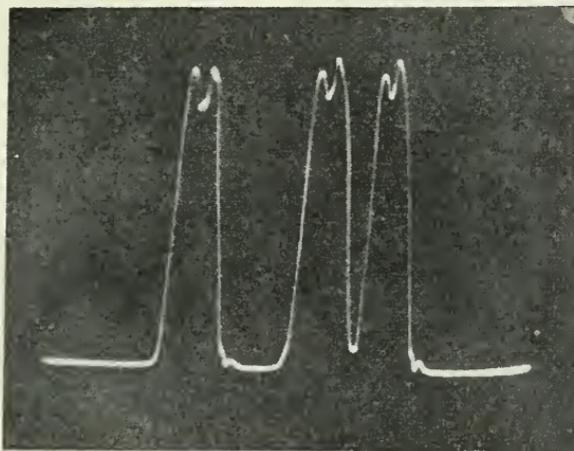


Fig. 8.—Linear trace oscillograph showing transmitted signal with 3 channels in operation 2 of which are adjacent.

suppressors in the manner described above and in the receiver by means of a combination of d-c. and two-phase a-c. potentials applied to the individual anodes. The potential of an anode was zero when the unwanted beam arrived and at or near 200 volts at the time of passage of the operating beam. The rotational frequency of the beam was sixty cycles and since both

stators were tied into the same source of power, no separate synchronizing means was necessary.

Figure 7 is a photograph of the cathode ray trace of the output of the transmitter tube when signals were being sent over three channels. A circular sweep circuit was used which distorted the signals somewhat. The shape of the pulses is shown better in Fig. 8 for which a linear sweep was employed. Signals were put on three channels, two of which were adjacent. The double-humped top of the pulse is caused by the window in the screen being slightly narrower than the beam width so that as the beam crosses the window, the greater densities in the edges relative to the center give this shape. A flat-topped pulse may be obtained by making the windows wider than the beam.

In conclusion the writer wishes to acknowledge his indebtedness to a number of his colleagues in the Laboratories for aid in the development of the tube. The 30-channel multiplex system was set up with the aid of Mr. W. H. T. Holden.

Abstracts of Technical Articles by Bell System Authors

*A Modification of Hallén's Solution of the Antenna Problem.*¹ M. C. GRAY. An alternative formula for the input impedance of a cylindrical antenna is derived from Hallén's integral equation. It is shown that the introduction of a variable parameter $Z(z)$ in place of Hallén's $\Omega = \log(4l^2/a^2)$ modifies the numerical results considerably, and leads to much better agreement with experimental evidence.

*Motor Systems for Motion Picture Production.*² A. L. HOLCOMB. The various types of motor systems and speed controls used in motion picture production are reviewed, evaluated, and the basic theory of operation described.

Motor drive systems are a fairly simple but important element in the production of motion pictures, but to many people who do not have direct contact with this phase of activities, the number of systems in use and their peculiarities are very confusing. Data on most of the different types of motors and motor systems in use have been published, but in different places and at different times so that no comprehensive reference exists. This paper is not intended as information on new developments or as a technical study, but rather as a review of all the major systems with an indication of their fields of greatest usefulness and with comments on both their desirable and undesirable features.

*A Dial Switching System for Toll Calls.*³ HOWARD L. HOSFORD. At Philadelphia, on the night of August 21st and the early morning hours of August 22, 1943, the cutover of the new #4 System was no mere episode; it was one of the milestones of telephone history. Intertoll dialing in itself is not new but this joint project of the Bell Telephone Company of Pennsylvania and the Long Lines Department is especially significant as it has been designed so as to extend the field of toll dialing by the operators to include the largest cities and joins together various types of dialing equipment. In its scope this project includes many points in an area reaching from Richmond, Va. to New York City and from Harrisburg, Pa. to Atlantic City, N. J.

From a traffic standpoint the #4 toll switching system actually comprises

¹ *Jour. Applied Physics*, January 1944.

² *Jour. S. M. P. E.*, January 1944.

³ *Bell Tel. Mag.*, Winter 1943-44.

three units, the switching equipment itself which is wholly mechanical, together with the so-called #4 and #5 switchboards. The #4 board is a cordless, key-typed call distributing board which is used in conjunction with the new switching system for such calls as must be given to an operator by offices not equipped for intertoll dialing. The operators at this board function as combined inward, through and tandem operators, thus eliminating the provision of separate units to provide these particular services. In brief, there is no basic difference between the essential operation of the #5 board and the conventional through board where delayed traffic is handled; however, operators handling calls at this board must make use of the new switching system to obtain both the calling and called offices by dialing.

Prior to the cutover the first trainees were given experience by handling some 300,000 test calls of every conceivable traffic characteristic. These were routed through the new system to break in the equipment and to shake down potential troubles. Two weeks prior to cutover a dress rehearsal was held, at which time about ten per cent of the circuits were put through their paces.

To provide information of value for future installations, arrangements were made for liberal provision of registers and meters to measure any and all phases of the various steps performed by the equipment. Some of these aids are not entirely new to telephone work but their application to toll, inward and through service is a departure.

The #4 System is running satisfactorily and both the equipment and the operators who use it deliver a high grade of service. Daily some 80,000 tandem, inward and through connections formerly handled by operators are routed through the equipment.

In connection with postwar planning, studies are now being made to determine future installations in order to take advantage of the possibilities of the new system. It is confidently expected that this will provide faster service on outward, inward and through calls and that transmission will be improved. These advantages should result in overall economies in outside plant and operating.

*Theoretical Limitation to Transconductance in Certain Types of Vacuum Tubes.*⁴ J. R. PIERCE. The thermal-velocity distribution of thermionically emitted electrons limits the low-frequency transconductance which can be attained in tubes in whose operation space charge is not important. A relation is developed by means of which this dependence may be evaluated for tubes employing electric and magnetic control. This relation is applied to deflection tubes with electric and magnetic control and to stopping-

⁴ *Proc. I. R. E.*, December 1943.

potential tubes. Magnetic control is shown to be inferior to electric control from the point of view of band-width and gain.

*Antenna Theory and Experiment.*⁵ S. A. SCHELKUNOFF. This paper presents: (1) a comparison between several approximate theoretical formulas for the input impedance of cylindrical antennas in the light of available experimental evidence; and (2) a discussion of the local capacitance in the vicinity of the input terminals, mathematical difficulties created by its presence, and methods of overcoming these difficulties. No exact solution of the antenna problem is available at present and so far it is impossible to set definite limits for errors which may be involved in various approximations. For this reason in appraising these approximations one is forced to rely on one's judgment and on experimental evidence. It is hoped that this paper will aid in correlating theory and experiment to the advantage of both.

⁵ *Jour. Applied Physics*, January 1944.

Contributors to this Issue

H. J. McSKIMIN, B.S. in Electrical Engineering, University of Illinois, 1937; M.S. in Physics, New York University, 1940. Bell Telephone Laboratories, 1937-. Engaged primarily in a study of electrical and electro-mechanical properties of piezoelectric crystals.

E. E. MOTT, Massachusetts Institute of Technology, B.S. 1927; M.S. 1928. General Electric Company, 1926-28. Bell Telephone Laboratories, 1928-. Mr. Mott has been engaged in telephone instruments research and development, particularly in connection with various types of telephone receivers and related devices. Since 1941 he has been engaged on war projects.

A. M. SKELLETT, A.B., 1924, M.S., 1927, Washington University; Ph.D., Princeton University, 1933; Instructor, 1927-28, Assistant Professor of Physics, 1928-29, University of Florida. Bell Telephone Laboratories 1929-. Dr. Skellett, formerly engaged in investigations pertaining to the transatlantic radio telephone, is concerned with applications of electronic and ionic phenomena.

R. A. SYKES, Massachusetts Institute of Technology, B.S. 1929; M.S. 1930. Columbia University, 1931-1933. Bell Telephone Laboratories, Research Department, 1930-. Mr. Sykes has been engaged in the applications of quartz crystals to broad-band carrier systems as filter and oscillator elements. Other work has included the application of coaxial lines as elements of filter networks and more recently the design and development of quartz crystals for radio frequency oscillators.

THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS
OF ELECTRICAL COMMUNICATION

- Effect of Telegraph Distortion on the Margins of Operation of Start-Stop Receivers *W. T. Rea* 207
- The Mounting and Fabrication of Plated Quartz Crystal Units *R. M. C. Greenidge* 234
- Effects of Manufacturing Deviations on Crystal Units for Filters *A. R. D'heedene* 260
- Mathematical Analysis of Random Noise . . *S. O. Rice* 282
- Abstracts of Technical Articles by Bell System Authors 333
- Contributors to this Issue 336

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

THE BELL SYSTEM TECHNICAL JOURNAL

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



EDITORS

R. W. King

J. O. Perrine

EDITORIAL BOARD

F. B. Jewett

M. R. Sullivan

O. B. Blackwell

O. E. Buckley

A. B. Clark

H. S. Osborne

S. Bracken

M. J. Kelly

F. A. Cowan



SUBSCRIPTIONS

Subscriptions are accepted at \$1.50 per year. Single copies are 50 cents each.
The foreign postage is 35 cents per year or 9 cents per copy.



Copyright, 1944
American Telephone and Telegraph Company

The Bell System Technical Journal

Vol. XXIII

July, 1944

No. 3

Effect of Telegraph Distortion on the Margins of Operation of Start-Stop Receivers

By W. T. REA

Recent practical and theoretical investigations of the effect of signal distortion on the margins of operation of start-stop telegraph receivers have led to the development of improved methods of testing and adjusting receivers, have enabled criteria of distortion tolerance to be set up for subscribers' and monitoring receivers and regenerative repeaters, and have made possible the application of more convenient and accurate standards of telegraph transmission. This paper describes the causes of distortion occurring both externally and internally to the receiver and the effects of such distortion on the operating margins. Methods of determining the internal distortion of a receiver are described and some of the more important considerations involved in establishing distortion tolerance criteria are discussed.

DURING the past decade the proportion of Bell System telegraph service operated on a start-stop teletypewriter basis has shown a continuous increase. Whereas in 1930 about 65% of telegraph long-distance circuit mileage was manual Morse, the present proportion of teletypewriter and teletypesetter service stands at 92%. The rapid growth of teletypewriter switching facilities has been an important factor in this development.

Naturally, this situation has made increasingly important a thorough understanding of the factors which affect the performance of start-stop receivers. In the present paper, an effort will be made to show some relationships between signal distortion and the operating margins of start-stop receivers.

A properly designed start-stop telegraph receiver requires only a small portion of the time of each signal element to permit a selection to be made; i.e. to determine whether the signal element in question is marking or spacing. The remainder of the signal element gives an operating margin, and serves as a reserve to take care of imperfections in the receiver or distortions which the telegraph signals may suffer in their passage over lines and through repeaters. The greater the signal distortion is, the smaller will be the margin which remains in the receiver to overcome the effect of such factors as wear of parts, variation of adjustments, or differences in speed between transmitter and receiver.

A consideration of the effects of telegraph distortion on the margins of

operation of start-stop receivers may well begin with a brief review of the nature and causes of the various types of distortion commonly experienced by telegraph signals. Telegraph distortion is generally considered to be divided into three types or components: bias, characteristic distortion, and fortuitous distortion.¹ The magnitude of the distortion is expressed in per cent of a unit pulse.

THE COMPONENTS OF TELEGRAPH DISTORTION

Bias, which is the simplest and most common component of distortion, may be positive (marking) or negative (spacing). Positive bias appears as a uniform lengthening of all marking pulses and an equal uniform shortening of all spacing pulses. Conversely, negative bias appears as a uniform lengthening of all spacing pulses and an equal uniform shortening of all marking pulses.

Bias is caused by an improper relation between the levels at which the relay or other receiving device responds and the steady-state marking and spacing levels of the signal. For example, Fig. 1(B) shows the signals of Fig. 1(A) as they might appear as a symmetrical wave on a line. With such a wave zero bias will be received when the currents at which the receiving relay operates from spacing to marking and from marking to spacing are symmetrically located with respect to the average of the steady-state marking and spacing currents. That is, zero bias will be received if the relay operates from spacing to marking and from marking to spacing at *B-B*, or if the relay operates from spacing to marking at *A-A* and from marking to spacing at *C-C*, or if the relay operates from spacing to marking at *C-C* and from marking to spacing at *A-A*. Negative bias will be received if the relay operates in both directions at *A-A*, and positive bias will be received if it operates in both directions at *C-C*.

In Fig. 1(C) is shown an unsymmetrical wave, in which the transient from space to mark is more rapid than that from mark to space. In this case, positive bias will result when the relay operates in both directions at *B-B* or at *C-C*, but no bias will result if the relay operates in both directions at *A-A*.

In the remaining diagrams of Fig. 1 it is assumed that the relay operates in both directions at a level midway between the steady marking and spacing levels. Fig. 1(D) shows a wave in which the transients are of such duration that the steady-state value is not attained in the shortest pulse length. It will be seen that the operation of the relay is delayed less after a short pulse than after a long one, and that this is true whether the pulse be marking or spacing. This effect is known as *negative characteristic distortion*, and it tends to shorten short pulses and lengthen long pulses. When a series of unbiased dots (called telegraph reversals) is transmitted, a steady-state condition is reached, in which the delays become equal on all transitions.

Hence, the signals are received as sent. When biased reversals are transmitted, the longer pulses are further lengthened and the shorter pulses are further shortened, causing the bias of the received signals to be of greater magnitude than that of the transmitted signals.

Fig. 1(E) shows a wave in which the current overswings the steady-state value, and fails to complete the return to steady state within the duration of the shortest pulse. It will be seen that the operation of a relay will be

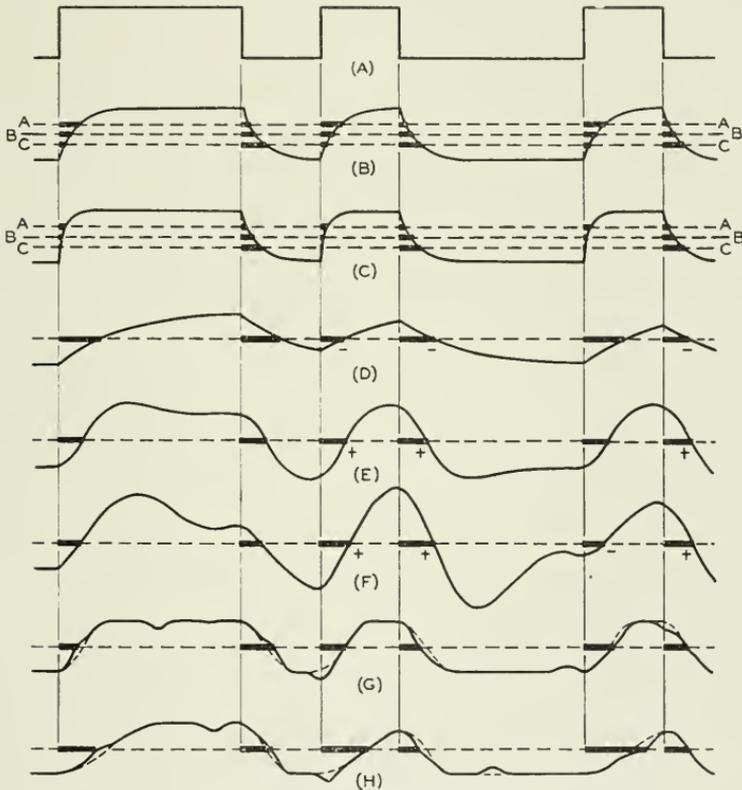


Fig. 1—Signal diagrams illustrating causes of distortion.

delayed more after a short pulse than after a long one, and that this is true whether the pulse in question be marking or spacing. This effect is known as *positive characteristic distortion*, and it tends to shorten long pulses and lengthen short ones. When unbiased reversals are transmitted, a steady-state condition is reached, in which the delays become equal on all transitions. Hence, the signals are received as sent. When biased reversals are transmitted, the shortening of the long pulses and lengthening of the short pulses causes the bias of the received signals to be less than that of the transmitted signals.

Fig. 1(F) shows a wave which performs a damped oscillation before settling to a steady state. This type of wave tends to produce a negative characteristic effect on certain transitions and a positive characteristic effect on others.

In general, if, on a given transition, the sum of all previous transients is such as to delay the operation of the receiving device, positive characteristic distortion is said to occur. If, on the other hand, the sum of all previous transients is such as to advance the operation of the receiving device, negative characteristic distortion is said to occur.

Bias and characteristic distortion, considered together, are called "systematic" distortion, because they occur with some regularity, and obey certain constant laws. There is another type of distortion that is not systematic. This is known as *fortuitous distortion*. It may be caused by the effect of various interfering currents on the receiving device. Fig. 1(G) shows a wave upon which interfering currents have been superposed. It will be noted that, for a given magnitude of interfering current, the more sloping the wave is in the region of the operating level of the receiving device, the greater will be the resulting fortuitous distortion.

Fortuitous distortion may also occur, in cases of extremely sloping wave-shape, due to the "indecision" of the receiving device, or, in other words, due to small variations of its effective operating level from signal to signal.

Fig. 1(H) shows a wave that is affected by interfering currents and in which the mark-to-space and space-to-mark transients have different slopes in the region of the operating level of the receiving device. The interfering current therefore causes fortuitous distortion of different magnitudes on mark-to-space and space-to-mark transitions. It will be shown later that distortion of this type affects a start-stop receiver in a particular manner which differs from the effect of distortion of the type illustrated in Fig. 1(G).

These, then are the generally-recognized components of telegraph distortion. More complicated effects ensue when characteristic distortion occurs on waves having dissimilar transients in the mark-to-space and space-to-mark directions, but a consideration of such phenomena is outside the scope of an elementary explanation of telegraph distortion, and is not necessary to an understanding of the effects of distortion on the margins of operation of start-stop receivers.

START-STOP DISPLACEMENTS

The basic principles of operation of start-stop receivers have been described in previous articles^{2,3}. A brief review of these principles will, therefore, suffice here.

The start-stop signal train consists of a start pulse, which is generally spacing, several selective pulses, each of which may be either marking or

spacing, and a stop pulse which is generally marking. The receiving mechanism is started by the transition at the beginning of the start pulse, and its speed is such that it arrives at the stop position before the end of the stop pulse occurs, and remains stopped until the succeeding start transition takes place. Thus any speed difference between the transmitter and receiver is prevented from cumulating for more than the duration of one signal train.

Since the receiving device starts anew at each start transition, and the instants of selection of the selective pulses are spaced in time relative to the instant of starting, as shown in Fig. 2(A), the start transition acts as a basic reference point to which all other instants of time during the selective cycle may be referred.

The advances and delays of the transitions of the start-stop signal train from their normal times of occurrence, relative to the start transition, are known as "start-stop displacements." Fig. 2(B) shows the four types of displacement that may occur: *MB* or "marking beginning displacement," which is the advance of a space-to-mark transition (beginning of a marking pulse) relative to the start transition; *SB* or "spacing beginning displacement," which is the delay of a space-to-mark transition relative to the start transition; *SE* or "spacing end displacement," which is the advance of a mark-to-space transition (end of a marking pulse) relative to the start transition; and *ME* or "marking end displacement," which is the delay of a mark-to-space transition relative to the start transition.

Effect of Bias on Displacement

Since bias affects all pulses alike, and since in the usual start-stop receiver the start transition is mark-to-space, the succeeding mark-to-space transitions of the signal train are not shifted relative to the start transition. Hence the total effect of the bias appears on the space-to-mark transitions. Positive bias causes MB displacement alone, as shown in Fig. 2(C). Negative bias causes SB displacement alone, as illustrated in Fig. 2(D).

The total range through which the selective periods may be shifted, relative to the start transition, without producing an incorrect selection is known as the orientation range of the receiver. Its limits are read on a scale calibrated from 0 to 100 in per cent of a unit pulse-length. Figure 3 is a graph of teletypewriter orientation range versus input signal bias, for a receiver whose range is from 10 to 90 on unbiased signals. Diagrams of this type are called "bias parallelograms."

Effect of Characteristic Distortion on Displacement

Characteristic distortion does not affect all pulses of miscellaneous signals alike, because, as explained above, the effect on each transition depends

upon the signal combinations that have previously been sent over the circuit. Hence the start transition and the transitions occurring between selective pulses are, in general, delayed by varying amounts. All four types of displacement shown in Fig. 2(B) occur, depending upon whether the transition in question is mark-to-space or space-to-mark and whether it has been

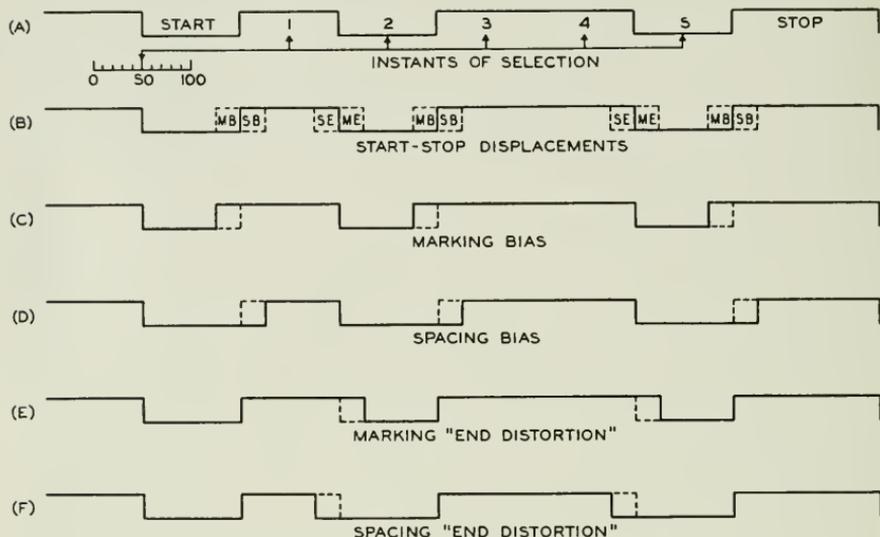


Fig. 2—Diagrams illustrating start-stop displacements.

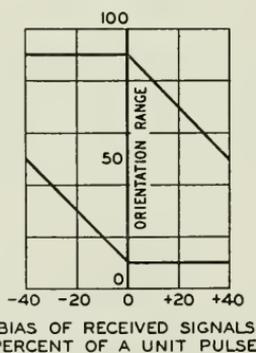


Fig. 3—The bias parallelogram.

delayed more or less than the start transition. For example, if a space-to-mark transition is delayed less (on an absolute time basis) than the start transition, *MB* displacement occurs; if more, *SB* displacement. If a mark-to-space transition is delayed less than the start transition, *SE* displacement occurs, if more, *ME* displacement.

Maximum Displacements Caused by Characteristic Distortion

The *maximum MB displacement* will occur when the start transition is delayed as much as possible and some space-to-mark selective transition is delayed as little as possible. This will take place, in the case of negative characteristic distortion, when as long a marking signal as is possible precedes the start transition and a combination of pulses as predominantly marking as possible precedes the space-to-mark transition in question. A marking signal sufficiently long to permit a steady state to be attained, followed by any signal train having the first selective pulse marking satisfies this condition, as shown at "X" in Fig. 4(B), but it will be noted that the *MB displacement* extends into the start pulse, where, in the case of a start-stop receiver, no selection is made. Hence it will not affect the margin of operation of the receiver, provided it is not so large as to prevent the receiver from starting. This particular distortion will, however, affect a start-stop distortion measuring set⁴ or regenerative repeater which is so designed that measurements or selections are made during both the selective pulses and the start pulse. As far as a start-stop teletypewriter, in which no selection occurs during the start pulse, is concerned, the *maximum MB displacement* occurs on the fourth transition of the letter *K* following as long a marking signal as possible, as shown at "Y" in Fig. 4(B). This space-to-mark transition, being preceded by a spacing pulse of unit length which, in turn, was preceded by signals which are predominantly marking, is delayed for a short time, whereas the mark-to-space start transition, which was preceded by a long marking signal, is delayed for a longer time. Except in the case of unusual wave forms, there will be very little difference between the magnitudes of the displacements shown at "X" and "Y" unless they are both very large, since the wave will usually attain steady state during the steady marking interval constituted by the first, second, third and fourth selective signal intervals.

In the usual case of positive characteristic distortion, the *maximum MB displacement* will occur when the start transition is preceded by a combination of pulses as predominantly spacing as possible, and some space-to-mark transition is preceded by the longest spacing signal possible in the start-stop code. These conditions are met by repeated, "BLANK" signal trains, shown in Fig. 4(E).

The *maximum SB displacement* will occur when the start transition is delayed as little as possible, and some space-to-mark selective transition is delayed as long as possible. This takes place, in the case of negative characteristic distortion, when a combination of pulses as predominantly spacing as possible precedes the start transition and the longest possible spacing signal precedes the space-to-mark transition in question. As noted in the

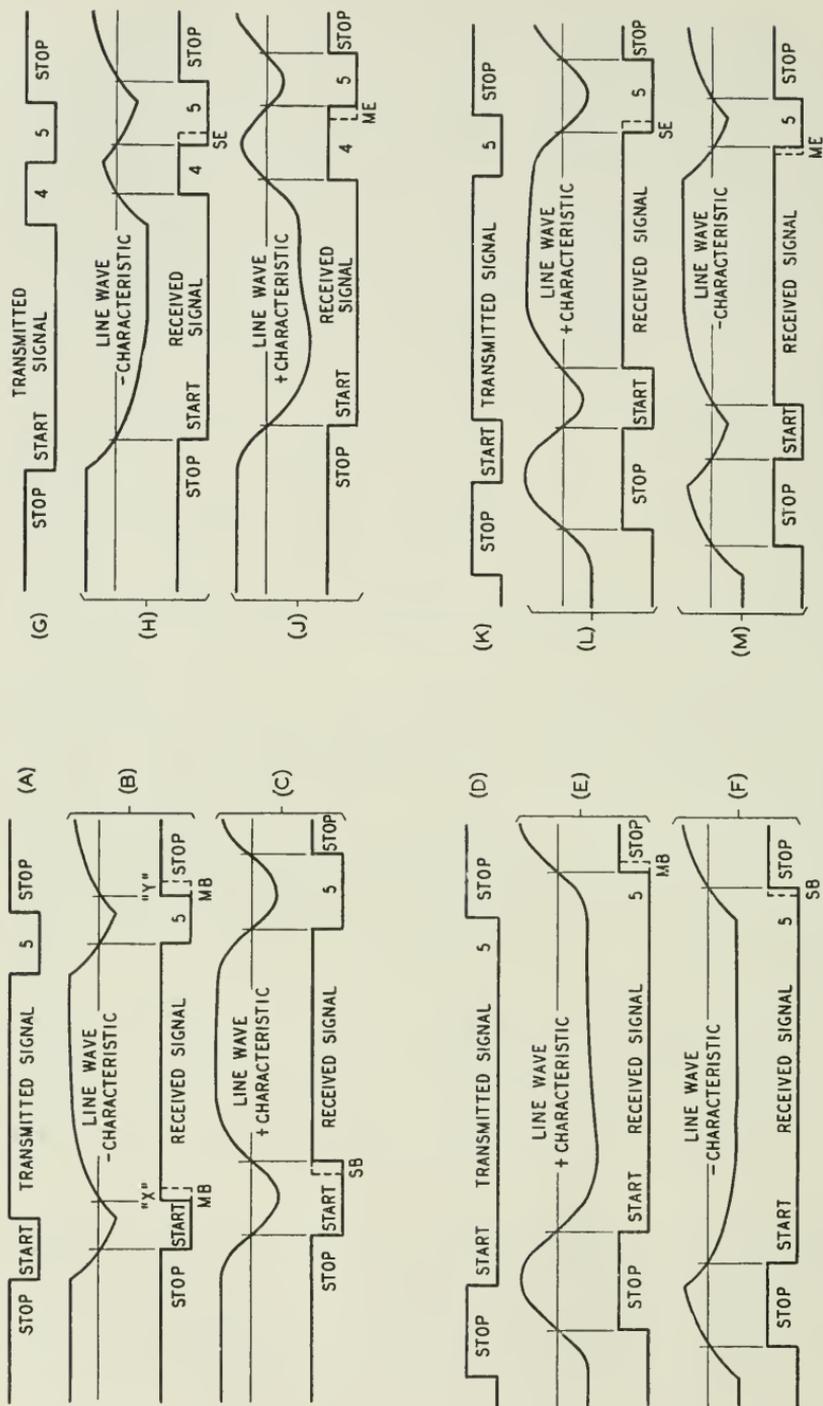


Fig. 4—Characteristic distortion.

preceding paragraph, repeated "BLANK" signals satisfy this description. Fig. 4(F) illustrates the effect of negative characteristic distortion on this signal combination. It will be noted that the resulting SB displacement extends into the stop pulse, where usually no selection takes place, and hence it would not affect the margin of operation of the start-stop receiver except, of course, in the case of a type of receiver, such as a regenerative repeater, in which selection of the stop pulse does occur. For the case of a receiver which does not select the stop pulse, the maximum SB displacement affecting the margin of operation occurs at the second transition of the letter "T" following repeated "BLANK" signals. Except in the case of very large distortions, this displacement will be of nearly the same magnitude as that illustrated in Fig. 4(F).

In the usual case of positive characteristic distortion, the maximum SB displacement will occur when the start transition is preceded by a long marking signal and some space-to-mark selective transition is preceded by a combination of pulses as predominantly marking as possible. As noted previously, this description is satisfied by a sufficiently long marking signal to permit the attainment of steady state, followed by any signal train having the first selective pulse marking. Figure 4(C) illustrates the effect of positive characteristic distortion on this type of signal.

The *maximum SE displacement* will occur when the start transition is delayed as much as possible and some mark-to-space selective transition is delayed as little as possible. This will take place, in the case of negative characteristic distortion, when a long marking signal precedes the start transition and a combination of pulses as predominantly spacing as possible precedes the mark-to-space transition in question. This description is answered by a long marking signal followed by a "CARRIAGE RETURN" signal train, as shown in Fig. 4(H). The SE displacement occurs at the end of the fourth selective pulse.

In the usual case of positive characteristic distortion, the maximum SE displacement will occur when the start transition is preceded by a combination of pulses which is as predominantly spacing as possible, and some mark-to-space selective transition is preceded by the longest possible marking signal. This description is satisfied by repeated "BLANK" signal trains followed by the letter "K," and, as shown in Fig. 4(L), the SE displacement occurs at the end of the fourth selective pulse.

The *maximum ME displacement* will occur when the start transition is delayed as little as possible and some mark-to-space selective transition is delayed as much as possible. This will take place, in the case of negative characteristic distortion, when the start transition is preceded by a combination of pulses which is as predominantly spacing as possible, and some mark-to-space selective transition is preceded by the longest possible marking

signal. As noted in the previous paragraph, the letter "K" preceded by repeated "BLANK" signal trains satisfies this description, and the ME displacement is as illustrated in Fig. 4(M).

In the usual case of positive characteristic distortion, the maximum ME displacement will occur when the start transition is preceded by a long marking signal, and some mark-to-space selective transition is preceded by a combination of pulses as predominantly spacing as possible. As seen previously, this description is answered by a long marking signal followed by a "CARRIAGE RETURN" signal train. Fig. 4(J) illustrates the ME displacement.

Effect of Characteristic Distortion on Orientation Limits

In the usual start-stop system which employs a stop pulse longer than the unit selecting pulse, characteristic distortion affects the upper and lower limits of orientation differently. This effect is due mainly to the longer stop pulse, although the fact that the start transition is always mark-to-space contributes to the effect.

In the case of negative characteristic distortion, the displacements (MB and SE) which affect the upper end of the orientation range are those in which the start transition suffers a long delay and a selective transition suffers a short delay. The delay of the start transition can be quite large, since it may be preceded by a long marking pulse. Moreover, the delay of the selective transition may be very short, since the pulse which precedes the transition can be of unit length, and this, in turn, may be preceded by a signal of the opposite type which may be of as much as four units duration. Hence these displacements, being the difference between a large and a small delay, are large.

On the other hand, the displacements, SB and ME , which affect the lower end of the range are those in which the start transition suffers only a fairly short delay and a selective transition suffers a long delay. The delay of the start transition can not be very short for two reasons: first, the start pulse cannot be preceded by a steady spacing pulse; and second, what is of more importance, the stop pulse is of greater than unit length. The delay of a selective transition can be long, as when the transition is preceded by a pulse of four or five units in length. (This delay may not be so long as that suffered by a start transition which follows a steady-state marking condition, but it is not much shorter.) Hence the SB and ME displacements, being the difference between a long selective transition delay and only a fairly short start transition delay, are smaller than the MB and SE displacements.

For this reason negative characteristic distortion affects the upper end of the range more than it does the lower.

In the case of what we have termed "the usual type of positive characteristic distortion," the displacements (SB and ME) which affect the lower end of the orientation range are those in which the start transition suffers a short delay and a selective transition suffers a long delay. The delay of the start transition can be quite short, since it may be preceded by a long marking signal. Moreover, the delay of the selective transition may be very long, since the pulse which precedes the transition can be of unit length and this, in turn, may be preceded by a signal of the opposite type which may be four or more units in length. Hence these displacements, being the difference between a short and a long delay, are large.

On the other hand, the displacements (MB and SE) which affect the upper end of the range are, in this type of distortion, those in which the start transition suffers only a fairly long delay and a selective transition suffers a short delay. The delay of the start transition cannot be very short for the two reasons mentioned previously. The delay of the selective transition can be short, as when the transition is preceded by a pulse four to six units in length. Hence the MB and SE displacements, being the difference between a short selective transition delay and only a fairly long start transition delay, are smaller than the SB and ME displacements.

For this reason positive characteristic distortion of this type affects the lower end of the range more than it does the upper.

In the case of a wave which oscillates, causing positive characteristic distortion on some transitions and negative on others, no such general statements as are made above are applicable. In practice, cases have been observed in which one end of the orientation range was cut and the other was actually extended.

Due to the fact that characteristic distortion delays the start transition by different amounts from character to character, it causes the character length to vary during continuous automatic transmission. The maximum variation in character length is roughly of the same magnitude as the maximum displacement affecting the selective pulses.

Effect of Fortuitous Distortion on Displacement

Fortuitous distortion causes the start transition to be delayed more or less than normal, and has the same effect on the selective transitions. Since it is usually equally probable that the maximum fortuitous effects will occur on mark-to-space or space-to-mark transitions and will increase or decrease their delay, this type of distortion generally produces the four types of displacement in equal magnitude, and this magnitude is equal to the maximum increase or decrease in the length of pulse.

An exception to the above statement occurs when the mark-to-space and space-to-mark transients give the wave different slopes at the point where the

receiving device operates. Then the magnitude of the fortuitous effect is different on mark-to-space and space-to-mark transitions. If the effect is greater on the space-to-mark transitions, MB and SB displacements are greater than SE and ME . If the opposite, SE and ME are greater than MB and SB . In all cases, however, the orientation range is reduced equally at both ends.

Fortuitous distortion also lengthens and shortens the character since it does not affect all transitions alike.

INTERNAL DISTORTION

Telegraph signal distortion may occur within the start-stop receiver and it should be expected that the components of distortion will have the same effect on the margins of operation as the same components external to the receiver. Consequently, it should be possible to determine the magnitudes of the various components of internal distortion by their effects on the margins of operation.

As mentioned previously, the upper end of the orientation range is determined by whichever of the displacements MB and SE is the greater; and the lower end by whichever of the displacements SB and ME is the greater. To discover the magnitude of the smaller type of displacement it is necessary to reduce the larger displacement by distorting the transmitted signals. For example, if a receiver has a large internal marking bias, the upper limit of orientation is determined by MB displacement, and hence the amount of SE displacement caused by internal distortion is concealed. However, by transmitting signals affected by SB displacement (in other words, signals biased to spacing), the total MB displacement is decreased until it is less than the internal SE displacement, whose effect on margin can then be found. Thus the internal distortion may be determined by observing the effect of external distortion on the margins of operation.

It is convenient to regard any start-stop receiver as a theoretically perfect receiver affected by certain types of internal distortion. The internal distortion is usually considered to be composed of bias, "skew" (defined later) and fortuitous distortion. (The internal characteristic distortion is generally included in "internal fortuitous distortion," since it is usually very small, and a fairly elaborate testing procedure is required to separate its effects from those of internal fortuitous distortion.) Internal bias and internal fortuitous distortion are of the same nature as the external effects previously described. Skew is said to occur when there exists the type of distortion, previously mentioned, in which the fortuitous effect on space-to-mark transitions differs in magnitude from that on mark-to-space transitions. When the former is greater the skew is said to be positive; when the latter, negative. Hence in positive skew, MB and SB displacements tend to

be larger; in negative skew, ME and SE displacements tend to be larger. The magnitude of the skew is defined as the difference between the magnitudes of the fortuitous effects on space-to-mark and mark-to-space transitions.

Figure 3 showed the bias parallelogram of a receiver which had a local margin of 10 to 90. Figure 5 shows the bias parallelogram of a perfect

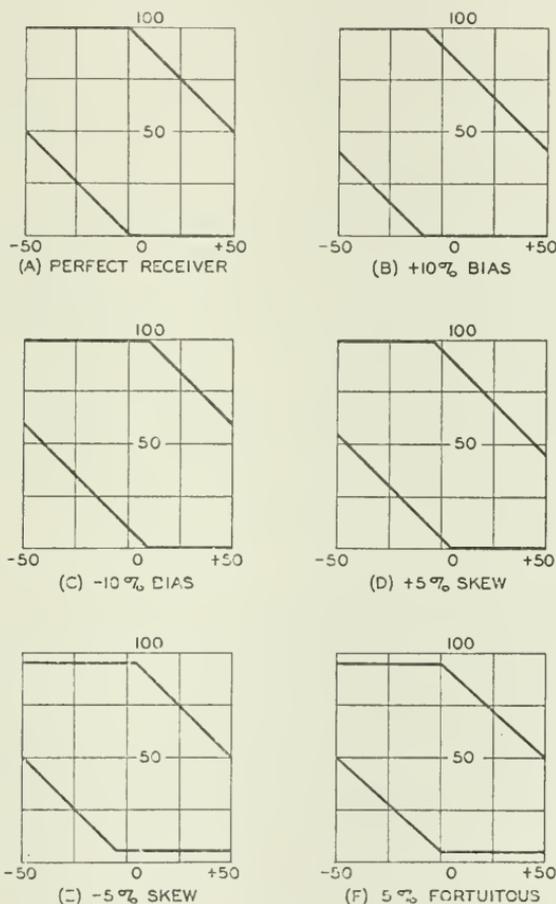


Fig. 5—Effect of internal distortion on bias parallelogram.

receiver and illustrates how the components of internal distortion affect the shape of the bias parallelogram. The skewing of the corners of the parallelograms shown in Fig 5(D) and (C) led to the use of the term “skew” for this effect.

In telegraph transmission systems skew may be caused by the effect of interference on a wave which has different slopes during mark-to-space and

space-to-mark transitions. It may also result from an equivalent electro-mechanical effect in a start-stop receiver, as will be described later.

Measurements of Receiver Distortion Tolerance

In measurements of the distortion tolerance of start-stop receivers there is used a distributor which is arranged to transmit signals having any of the four types of displacement *MB*, *SB*, *SE* and *ME*. Positively biased signals are transmitted for *MB* displacement and negatively biased signals for *SB* displacement. The test signals having *SE* or *ME* displacement are said to be affected by "end distortion." These differ from any experienced on transmission circuits in that only the mark-to-space transitions of the selective pulses are shifted relative to the start transition, being delayed for *ME* displacement and advanced for *SE* displacement, as shown in Fig. 2(E) and (F). "End distortion" simulates the mark-to-space displacements produced by characteristic and fortuitous distortion, and it has been found in practice that it yields results which enable a receiver's tolerance to these components of distortion to be predicted with a high degree of accuracy.

When fixed values of displacement are transmitted, the limits of orientation are measured by means of the range scale of the receiver. Alternately, a distributor may be used in which the magnitude of displacement may be continuously varied, and this enables measurements of internal distortion to be conducted with the orientation fixed, or, indeed, on receivers having no means or a limited means of varying the orientation.

Orientation Settings for Best Tolerance to Test Distortions

Obviously, the best orientation setting is that which permits the receiver to tolerate the greatest amount of any distortion which is expected. If all four types of displacement are considered equally likely, the orientation should be set at that point at which the minimum tolerance to any type of displacement is as large as possible. For example, consider a receiver which, with an orientation setting of 49, has the following tolerances to test displacements:

<i>MB</i>	44
<i>SB</i>	38
<i>SE</i>	42
<i>ME</i>	44

Let the orientation setting be raised 2 per cent, to 51. Then the tolerances are as follows:

<i>MB</i>	42
<i>SB</i>	40
<i>SE</i>	40
<i>ME</i>	46

The shift of orientation has increased the minimum tolerance (spacing bias) from 38 to 40. Any further shift would make the tolerance to spacing "end distortion" less than the tolerance to spacing bias. This setting is called the "center of fortuitous distortion tolerance," since at this point the receiver will tolerate the maximum amount of fortuitous distortion.

If, on the other hand, bias is considered more probable than distortions which produce "end distortion" effects, the orientation might be adjusted to the point at which the tolerances to marking and spacing bias are equal. For example, suppose the orientation setting of the receiver under consideration were raised 1 per cent to 52. The tolerances would then be

<i>MB</i>	41
<i>SB</i>	41
<i>SE</i>	39
<i>ME</i>	47

This setting is called the "center of bias tolerance," since at this point the receiver will tolerate the maximum amount of bias regardless of the sign of the bias.

There is one more setting that is of interest. It is that at which the tolerances to marking and spacing "end distortion" are equal. Suppose the orientation of the receiver were lowered 4 per cent to 48. The tolerances would then be

<i>MB</i>	45
<i>SB</i>	37
<i>SE</i>	43
<i>ME</i>	43

This setting is called the "center of end distortion tolerance," since at this point the receiver will tolerate the maximum amount of "end distortion" regardless of its sign.

Calculation of Components of Internal Distortion

Figure 6 illustrates how the components of internal distortion are determined from measurements of distorted signals. Each diagram shows a portion of a teletypewriter character consisting of a start pulse, a marking selective pulse and a spacing selective pulse. The solid lines show an undistorted signal. The dashed lines show the displacement of a transition due to internal bias. The shaded area defines the fortuitous effect which is skew; that is, the transition in question may fall anywhere within the shaded area during repeated transmission of the signal. The arrows below the figure show the extent of the displacement occurring on each transition due to the presence of a given displacement of the transmitted signals. The four types of displacement are of equal magnitude D . The arrows above the diagram designated L_B and L_E show the lower limits of orientation with, respectively,

spacing bias and marking "end distortion" (*SB* and *ME* displacements). The arrow U_B and U_E show the upper limits or orientation with, respectively, marking bias and spacing "end distortion" (*MB* and *SE* displacements).

Figure 6(A) shows the case of positive internal bias and positive skew; Fig. 6(B), positive bias and negative skew; Fig. 6(C), negative bias and posi-

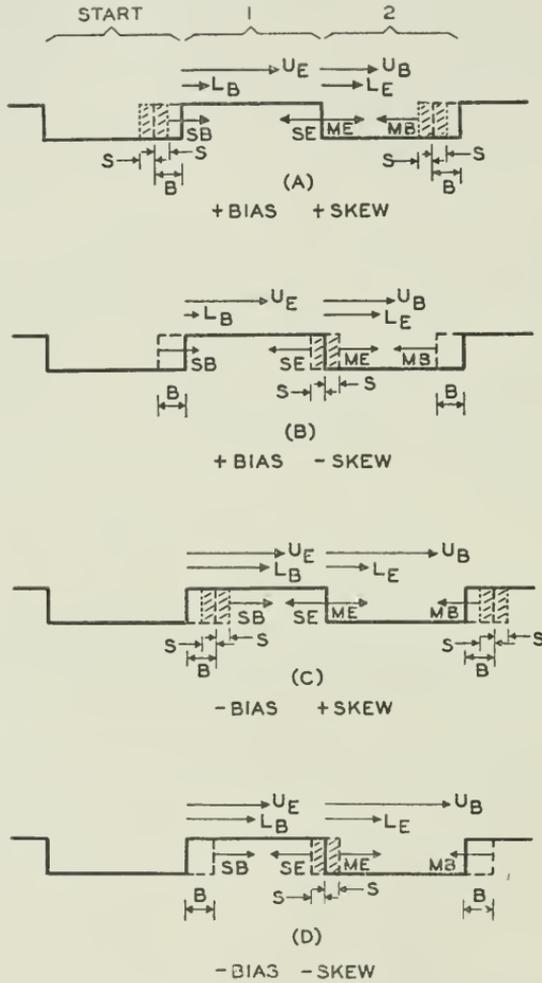


Fig. 6—Use of distorted test signals in measuring internal distortion.

tive skew; and Fig. 6(D), negative bias and negative skew. The following relationships hold, bearing in mind that $MB = SB = SE = ME = D$:

Fig.	Eias	Skew	L_B	U_B	L_E	U_E
(A)	+	+	$D+s-b$	$1-b-s-D$	D	$1-D$
(B)	+	-	$D-b$	$1-b-D$	$D+(-s)$	$1-(-s)-D$
(C)	-	+	$D+s+(-b)$	$1+(-b)-s-D$	D	$1-D$
(D)	-	-	$D+(-b)$	$1+(-b)-D$	$D+(-s)$	$1-(-s)-D$

In any figure $L_B - L_E = s - b$

and $U_B - U_E = -s - b$

Adding and subtracting, we find that:

$$\text{Internal bias} = \frac{U_E + L_E}{2} - \frac{U_B + L_B}{2}$$

$$\text{Skew} = \frac{U_E - L_E}{2} - \frac{U_B - L_B}{2}$$

Any $\frac{U + L}{2}$ is the center of an orientation range. Hence it may be stated that the internal bias is equal to the difference between the *centers* of tolerance to "end distortion" and bias. It will also be noted that any $\frac{U - L}{2}$ is half of an orientation range. When the test signal displacements determining the range limits are equal, the amount of tolerance equals $\frac{U - L}{2} + D$ (assuming no curvature in the distortion parallelogram). Hence the skew is equal to the difference between the *amounts* of tolerance to "end distortion" and bias.

For example, the receiver cited previously has the following characteristics

$$\text{Internal bias} = 48 - 52 = -4\%$$

$$\text{Skew} = 43 - 41 = +2\%$$

Incidentally, this means that internal bias does not reduce the total bias tolerance of a receiver, but merely shifts the center of bias tolerance with relation to the center of "end distortion" tolerance. Hence the effects of internal bias may be compensated for, as far as the bias tolerance of the receiver is concerned, by setting the orientation at the center of bias tolerance. However, internal bias does reduce the minimum "end distortion" tolerance of a receiver whose orientation is adjusted to the center of bias tolerance.

"Switched" Bias

When biased signals are produced by the action of a biasing current on a relay driven by a symmetrical wave, and the sign of bias is suddenly reversed during the transmission of a teletypewriter character, all the succeeding transitions of that character are affected, not by bias, but by "end distortion." This is shown in Fig. 7, of which (A) shows the original unbiased signals, (B) shows the signals affected by bias which changes from positive to negative at time T , and (C) shows the effect on the same signals when the bias is changed from negative to positive.

Signals such as these, in which the sign of bias is changed at intervals, are said to be affected by "switched bias." Since all four types of displace-

ment are present in equal magnitude in switched bias signals, the effect on a start-stop receiver resembles that of fortuitous distortion. Thus the center of switched bias tolerance is the center of fortuitous distortion tolerance and the amount of switched bias tolerance is the amount of fortuitous distortion tolerance. This center is also the center of orientation in a receiver having no curvature or symmetrical curvature of the displacement-vs.-orientation-limit characteristic. The switched bias tolerance is, of course, one-half the orientation range in a receiver having no curvature of the characteristic.

In actual field practice, switched bias signals, applied at a central office, are used as a test of tolerance of the teletypewriter at a subscriber station in combination with the subscriber loop. They provide a more accurate measure of transmission capabilities than an orientation range measurement with undistorted signals from the central office, since not only is the curvature of the distortion parallelogram taken into account, but the character

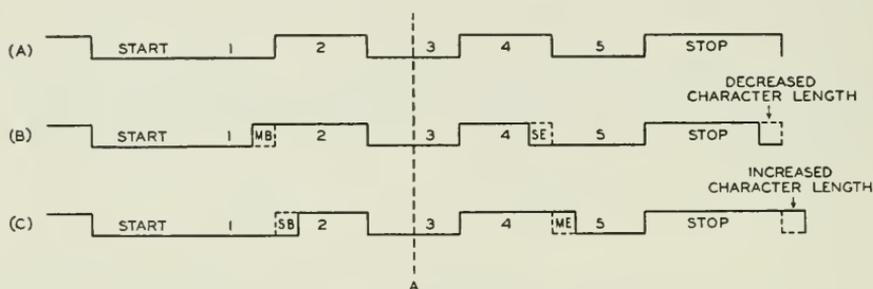


Fig. 7—Switched bias.

length changes in much the same manner as in signals affected with characteristic or fortuitous distortion.

The components of internal distortion of a receiver may be estimated from bias and switched bias measurements, but they cannot be accurately specified thereby. Figure 8 illustrates the difficulty in separating bias and skew by means of measurements of the difference between the amounts and centers of tolerance to steady bias and switched bias. Figure 8(A) shows the bias and end displacement parallelograms of a receiver having +24 per cent bias and +16 per cent skew. The center of tolerance to switched bias is 4 per cent above the center of steady bias tolerance and the steady bias tolerance is 4 per cent greater than the switched bias tolerance. Figure 8(B) shows the parallelograms of a receiver having +4 per cent bias and -4 per cent skew. Again, the center of tolerance to switched bias is 4 per cent above the center of steady bias tolerance and the steady bias tolerance is 4 per cent greater than the switched bias tolerance.

Of course, the components of internal distortion can be measured by

observing both ends of the orientation range with positive and negative bias rather than observing the upper end with positive bias and the lower end with negative bias. This type of measurement is merely equivalent to using a fairly large percentage of bias and zero per cent of end distortion. The disadvantage of this measurement is that no account is taken of the curvature

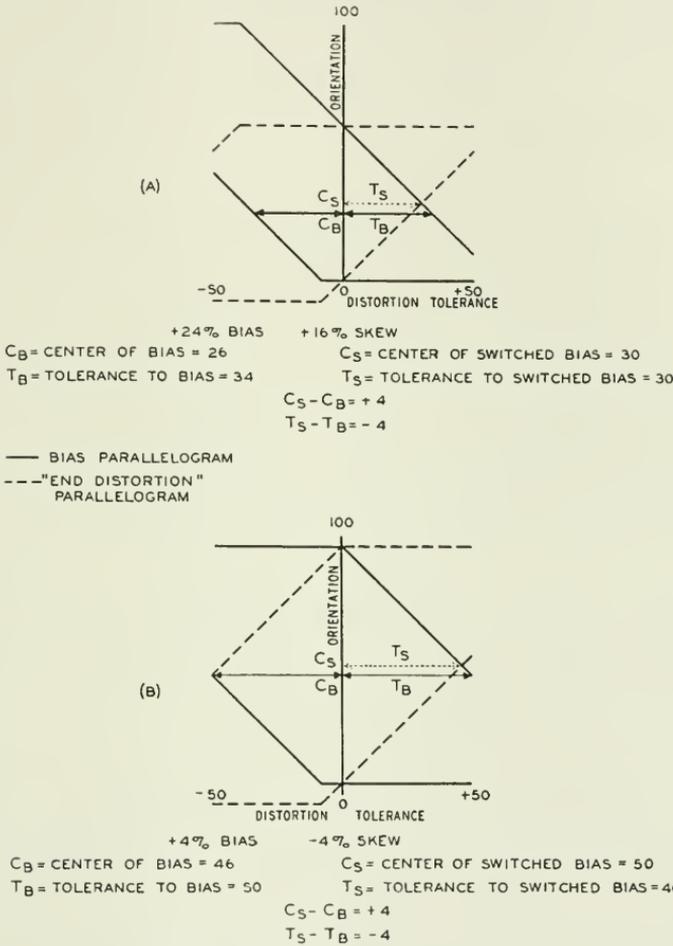


Fig. 8—Switched bias measurements

of the end displacement parallelogram, and hence the indicated values of tolerance may not be an accurate measure of the receiver's ability to receive distorted signals.

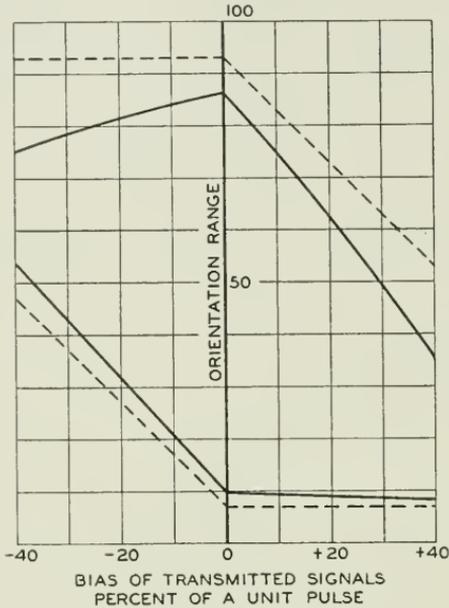
Internal Fortuitous Distortion

It is usually considered, in measurements of miscellaneous signals, that the difference between the maximum distortion tolerance and 50 per cent (the

latter being the tolerance of a perfect receiver) is due to internal fortuitous effects, even though part of it may be due to the effects of internal characteristic distortion. Hence the internal fortuitous distortion is usually defined as the difference between 50 and the tolerance to bias or end distortion, whichever of the latter may be the larger.

For example, in the sample receiver considered on page 220, the internal fortuitous distortion is:

$$50 - 43 = 7 \text{ per cent}$$



----- WITHOUT CHARACTERISTIC DISTORTION
 ————— WITH NEGATIVE CHARACTERISTIC DISTORTION

Fig. 9—Effect of negative characteristic distortion on bias parallelogram.

Internal Characteristic Distortion

In practice it is found that the relation between displacement and reduction of margin is sometimes not strictly linear. Especially at large values of displacement, the reduction in margin is often greater than the displacement causing it. This effect is due to internal characteristic distortion, which causes an increase in the distortion of shortened pulses. Internal characteristic distortion, like any other form of characteristic distortion, is caused by the failure of some circuit or mechanical element to attain steady state before the occurrence of a succeeding transition. Figure 9 shows an example

of the bias parallelogram of a receiver suffering from internal negative characteristic distortion.

SOME CONSIDERATIONS INVOLVED IN THE MEASUREMENT AND ADJUSTMENT OF START-STOP RECEIVERS

Because of the effects of characteristic distortion, it cannot be assumed that the ultimate tolerance of a receiver is equal to the sum of the displacement of the received test signals and one-half the remaining orientation range, especially if the latter is large. To attain accurate results, the ultimate tolerance must be measured with the orientation adjusted to the center of tolerance.

For the same reason (the curvature of the "parallelogram" caused by internal characteristic distortion) measurements of internal distortion on a receiver which is, itself, to be used to measure distortion should be made with displacements of approximately the same magnitude as the distortions which the receiver is to measure. In a receiver which is to be used to measure small distortions, we are interested in the properties of the linear portion of the parallelograms. Hence we measure the receiver's internal bias and skew using small amounts of displacement in the measuring signals. The internal fortuitous distortion may generally be neglected, since it does not affect the shape, but only the size, of the distortion-vs-margin characteristic.

On the other hand, in a receiver which is to be used for receiving signals we are interested not so much in the shape of the characteristic as in the ultimate tolerance to telegraph distortion at an optimum setting of the orientation mechanism. For this reason, a receiver destined for service use is best tested with signals containing fairly large displacements. Internal fortuitous distortion is deleterious in such a receiver, since it decreases the tolerance to displacement of all kinds. Skew, depending upon its sign, affects the tolerance to either space-to-mark or mark-to-space displacements.

It should be realized that the removal of skew does not necessarily improve a service receiver. In the case of bias or characteristic distortion the introduction of distortion of a given sign will remove internal distortion of the opposite sign, and thus improve the performance of the receiver. But since skew is the difference between two fortuitous distortion effects, it may be removed either by reducing the larger or increasing the smaller effect. The former procedure will increase the receiver's total tolerance to distortion, whereas the latter will reduce it.

In practice bias tolerance is generally considered to be more desirable than "end distortion" tolerance. The reason for this is that most transmission circuits suffer from some bias (of unpredictable sign and amount) which uses up some of the receiver's bias tolerance but none of its "end distortion"

tolerance. This is why the orientation of a service receiver is generally adjusted to the center of bias tolerance, and small amounts of internal bias or negative skew are not considered objectionable, since they do not affect the tolerance to bias at the center of bias tolerance. By the same token, the presence of positive skew, which indicates a lowered bias tolerance, usually calls for a readjustment of the receiver to reduce the fortuitous effect on the space-to-mark transitions. As explained above, removing the skew by introducing a fortuitous effect on the mark-to-space transitions will not, of course, improve the bias tolerance.

It is the present practice in the field to specify a minimum bias tolerance about 5 per cent greater than the minimum permissible "end distortion" tolerance, the orientation being adjusted to the center of bias tolerance for both measurements.

SOME CAUSES OF INTERNAL DISTORTION

Up to this point internal distortion has been considered without regard to its probable causes. The more obvious causes will be found to be analogous to those which produce equivalent distortions in telegraph transmission circuits.

Bias will result when an element (whether electrical, mechanical, or electronic) of a receiver possesses dissymmetry toward marking or spacing. For example, a mechanical element may travel more slowly from spacing to marking than from marking to spacing and thus cause spacing bias, or its range of travel may be divided unequally into marking and spacing portions, thus producing an equivalent effect.

Characteristic distortion will result when an element (whether electrical or mechanical) of a receiver fails to attain a steady state before being acted upon by a succeeding transition, or otherwise depends, in its action, upon the previous history of the signal train. An example of characteristic distortion is found in the 20-milliampere holding magnet selector when it is equipped with a resistive shunt. In this type of selector the armature is actuated by a cam, which presents it to the pole-face at about the middle of each pulse, and then disengages it. The armature is then free to release or remain operated, according as the received pulse is spacing or marking. The shunt that is normally used presents so low an impedance to the magnet winding that the motional impedance effect which is produced by the sudden mechanical presentation of the armature to the pole-faces causes a sizeable reduction in the magnet current. In the case of a short marking pulse, the current fails to attain steady state before the next mark-to-space transition occurs. The magnet therefore releases sooner than it does at the end of a long marking pulse, during which the current has had time to attain steady state. It will be seen that this is really a characteristic distortion effect, since it is due to a failure to reach steady state and depends upon the previous history

of the signal train. However, when miscellaneous signals are being received the effect appears similar to a fortuitous distortion occurring on mark-to-space selective transitions, and hence it is usually thought of as negative skew.

Fortuitous distortion will result when an element is irregular in its action, and if such action is more irregular on one type of transition than on the other, the result will appear as skew. For example, irregular action of the receiving clutch affects the selector alike in regard to all selective transitions, and appears as internal fortuitous distortion. Another source of internal fortuitous distortion is the period of indecision that occurs during the passage of a selective element past a locking member, at which time the choice between marking and spacing is largely fortuitous.

A common cause of skew in teletypewriters may occur in the following manner: If the armature stops are so adjusted that, for example, the armature travel is greater on the marking side than on the spacing side of the armature lock, positive internal bias results. If, now, this bias is compensated for by so adjusting the armature air-gap and retractive spring tension as to cause the receiving magnet to operate in a negatively biased manner (rather than by correcting the improper armature travel), the armature will be forced to operate in a region of the operating wave that is more sloping than the region in which it releases. Hence, it will operate more irregularly than it releases, and thus will be affected by positive skew.

SELECTOR ACTION

Over and above the sources of internal distortion which are analogous in effect to sources of distortion encountered in telegraph transmission circuits, there is another whose action in causing internal distortion is not so obvious as those just described. This source of internal distortion may be termed "*selector action*," and it depends upon the relation between the operating time of a selector element and the period of time allowed for said element to act. For the purpose of explaining the effect of time relations within the selector on internal distortion, selector mechanisms may be classified as of three basic types: *M*, *S*, and *P*.

In a mechanism of type *M* each selector is initially in the spacing condition and either remains spacing or operates to marking when subjected to the action of the corresponding received signal element. When it attains the marking condition it becomes locked for the duration of the character. Early types of start-stop printers having an individual selector magnet for each pulse of the code and employing a separate receiving distributor,² are illustrative of type *M*.

In a device of type *S* each selector is initially in the marking condition and either remains marking or operates to spacing when subjected to the action of the corresponding received signal element. When it attains the

spacing condition it becomes locked and cannot again operate to marking during that character. The Siemens-Halske five-selector teleprinter⁵ is an example of this type.

In a mechanism of type *P*, the selector may be in either the marking or spacing condition initially, according to the type of the previous signal element to which it has responded. When subjected to the action of a received pulse the selector may go in either direction, and it remains responsive to the action of the signal during the entire selecting interval. The No. 14 and No. 15 teletypewriters² (not equipped with holding magnet selector) of the Teletype Corporation are examples of type *P*.

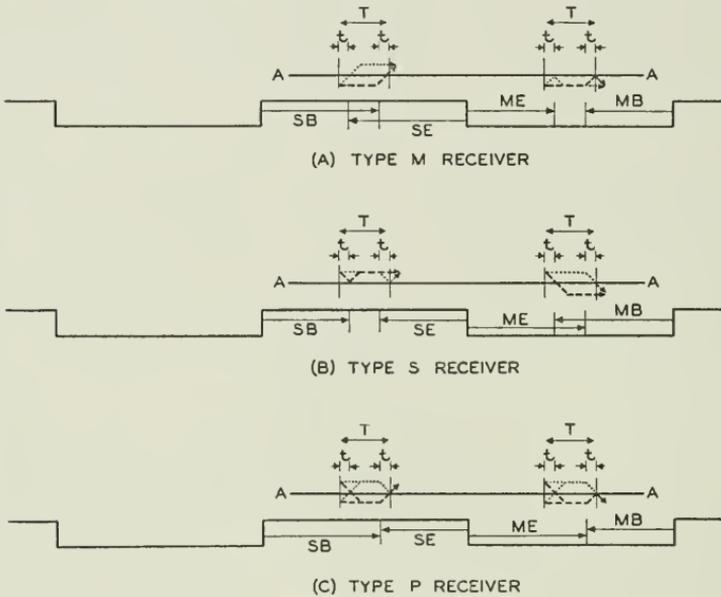


Fig. 10—Effect of selector action on internal distortion.

Figure 10 (A) illustrates the action of a type *M* selector. A portion of a teletypewriter character is shown, consisting of the spacing start pulse, a marking first selective pulse and a spacing second selective pulse. The undistorted signal is shown in solid lines. The maximum amounts of marking and spacing bias that the receiver will tolerate are shown by dashed lines and are designated *MB* and *SB*. The limiting amounts of marking and spacing end displacement are shown by dotted lines and are designated *ME* and *SE*. Above the signal train is shown a schematic representation of the action of the selective system. The periods of time *T* are those during which the selector is subject to the action of the received signal, and *t* is the time that the selector must be subjected to the operative force in order that it

operate. The line *A-A* indicates the boundary between the marking and spacing positions of the selectors. In this type of receiver, as mentioned previously, when the selector crosses to the marking or upper side of line *A-A* it becomes locked and cannot again go to spacing even though the signal should subsequently become spacing during the selective period T .

It will be noted that the limits of end displacement tolerance occur at time t after the beginning of the selective period. This instant is sometimes called the "instant of decision for end displacement." On the other hand, the limiting tolerances to bias are determined at a time t before the end of the selective period, sometimes known as the "instant of decision for bias." If the selective periods were advanced relative to the start transition by lowering the orientation until the bias tolerances were equal, the instants of decision for bias would correspond with the center of bias tolerance. If, then, the selective periods were delayed, by raising the orientation, by an amount $T - 2t$, the instants of decision for end displacement would correspond with the center of end displacement tolerance. Since the difference between the center of end displacement tolerance and the center of bias tolerance is equal to the internal bias of the receiver, it will be obvious that the internal bias is also equal to the difference between the instant of decision for bias and the instant of decision for end displacement. In this type of receiver the internal bias is $T - 2t$, and will be positive, zero, or negative according as $2t$ is less than, equal to, or greater than T .

Figure 10 (B) shows the action of a type *S* selector. Here the instant of decision for bias occurs at time t before the end of the selective period and that for end displacement at time t after the beginning of the selective period. Hence, the internal bias is equal to $2t - T$.

The action of a type *P* receiver is illustrated in Fig. 10 (C). It is assumed in this figure that the selector operates toward marking at the same rate as toward spacing, since the effect of unequal rates of operation has been described previously. In a selector of this type, both instants of decision occur at time t before the end of the selective period and hence the internal bias is not dependent upon the relation between T and t . If, however, t is so long that the selector cannot pass from one extreme of travel to the other, attain a steady state, and return to the center position within time T , a sort of characteristic distortion occurs, in which the instant of decision depends upon whether the selector began the selective period in the same or the opposite condition from that finally selected. In measurements of miscellaneous signals this appears similar to a fortuitous effect, since it decreases all tolerances equally. Hence it is usually considered as internal fortuitous distortion.

Receivers equipped with holding magnet selectors are of Type *S*, since the armature may be released, but not operated, by the magnet. In this

type of mechanism, the armature generally drives a subsidiary selective member, and the time T extends from the instant at which the armature is disengaged by its operating cam until the instant when the subsidiary selector becomes locked. As this period is often long in relation to the magnet releasing time t_1 and the subsidiary selector operating time t_2 , holding magnet selectors are often subject to negative internal bias. In those mechanisms in which the subsidiary selector is flexibly coupled to the magnet armature, the former's operation is of type P . It, therefore, may be subject to the characteristic distortion effect noted in the description of type P operation, except that the effect, when it occurs in this type of mechanism, affects only the instant of decision for end displacement and hence resembles negative skew rather than internal fortuitous distortion.

An interesting, but somewhat unusual, effect occurs in any receiver, of whatever type, in which the lengths of selective period or selector operate time, or both, differ for the various selective pulses, or in which the spacing of the selective periods is improper. In a case of this sort, the receiver exhibits an internal bias equal to the difference between the average instant of decision for bias and the average instant of decision for end displacement, an internal fortuitous distortion equal to the variation of the instant of decision having the smaller variation, and a skew equal to the difference between the variations of the instant of decision for bias and the instant of decision for end displacement.

CONCLUSIONS

A working knowledge of the effect of telegraph distortion on the margins of operation of start-stop receivers is essential in dealing with a plant in which the use of teletypewriters, regenerative repeaters and start-stop distortion measuring sets is as widespread as it is in the Bell System. When a major portion of the communication system operates on a start-stop basis, it is desirable that transmission measurements be made on the same basis.

The knowledge of this subject that has been gained in recent years has made possible many improvements in technique both in the field and in the laboratory, and these have led to corresponding improvements in the mechanisms used in telegraph service. The analysis of new start-stop devices may now be carried out efficiently and accurately, and this often permits the formulation of suggestions leading to improved operation of the devices.

The general level of service excellence has been raised by the setting up of criteria for the distortion tolerances of station teletypewriters, regenerative repeaters and other start-stop devices used in service, including those provided for switching. The sources of distorted test signals that are now available are useful not only in measuring the tolerances of service receivers,

but also in determining the characteristics, and hence the accuracy, of start-stop distortion measuring sets and monitoring teletypewriters.

Finally, there has resulted an improved ability to analyze and predict the performance of transmission links from the results of distortion measurements made on a start-stop basis.

REFERENCES

1. "Measurement of Telegraph Transmission," H. Nyquist, R. B. Shanck, S. I. Cory, *Jour., A. I. E. E.*, March 1927, p. 231.
2. "Fundamentals of Teletypewriters Used in the Bell System," E. F. Watson, *Bell Sys. Tech. Jour.*, October 1938, p. 620.
3. "A Transmission System for Teletypewriter Exchange Service," R. E. Pierce and E. W. Bemis, *Bell Sys. Tech. Jour.*, October 1936, p. 529.
4. "Recent Developments in the Measurement of Telegraph Transmission," R. B. Shanck, F. A. Cowan, S. I. Cory, *Bell Sys. Tech. Jour.*, January 1939, p. 143.
5. "Der Spielraum des Siemens-Springschreibers," M. J. de Vries, *Telegraphen-und Fernsprech-Technik*, January 1934, p. 7.

CHAPTER XIII*

The Mounting and Fabrication of Plated Quartz Crystal Units

By R. M. C. GREENIDGE

13.1 INTRODUCTION

THIS paper is one of a series on piezoelectric quartz plates and deals primarily with the methods employed in mounting crystal plates operating up to approximately one megacycle for practical utilization in communication equipment. The theoretical aspects of mounting crystals have been covered in Chapter VII. The discussion is confined to plates¹ having definite nodal lines or points, such as $+5^\circ$ and $-18^\circ 25'$ X cuts, GT, CT, DT, MT and NT cuts. The mounting of high-frequency crystal plates such as AT and BT cuts, which vibrate in thickness shear modes, is not included. It should also be noted that the subject matter is treated descriptively and that no attempt is made to go into the more intricate details of design or to give performance characteristics. These matters will be dealt with fully in a later paper. The designs and methods outlined are up to date for each type of unit, the results of many years of development on the part of Bell System engineers to evolve practical designs for commercial manufacture and use. Expanding on the contributions of the early investigators mentioned by W. P. Mason in Chapter I,¹ these engineers had, in the ten years prior to 1939, worked out practical designs and developed suitable tools and processes for wide commercialization in telephone applications. In the last five years, under the impetus of war, further improvements have been made in the design and manufacture of crystal units, particularly those for use by the Armed Forces.

The term "Crystal Unit", originally adopted by the Bell System to designate the complete assembly of a crystal plate in its mounting and case, has now been standardized quite generally in the art, replacing a variety of names by which these devices were formerly called. The basic design features of a crystal unit involve the use of:

1. Electrodes, on or near to the crystal surfaces for impressing voltage across the plate,

* Chapters IX, X, XI and XII, which will be included in a forthcoming volume are omitted from the *Technical Journal* because they deal largely with details of manufacturing operations.

¹"Quartz Crystal Applications", W. P. Mason, *B.S.T.J.*, Vol. XXII, Page 191, July 1943.

2. Supports for holding the crystal plate in its mount, and
3. A sealed outer case having the necessary terminals, and provisions for incorporating the unit electrically and mechanically into the apparatus.

Two distinct types of crystal units have been evolved, one embodying the use of pressure pins or anvils for supporting and holding the crystal plate and the other involving the suspension of the crystal plate by means of fine wires.² These designs are known, respectively, as the Pressure Type and Wire Supported Type, and will be discussed later under these headings. However, there are several details of fabrication common to both types which can best be discussed at this point.

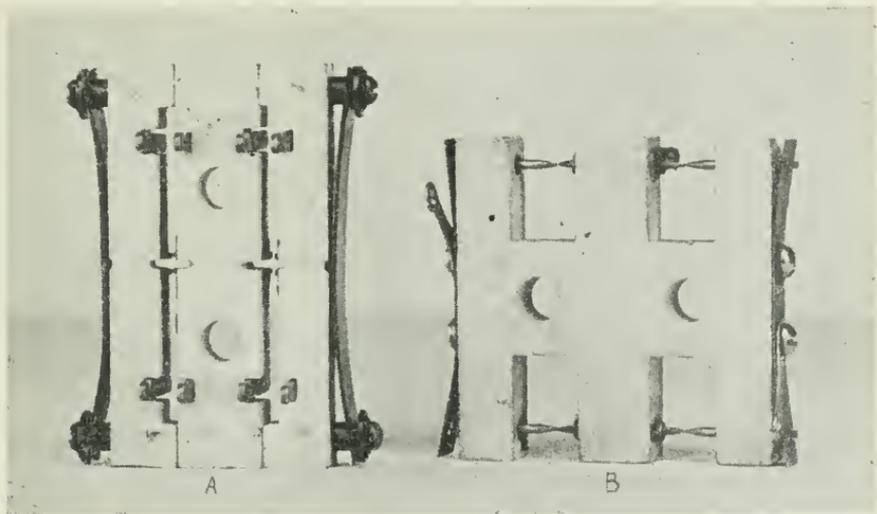


Fig. 13.1—Pressure-type holders.

Irrespective of the type of mounting, realization of the desired performance in a crystal unit depends to a considerable extent on the processing of the quartz plate itself. Previous articles^{1, 3} have brought out the significance of such factors as the precision of angular orientation and linear dimensions on the fundamental characteristics of the plate. The plate must also be virtually free of impurities or imperfections.⁴ In the preparation of a quartz plate it must be lapped using increasingly finer abrasive materials until the final dimensions are reached. Depending upon the type of crystal unit, No. 400, No. 600 carborundum or finer abrasives are now

² A. W. Ziegler, Patent 2,275,122, March 3, 1942.

³ "The Use of X-Rays for Determining the Orientation of Quartz Crystals", W. L. Bond and E. J. Armstrong, *B.S.T.J.*, Vol. XXII, Oct. 1943.

⁴ "Raw Quartz, Its Imperfections and Inspection," G. W. Willard, *B.S.T.J.*, Vol. XXII, Oct. 1943.

employed for the final stage grinding. Following this, the plate is thoroughly cleaned by acid treatments or by the use of solvents and detergents followed by copious washing. It is then etched in commercial hydrofluoric acid to remove all loose particles of quartz that might have remained on the surfaces or in the crevices after cleaning. Etching also smooths off the roughness of the ground surfaces. The effect of this treatment reduces energy dissipation in the plate itself and increases by many times the efficiency of the crystal units. Etching also improves the stability of performance of the crystal unit. Standard designs of crystal units require etching of the plates, uniformly on all surfaces, for a period of thirty to forty minutes. For units of highest precision and efficiency longer etching periods are employed.

The electrodes employed with types of crystal units being described consist of metallic coatings, generally aluminum, silver, or gold deposited over the major surfaces of the crystal plate. These coatings are applied by the evaporation process which results in an extremely thin and uniform coating of metal having excellent adherence to the quartz.

With reference to the mechanical supporting members for the plate, it has been brought out that such supports should be confined as closely as possible to the nodal points or nodal lines where the motion for all practical purposes is zero. It is common practice for the supporting members to be made of metal so that they will serve also as a means of making electrical connections to the electrode coatings on the surfaces of the plates.

13.2 PRESSURE TYPE CRYSTAL UNITS

This type of crystal unit was initially developed for use in telephone filters. Up to about five years ago it was employed in virtually all commercial designs of filters. Depending upon the mode of vibration and size of the crystal plate the design of the mounting varies. However the principles employed for clamping are essentially the same in all cases. Where small longitudinal or face shear plates are involved one pair of pressure pins is used unless two are required for electrical reasons as explained later. For medium size plates of the same type or for face flexure plates, two pairs of pins are employed. In the case of large low-frequency longitudinal plates double anvils are used instead of pins in order to obtain firmer clamping of the plate to prevent translation or rotational movement which might cause wear in the electrode surface at the pressure point with resultant variations in frequency and resistance. The blocks are usually composed of molded steatite and the springs for exerting the necessary pressure are of phosphor-bronze.

Figure 13.1 (B) shows a pressure mounting for holding four crystals which have single coatings on each of their major surfaces. The main require-

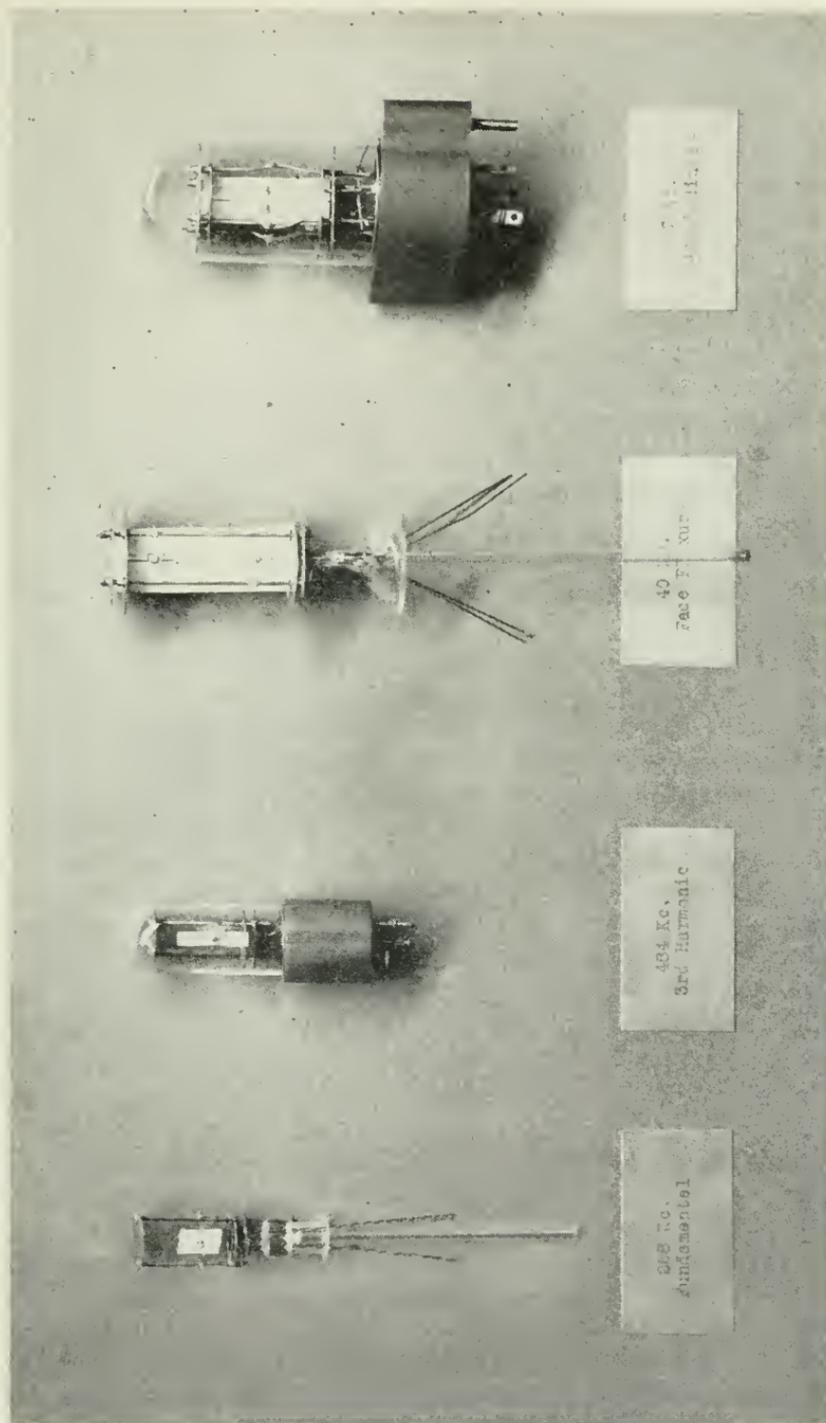


Fig. 13.2—Wire-supported crystal units.

ments which must be met for such a mounting are small areas, accurate alignment, and adequate pressure to hold the plate in place. Some designers make slight indentations in the quartz at the point of contact to improve the mechanical stability of the plate. For crystal plates of the order of one-half inch square or smaller, points having an area of about 10 mils in diameter are employed and the pressures used for holding the plate range from one to two pounds. For larger plates correspondingly larger areas of points and increased pressures are employed. The accuracy of alignment required for ten mil points is of the order of two or three mils. This is obtained in the mountings shown in Figure 13.1 by using concentric sleeves for holding the points which are brought into alignment by means of a straight rod and then cemented in place. One of the points is fixed while the other point slides in its sleeve and the pressure required is obtained by the spring which presses on the outer end of the sliding point. In balanced filter structures it is desirable to use crystal plates with the coating on each side divided into two equal areas. This reduces to one-half the number of plates that would otherwise be required. In mounting plates with divided coating, it is necessary to provide a mounting which makes double contact on each side of the plate. Figure 13.1 (A) shows a pressure type mounting which accomplishes this. This is the mounting which has been used for several years in holding the plates used for the 75-type crystal channel filters⁵ for the standard terminal common to all broad-band telephone systems. The crystal is mounted in the holder in such a way that the two pairs of points clamp the crystal along the nodal line. The rectangular dimensions of the points used for this type mounting for crystals operating in the frequency range from 60 kc up to 120 kc are about 35 mils long in the direction of the nodal line and from 10 to 15 mils wide. A very important requirement for such a mounting is that the flat area of the points on each side of the plate fall in the same plane. This is accomplished by a precise milling operation after the points are assembled in the mounting. The pressure applied to the pair of points is furnished by the flat spring shown and is equalized by the action of the roller centrally located under the springs. The pressure employed is of the order of four to five pounds for each pair of points.

The most commonly used coating for crystal plates held in pressure-type mountings is aluminum.⁶ Aluminum has been found to be most satisfactory for this type of unit because its hard surface is more resistant to wear at the points of clamping than other metals such as silver, or gold.

Except for a few designs, which are mounted in sealed metal or glass

⁵ "Crystal Channel Filters for Carrier Cable Systems," C. E. Lane, *B.S.T.J.*, Vol. XVII, Page 125.

⁶ The details of processing aluminum-coated crystals are similar to those described for silver-coated crystals in paragraphs 13.42 and 13.43.

containers, pressure-type crystal units are not sealed in individual containers. However, the entire filter in which they are employed is dried and sealed off after filling with dry air.

In pressure-type units of this type, variations in frequency of the order of .01% may be expected if crystals are transferred from one mounting to another or relocated in the same mounting. Consequently, if high-frequency precision is desired, it is necessary to make the final frequency adjustment with the crystal located in its final position. Due to the inherent difficulties of adjustment, coupled with the close manufacturing tolerances on parts, and precision adjustments necessary for the holders during assembly, these designs have been virtually discarded in favor of wire supported designs. However, more recent developments by J. F. Barry on pressure-mounted-type units have brought forth some new ideas which might prove in for wider future application.

13.3 WIRE SUPPORTED CRYSTAL UNITS

This type of mounting is being used extensively on crystal applications and has superseded the earlier pressure-type units. Various designs of this type of crystal are shown on Fig. 13.2. The wire-mounted crystal possesses the definite advantage in that after the supporting wires are attached to the plate, they remain fixed in position throughout the subsequent manufacturing process thus facilitating adjustment of the frequency or the frequency-temperature characteristic. Moreover, the supporting wires can be formed so as to provide a spring mounting for the crystal plate which protects it from any shocks or vibration it may encounter in shipment or use. In the wire-supported design the suspension wire is also employed as a means of connecting the electrical circuit to the electrode plating on the crystal. By virtue of the solder bond between the wire and the electrode, this type of unit is free from the possibility of instability in frequency performance due to slight changes in position and variations in contact resistance prevalent in pressure type designs, and for this reason the stability and frequency precision of wire supported crystals is superior.

From a manufacturing standpoint the wire-supported unit involves a greater number of processing operations than the pressure-type unit, but the adjustment operations are considerably easier. Moreover, it is possible to realize greater precision of frequency adjustment by a factor of at least two or even three. The crystals are also more uniform in their effective resistance. Since the mount or cage in which the crystal is suspended is comparatively inexpensive, there should in general be little difference in the manufacturing costs of the two types. Consequently, in the wire-supported crystal units a very appreciable improvement should be gained in performance without increasing the cost of the unit.

13.4 FABRICATION OF WIRE SUPPORTED UNIT

*13.41 Silver Spotting**13.411 Application of Silver Paste*

Starting with the crystal plate, the first step in manufacture is to apply silver spots to the surfaces of the plate. These spots serve as footings to which the supporting wires are ultimately soldered or sweated. They are placed on the nodal points or along the nodal lines of the plate in order to detract as little as possible from the intrinsic characteristics of the plate itself. The areas of the silver spots cover the range from about 40 to 90 mils in diameter depending upon the amount of solder to be used in attaching the wire to the plate. Before spotting the plates it is essential that they be free from any contamination such as grease or organic material that might affect the fusion of the silver spots into the surface of the quartz. One of the best methods to ensure cleanliness is to boil the plates in aqua regia, followed by copious rinsing in water. Detergents such as sodium meta-silicate are also employed followed by a rinsing. The plates may finally be boiled in distilled water and carefully dried. Throughout the subsequent processes the plates should be handled with clean tweezers or gloved fingers and kept away from any source of contamination. Prefiring of the plates at 950°F prior to spotting has also been used as a positive way of ensuring freedom from any contamination that would affect the fusion of the spots, but this process is not necessary if the first mentioned process is properly controlled.

In spotting, small quantities of a prepared silver paste are placed on the areas of the plate to which the wires will ultimately be attached. The paste consists of a compound of finely divided silver and low melting point glass (lead borate) thoroughly mixed with a suitable vehicle to facilitate application. For spotting purposes it has been found that a paste having a specific gravity of between 2.3 and 2.6 gives best results. In use, the materials must be constantly agitated or stirred in order to prevent the solid ingredients from settling out. This is important, for, unless the concentration of silver is maintained around 90 to 95 per cent of the solid matter, it will not be possible to obtain good wetting of the solder in making the wire attachment.

The placement of the semi-liquid material on the plate is accomplished by means of a small stylus, the crystal plate being held in a clamp or vise and the stylus guided so as to place the material at the exact location on the plate as desired. A typical tool for doing this work is shown in Figure 13.3. The point of the stylus should have a slightly rounded end. With the rounded point the tendency of the paste to spread out is minimized and consequently the diameter of the spot is substantially the same as that of

the stylus. The rounded stylus also results in a more uniform distribution of the material. The material is applied to the end of the stylus by spreading a small amount of the paste on a glass plate from which it is transferred to the stylus and then deposited on the crystal. The material on the glass plate should be wiped off and replaced quite often due to settling and drying out of the mixture, in order to insure uniformly good spots. Generally speaking, anywhere from two to six spots at the most should be possible from one loading of the transfer plate depending upon the speed of the operator.

Regarding the character of the crystal surface, aside from cleanliness, and its effect on the ultimate strength of adhesion of the silver spot, experience

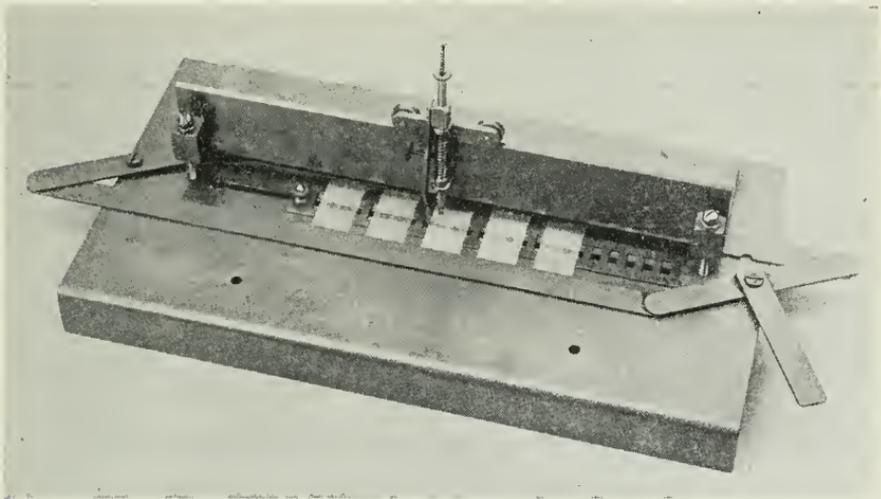


Fig. 13.3—Tool for applying silver spots.

so far with all the various cuts of plates does not indicate that this is a factor. Spots have been found to adhere to polished surfaces as well as ground and etched surfaces with about the same degree of strength.

13.412 *Firing of Silver Spots*

Following the application of the paste to the plates it is desirable to pre-dry the spots in order to remove the low volatile constituents of the vehicle prior to firing at high temperature. This may be done in a ventilated oven or over a hot plate at approximately 300°F for about 15 minutes. The plates are then placed in a furnace and heated up to between 975 to 1010°F and held at that temperature for a sufficient length of time to obtain good fusion of the spot to the plate. Ordinarily this reaction takes only a few minutes after the plate has reached the proper temperature. After firing, the plates are allowed to cool in the furnace to the point where they can be

removed without danger of the crystal cracking due to cold shock. It is essential to control the temperature of the furnace so that the temperature of the crystal plates does not reach 1063°F , otherwise the crystals may become electrically twinned and consequently useless. In order to avoid shattering of quartz plates due to thermal shocks while heating up and cooling, fairly long cycles have heretofore been specified. However, more recent experience has shown that much shorter cycles can be employed especially where small crystals are involved. Moreover, there are indications that the faster heating, particularly during the last two or three hundred degrees temperature rise, results in better spots. During the firing

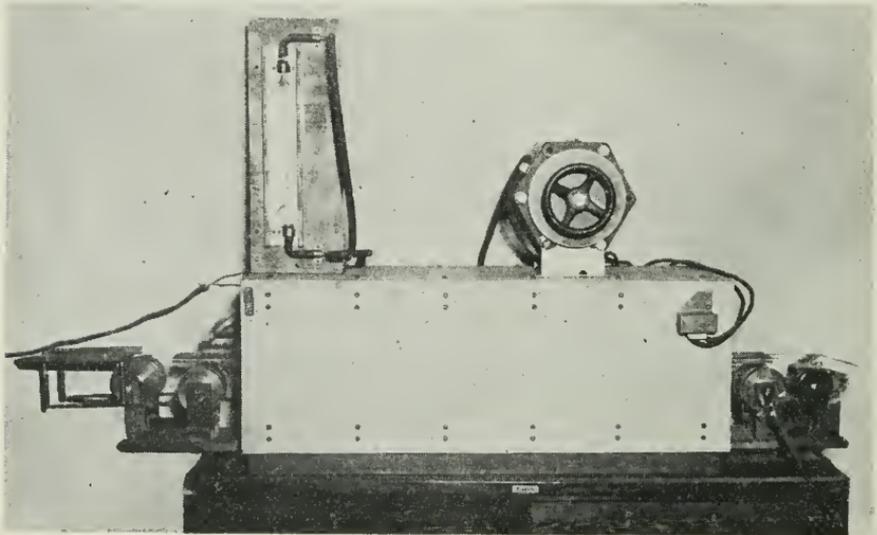


Fig. 13.4—Continuous-belt furnace for firing silver spots.

operation the crystal plates may be placed on nichrome wire mesh trays or Pyrex dishes provided ample provision is made for air to circulate around the spots. This precaution is essential with the types of pastes employed as the baking reaction must take place in the presence of oxygen so that the lead borate will not be reduced to lead, leaving only a partially bonded mixture of silver and lead on the plate. This type of spot when encountered usually has a dull appearance after burnishing as compared with the bright surface obtained with a good silver spot, and is quite difficult to wet with solder. For firing silver spots, ventilated continuous-belt-type furnaces with open ports are being used with very satisfactory results. Figure 13.4 shows a furnace of this type developed by C. J. Christensen for this purpose.

After the firing operation, the spots should be examined to ensure that they are satisfactory. Besides a visual inspection, it is desirable that a

small percentage of the plates be used as a control sample to which mounting wires are attached and pull tested. Satisfactory spots should withstand for a few seconds a force of at least two pounds. The average pull-off strength of commercial attachments using 6-mil hooked or headed wires is between three and four pounds. Unsatisfactory plates can be reclaimed at this stage by stripping off the spots by means of aqua regia and ammonium hydroxide, and reprocessing in the manner described.

13.42 Silver Plating

At this point of the process the surfaces of the plates are coated with silver electrodes by the evaporation process previously mentioned. Four milligrams per square inch of silver is the weight of coating generally employed, which amounts to a thickness of .024 mil. Except for harmonic and other special types of crystal units, these coatings are required only on the major surfaces of the plates. However, during the evaporation process, the silver is deposited to some degree on the minor surfaces or edges as well, and it is necessary to remove it. This process called "edge cleaning", is done by lapping the edges of the crystal on a flat plate covered with a mixture of pumice or a finely divided abrasive such as No. 600 carborundum and water or kerosene in the form of a paste. Rubbing the edge of the plate lightly over very fine abrasive cloth is also satisfactory. The pumice is preferable, however, since while it readily removes the silver, it is much softer than quartz and consequently does not remove any material from the plate. The use of harder abrasives has a tendency to chip the edge of the quartz unless the operation is performed very carefully. After the edge cleaning is completed the plates are washed, dried and inspected by testing for insulation resistance at 500 volts d.c. to make certain that no conducting material remains between the silver coatings on the major surfaces.

13.43 Division of Coating

For circuit reasons all but a few types of crystal units require a balanced pair of electrodes on each side of the plate. Division of coatings along the longitudinal axis is essential on flexure mode crystals in order to make the plate vibrate in flexure. Typical divisions of coating can be noted on the crystals shown in Figure 13.2. In the case of the flexural crystal the dividing line is carried around the wire attachments in such a manner that each of the divided surfaces is connected to one of the wires. One method for dividing coating involves the use of a low voltage (two to three volts) impressed between the coating to be divided and a stylus.⁷ When the fine point of the stylus is brought in contact with the silver plating and moved along the desired line of division, the silver is burned away, leaving a small

⁷ W. L. Bond, Pat. #2,248,057.

gap in the plating between 8 and 18 mils wide depending upon the point of the stylus. Following the burning operation, the plate is immersed in a photographic hypo solution to remove all traces of the burned residue after

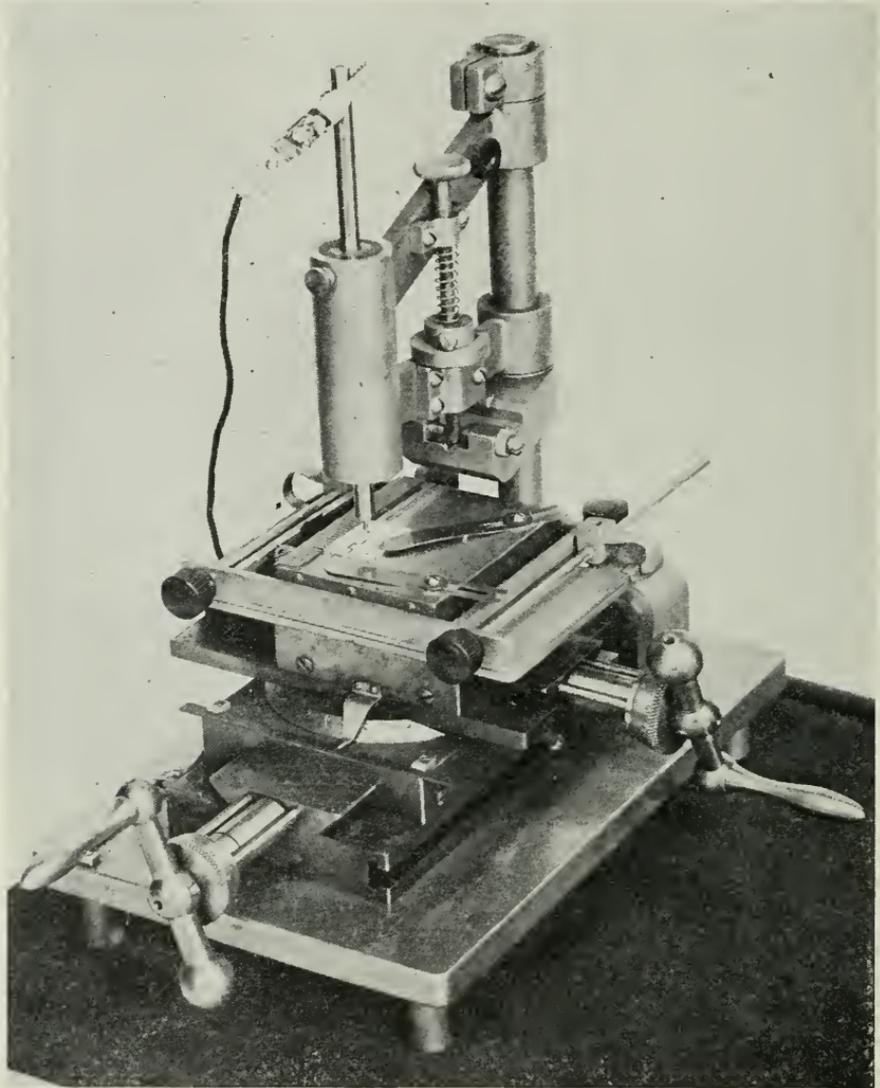


Fig. 13.5—Tool for dividing plating (electric stylus).

which it is carefully washed and rinsed in water and dried. In this process it is important that the plates are not kept in the hypo solution for longer than two or three minutes, otherwise discoloration of the silver coating will

result. The gap is then tested for presence of metallic particles by gradually impressing voltages up to 1000 volts a-c across it. If no flashover occurs the division is satisfactory. If flashover occurs the voltage is maintained until the slivers are burned out. The hypo and burning treatments are repeated until a good division is obtained. Figure 13.5 shows an electric dividing tool developed for this purpose. In using this method it has been found that the arc at the point of the stylus may cause twinning of the quartz to a minor extent along the dividing line. This effect is usually insignificant although it may be objectionable especially where precise values of crystal inductance or frequency-temperature performance are required. Where more complicated divisions are necessary, as in the case of face flexure and harmonic plates, the electric stylus method is employed, although methods and tools for performing this operation by other means to avoid twinning are being developed.

13.44 Attachment of Wire Supporting Leads

Phosphor-bronze wire is employed in wire supported crystal units primarily because of its high tensile strength, and excellent fatigue resistance characteristics. Five- and six-mil diameter wires are the most widely used sizes, depending upon the mass of the crystal plate, the desired electrical performance, and the severity of treatment it is likely to encounter in use. To facilitate soldering the wires to the spot on the crystal plate and to the crystal support system the phosphor-bronze wire is given a heavy electro-tinned finish. 59.5-34.5 per cent tin-lead eutectic solder saturated with approximately 6 per cent silver at 570°F is employed for attaching these fine wires to the silver spots of the crystals. This solder solidifies at approximately 360°F with practically no mushy stage. The reason for saturating the solder with silver is to discourage migration of the silver in the spot to the solder during the soldering operation. Even with this solder it is advisable to limit the time for heating of the joint to a minimum.

One method of attaching the wires to the crystal plate is by means of a special machine developed for the purpose. Such a machine is illustrated in the photograph on Fig. 13.6. The wire is fed from a spool through the head in the movable arm. The head contains a wire guide having a hole only slightly larger than the diameter of the wire and a small vise for firmly clamping the wire. The crystal plate is clamped in the vise on the hot plate which is thermostatically controlled at approximately 240°F. The position of the arm carrying the wire is lined up with respect to the crystal plate by means of guides so that the wire will be placed exactly on the nodal point or line of the crystal. In making the attachment, with everything lined up, the wire is fed through the guide until it touches the spot on the plate and the vise closed. Since the curvature of the wire can never be entirely

eliminated the distance between the tip of the guide and the crystal plate is kept as small as possible. A small disc of solder is then punched in the press at the left, the little disc remaining in a round slot whose position is also lined up with respect to the arm carrying the wire. The movable arm

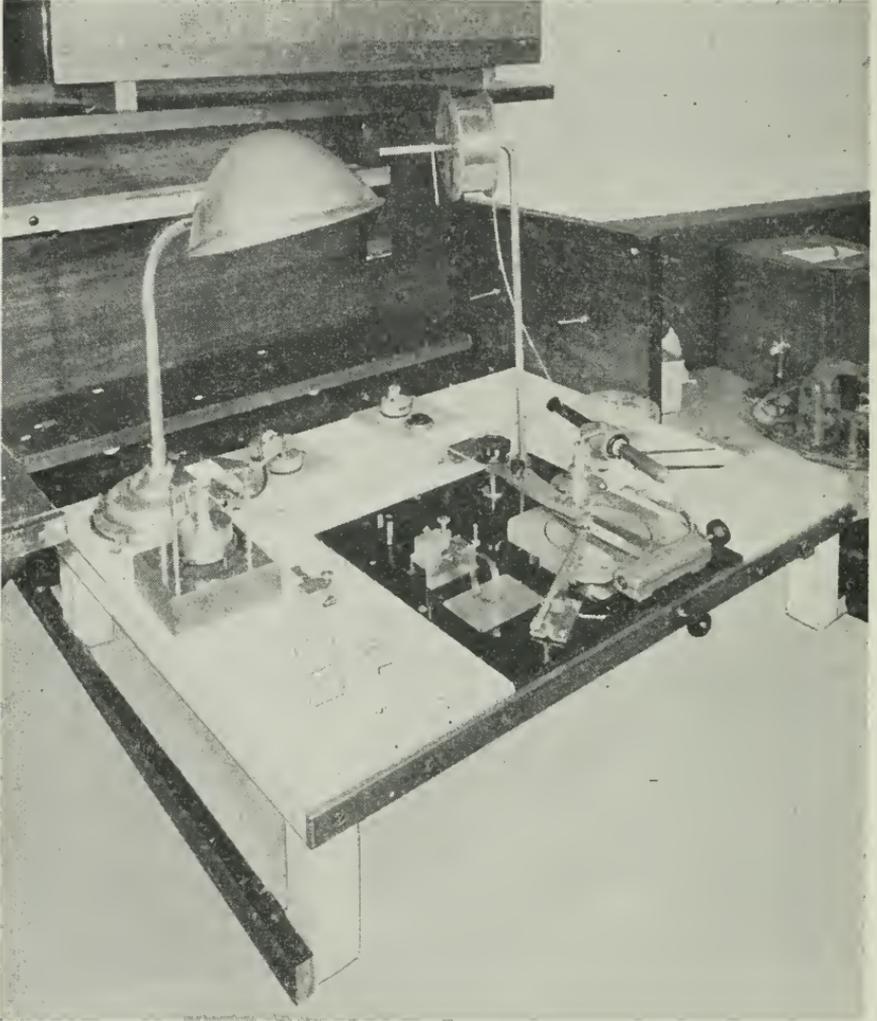


Fig. 13.6—Wire-soldering machine for straight or hooked wires.

is then rotated until the wire is directly over the solder disc at which point it falls into a guide and comes down and spears the disc. The arm is then lifted, picking up the solder at the end of the wire. The solder and the end of the wire is then wetted with rosin-alcohol flux and the arm rotated

until it falls into its original position over the crystal plate. The solder is then fused to the wire and plate by means of a special aluminum tipped soldering iron as shown in the illustration or by a controlled hot air blast focused on the joint to melt the solder. In this operation a fillet or conical button is formed around the wire attaching it to the silver spot. To promote good wetting of the solder, the spot should be clean and well burnished. Rubbing the spot on a hard polished metal surface or burnishing with a blunt pointed tool of agate, are the best methods found so far.

The hot air blast has now replaced the iron entirely in commercial use. It consists of a tube through which air at about one inch water pressure is passed over a hot filament and through a nozzle directed at the solder. The head of the filament is adjusted so that the temperature of the blast is just hot enough to melt the solder and complete the attachment in 10 to 15 seconds time. In using this method with large plates care must be taken to insure that the temperature of the crystal has reached that of the hot plate and that the blast is brought up to the plate slowly, for otherwise the heat shock of the localized blast may cause the crystal to crack. For very small plates the use of a hot plate may be dispensed with if the hot blast is brought up slowly enough to preheat the crystal plate. Other means of melting the solder such as a hot radiant wire or ribbon or the use of a minute flame have been considered, but so far no extensive trials of these methods have been made. The advantage of the hot blast over the other methods mentioned is that it can be better controlled since little is left to the judgment of the operator. If an iron is used it must actually be touched to the solder with the possibility of displacing the position of the wire. Moreover, as already mentioned, the iron must be equipped with a special aluminum tip to prevent removal of solder from the joint on withdrawal of the iron. Considerable maintenance is required to keep such irons in satisfactory operating condition.

13.45 Type of Wire Attachments

The type of attachment described above wherein the part of the wire embedded in the solder cone is straight was used in the first designs of wire-supported crystals. However, it was found that with such attachments vibration of the crystal plate caused breakage of the bond between the solder cone and the wire with resultant failure of the attachments, especially in large plates. Because of this the use of straight wires is recommended only for small size plates. In order to eliminate the above difficulty a little hook has been placed at the end of the wire embedded in the solder in order to obtain a better anchorage. The hook is formed in the wire by means of a special tool affixed to the soldering machine. The basic methods described for straight wires are otherwise used for this type of attachment. Instead

of spearing the little solder discs as with the straight wire, the solder is punched in the shape of a horseshoe and squeezed in place on the hook or positioned by tool with the hooked wire in place on the spot. Hooked wire attachments will withstand pulls of the order of three to four pounds before pulling off. Under severe vibration hooked-wire attachments have the same tendencies as straight wires towards breaking away of the wires from the top of the solder cone forming a small crater in the latter. However, the crater does not progress deeply enough into the solder cone to impair their strength or cause failure under ordinary conditions.

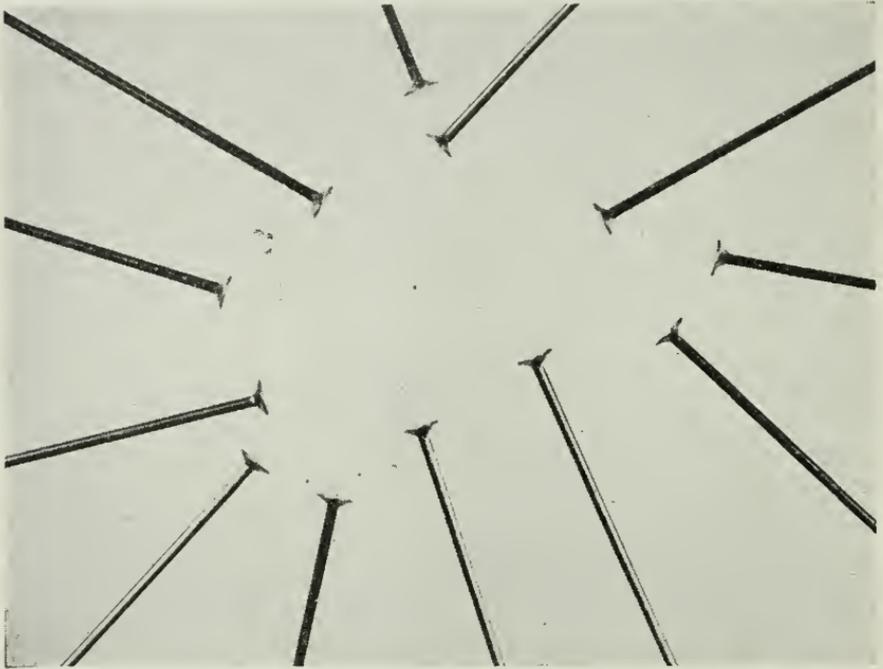


Fig. 13.7—Examples of headed phosphor-bronze wire.

The most recent development for wire supports involves the use of headed phosphor-bronze wires as worked out by A. W. Ziegler. In this procedure individual wire lengths are cut and one end upset in a cold heading tool which provides a little cone-shaped head with a base of about 22 mils as shown in Figure 13.7 for 6-mil diameter wires. The head is carefully pre-tinned, leaving a small globule of solder at the end. Depending upon the size of the crystal plate, globules of 1000, 3000 or 7000 cubic mils of solder are used. The attachments are made in a wire soldering tool which attaches the wires to both sides of the plate simultaneously. This tool is illustrated in Figure 13.8. The prepared wires are fed into positioning guides, which

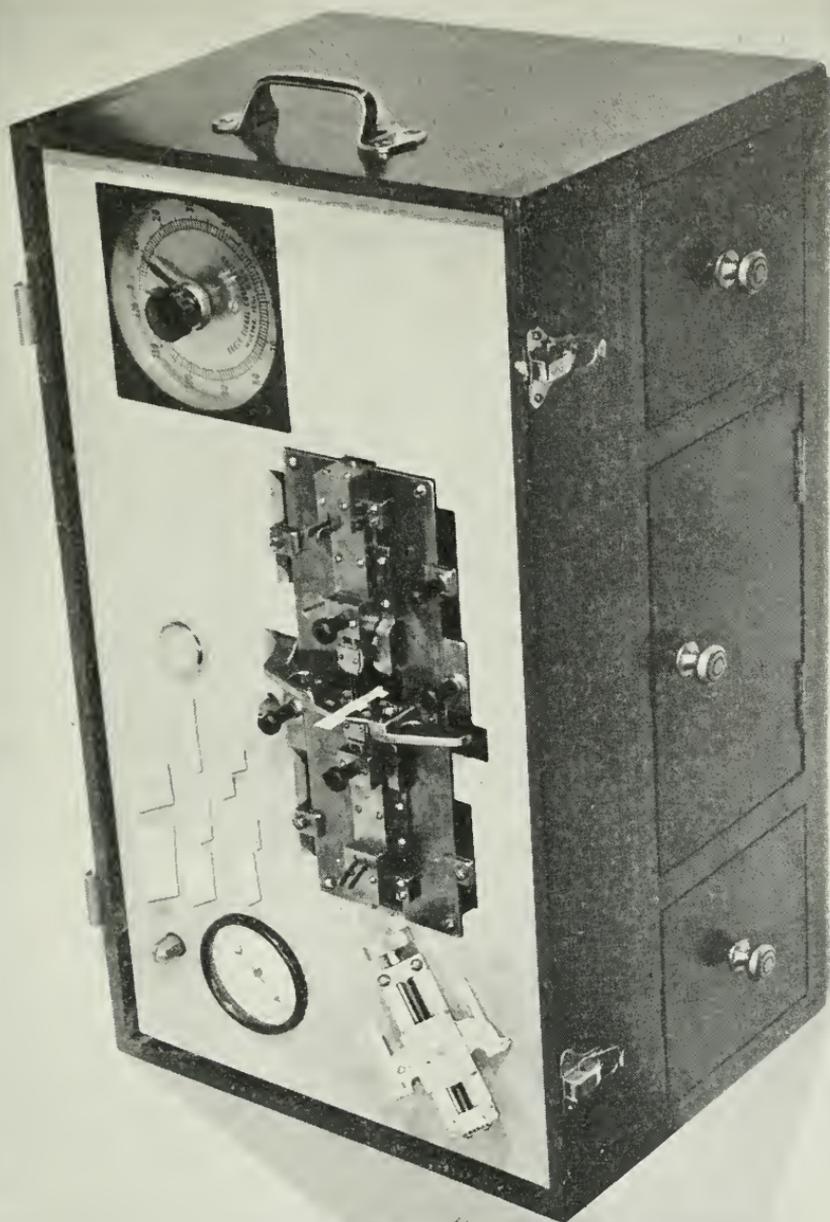


Fig. 13.8—Machine for attaching headed wires.

are aligned with respect to the crystal plate, which is properly located and held between sliding jaws as shown. Prior to the operation the silver spots are burnished and fluxed. The wires in their guides are then slid into contact with the plate. The hot blasts are raised from below so as to aim directly at the work. The solder is melted and the attachment completed in about 10 seconds when the blasts are withdrawn. Slight pressure is maintained on the wires during this operation by springs in the guides to force the head of the wire to seat on the spot. A small fillet is obtained around the head making the solder cover an area of about 35 to 40 mils diameter.

Crystal units using headed wire attachments have many advantages over those made with straight or hooked wires. The pull-off strength is more uniform and averages slightly better than that of hooked wires, despite the fact that only a fraction of the amount of solder used with hooked wires is employed. With this type of attachment the cratering effects encountered with previous methods have also been eliminated. The reduced quantity of solder on the face of the crystal plate effects a decided improvement in the temperature coefficient of the crystal unit as well as in its efficiency and stability. Although the heading of the wires and the subsequent cleaning and tinning operations involve more work, the process of making the attachments is simpler and quicker, since the use of individual wires is better adapted to making all the attachments in one operation. Headed wire attachments are more uniform in size and shape and give a more workmanlike finish to the job. This type of attachment has now replaced those using straight and hooked wires in virtually all designs of telephone type crystal units using wire supports. While the potential advantages of a headed wire type of attachment for crystal support wires had been known for many years, the practical exploitation of the idea depended on finding commercial means for producing the headed wires. The development of a suitable machine for this purpose was carried out by the Western Electric Company in close collaboration with the Laboratories. Figure 13.9 shows such a machine. The fine wire is fed through the lower mechanism to a die in which it is firmly clamped with a predetermined amount extending above the plate. This part of the wire is then cold-worked by multiple punches in the head of the machine until a conical shaped head is formed in the die cavity. As the individually headed wires are formed they are cut off to a definite length and expelled as the vise is released and the next wire brought into position. Cold heading of the wires is necessary in order to retain the elastic properties of the phosphor-bronze springs employed in the suspension. The operation of the tool is simple after the precise alignments of the die and punches have been made.

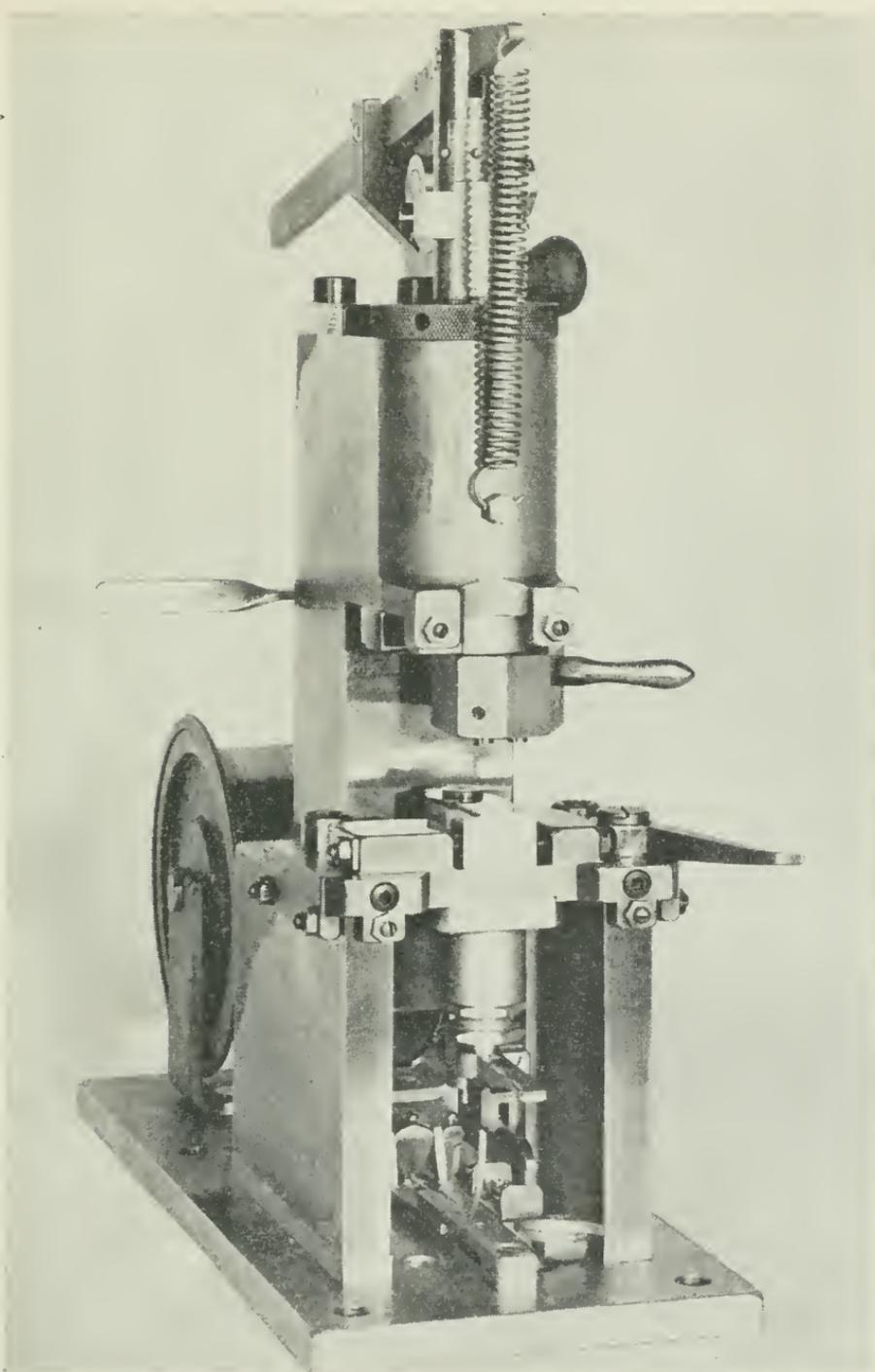


Fig. 13.9—Tool for cold-heading phosphor-bronze wires.

13.46 Mounting the Crystal Plate

After the suspension wires are affixed to the plate, they are then bent to serve as springs and to permit soldering into the cages as illustrated in Fig. 13.2. Two different types of springs are used, one of them involving one bend and the other two. The direction of the bends and the distances between them have been worked out so that the crystal will be displaced to about the same extent in all three directions for equal forces. The cages are of simple construction being made up of mica stampings and metal rods. The assembly of these parts is performed by welding little

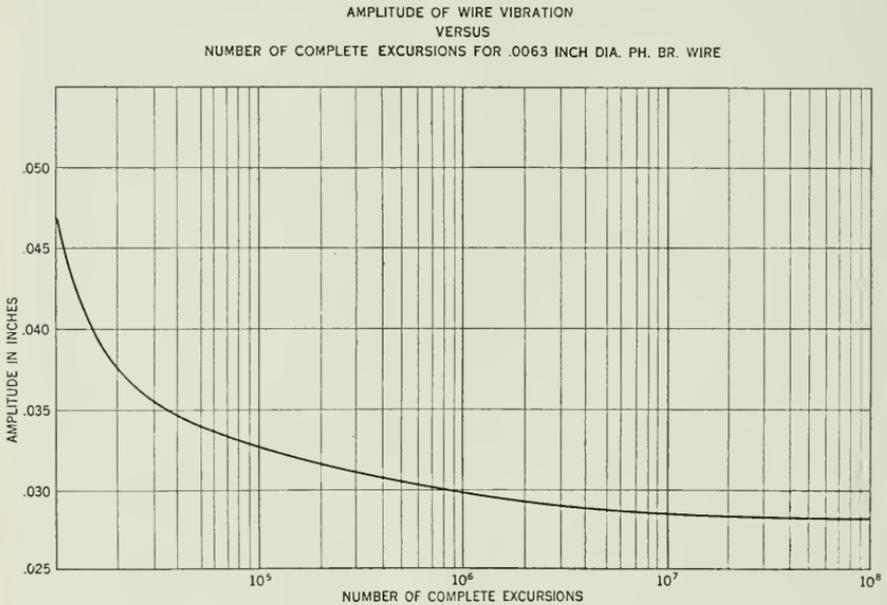


Fig. 13.10—Characteristic performance of phosphor-bronze spring wires.

eyelets, which are staked into the micas, to the rods. In the structures shown, the inside micas are provided with rectangular slots which limit the sidewise movement of the crystal plate from 25 to 30 mils. The end micas are spaced so as to limit the movement of the plate in the lengthwise direction by the same amount. Aside from being used as parts of the cage, the micas therefore serve as “bumpers” to prevent excessive displacement, and possible breakage of the wires or plate if the crystal units are subject to extreme vibration or shock. Figure 13.10 is an experimental curve showing the minimum number of excursions made by wire-mounted crystal plates vibrated at different amplitudes before wire failure occurs for 6.3 mil phosphor bronze spring wires with single bends. On the basis of these data, the

chosen spacing of 25–30 mils between the crystal plate and the bumper should ensure against any service failure of the unit in this regard. In order to center the crystal laterally and longitudinally in the bumper system, the plate is assembled first in the cage by means of spacers. The fine wires are then soldered to the vertical rods or “straights” as they are called, and the spacers removed leaving the plate suspended in position. In order not to set up any strains in the junctions of the wire to the straight which might tend to displace the plate after this operation, the spring wires are usually pre-formed to come within about 5 to 20 mils of the straight. The junction is made by immersing the intersection of the wire and straight in a ball of molten solder. As the wires are withdrawn the ball of molten solder comes with them, solidifying in the air and thus joining the fine wire to the straight without strain.

It will be noted from Fig. 13.2 that the wires to the longitudinal crystal are equipped with little weights close to the plate. This practice has been found desirable on virtually all types of crystal units to alleviate problems of wire resonance⁸ which arise in occasional units thereby causing high resistance as well as a shift in the frequency of the plate. Initially, while these effects were noted to some extent in the course of laboratory developments, it was not thought that they would be prevalent enough to warrant taking precaution to eliminate them by loading the wires, since they can usually be corrected by refloating and resoldering the crystal plate thereby changing the effective length of the wire. However, it has turned out that in manufacture a large enough percentage of crystals contain resonant wires to warrant the use of weights. For low-frequency crystals (up to 200 kc) solder balls are placed on the wire at the desired location using a method worked out in conjunction with the Western Electric Company. The process is performed in somewhat the same manner as that described above for connecting the crystal support wires to the straights, except that the weight of the solder deposited and the distance from the plate is more critically controlled. For higher-frequency crystals above 200 kc in which more precise positioning of the weight is essential, small metal discs are employed. They are threaded onto the mounting wire and held in the correct position by a definite amount of solder on the back to obtain the desired loading. Since the free length of wire must be accurately controlled, the manufacturing aspects of this job have been greatly simplified by the use of headed wires in which the variation in height of the solder cones is very small. The chart shown in Fig. 13.11 shows the weights of solder balls or discs and the position they should take on crystals having frequencies up to about one megacycle. It should be noted that the chart covers .0063” phosphor

⁸“Principles of Mounting Quartz Plates,” R. A. Sykes, *B.S.T.J.*, April 1944.

bronze wire. For any other diameter, d , of phosphor bronze wire, the new distance

$$X' = X \sqrt{\frac{d}{.0063}}$$

Earlier, it was mentioned that the supporting wires for the plates were formed with one or two bends. In addition to the function of suspension these bends also introduce changes in impedance along the wire thus mini-

LOCATION OF WEIGHTS ON MOUNTING WIRES OF QUARTZ CRYSTALS TO SUPPRESS WIRE VIBRATION DISTURBANCE

NOTE: Information shown is for 6.3 mil Phosphor Bronze Wire
 (For 3.5 mil P-b wire, weight should be multiplied by .50 and located at .75X)
 For 5 mil P-b wire, same weight should be located at .89X
 For 8 mil P-b wire, weight should be multiplied by 1.8 and located at 1.12X

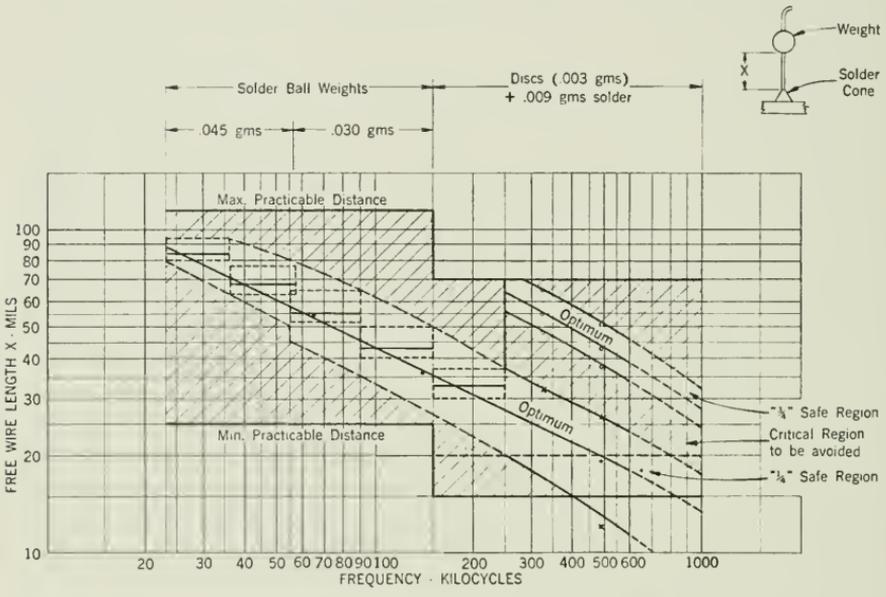


Fig. 13.11—Graph for determining placement of weights on wires for damping vibration.

mizing the possibility of trouble due to wire resonance. The use of a greater number of bends in the wire would tend to accomplish the same result as that of weights. However, the use of weights is considered more practical and has been adopted. As a result of this change, it is possible to employ wire supports having only one bend in virtually all crystals. In low-frequency crystals (below 2 kc) where the wave-length of the flexural wave in the wire is relatively long it is unnecessary to use weights since the wire length can be controlled adequately by the termination of the support wire at the straight. Depending upon the frequency, the desired length of wire is obtained by using either two or three direction bends.

13.47 *Housing of Crystal Units*

For the pressure-type units first discussed no provision was made for protecting or sealing them other than the hermetically sealed containers in which all the other associated components of the filter were enclosed. However, the wire-supported designs have been worked out so that each unit is sealed in its own individual container. Fortunately, the sizes of virtually all crystal units are in the range which permits the use of relatively inexpensive radio tube parts for these housings. There are many obvious advantages to the individually sealed unit. After adjustment and sealing it can be handled more readily in subsequent assembly operations. It is not subject to variations due to changes in ambient humidity and consequently does not restrict the assembly of apparatus to conditioned space. It can be made up and stored or shipped as an individual unit. It has a higher degree of stability. There is one small effect, however, in the case of units which are sealed in vacuum. Due to the absence of any gaseous medium around the crystal, a slight change in frequency is encountered when the tube is evacuated. However, this change is always the same for each particular type and size of crystal and can be allowed for in the final adjustment before sealing.

Most designs of crystals can be sealed in an atmosphere of dry air although better performance results from the use of vacuum. Some crystals must be sealed in vacuum for this reason. A decided advantage in favor of vacuum-sealed crystals is the elimination of acoustic effects from air resonance.

Both metal and glass tubes are used for housing crystal units. Initially it appeared that metal tube radio parts were ideally adapted to crystal use, and it was felt that, instead of welding the stem to tube, this sealing operation could be done by soldering. However, it was found that while sound solder joints could be obtained, extreme precautions were necessary to protect the button-type glass seals, through which the leads emerge, during the pre-tinning and soldering operations. Even with such precautions, it would have been essential to include in every vacuum type tube a means of detecting whether or not a leak had developed. For air-filled tubes at atmospheric pressure this would not have been necessary since minute leaks can be tolerated with little likelihood of the crystal being affected over a long period of time. The possibility of welding as is done in the case of radio tubes was considered but did not appear justified on the basis of equipment cost. Moreover, even with welding there still appeared to be problems from leakage and outgassing of the metal since, after the crystal is enclosed, the assembly cannot be exposed to high temperature to drive off adsorbed gases during the evacuation process. In view of these draw-

backs the use of metal tubes has been discarded in favor of glass tubes except in the case of a few special designs.

The procedure of mounting a crystal unit on a stem and sealing it in glass is much the same as for a radio tube. Figure 13.2 shows crystal units mounted on stems ready for sealing and also shows units sealed in glass and based. The extensions of the straights through the bottom micas are welded to the formed wires emerging from the glass seals. In the glass-sealing operation care must be taken not to heat up the assembly to the point where the solder attachments will be melted or even softened enough to permit the crystal to change position. To accomplish this it is necessary to use hot, sharp-pointed fires localized to the region where the seal will be made. The use of oxygen-gas flames is virtually essential to accomplish the seal quickly. Having the fires strike the bulb at tangency is also desirable. The ordinary type of glass-sealing head for use with gas-air fires is not well adapted to this work since the rotating pillars require the fires to be held too far away from the work thereby necessitating larger flames and consequently more heating up of the crystal unit assembly. The screening effect of the pillars as they revolve also slows up the work of the fires thus increasing the over-all heating of the assembly. A special glass-sealing machine developed for sealing crystal units is shown in Fig. 13.12. Immediately following the sealing operation the glass units should be placed in a suitable annealing box or leer where they can cool off very slowly. A large wooden block equipped with holes to admit the individual bulbs is convenient. The holes may be covered with a cloth to prevent air circulation.

After the units have cooled they are placed on a vacuum pumping station and evacuated. During the first half hour of pumping they are enclosed in a heated oven in which the ambient temperature is maintained at about 240°F. This drives off any traces of moisture that might have entered the tube prior to sealing. Following the heating interval, the tubes are pumped for another half-hour during which time they will have cooled down to room temperature. At this point the pressure in the tubes should be at the minimum of which the pump is capable of attaining. This value should be at most 20 microns and preferably less. However, with a six or eight-tube station better than 15–20 microns is not likely to be attained unless a liquid nitrogen trap is employed in the system for eliminating moisture.

After the pumping period, vacuum-type units are sealed off, with pump running, by melting the glass tubulation with a fine-pointed oxygen-gas flame as close to the stem as possible. If air is to be admitted, the pump is closed off from the system and dry air admitted to the tubes after which they are sealed off. After testing the crystal unit to see that it meets its requirements, the unit is equipped with a base in the same manner as followed for radio vacuum tubes.

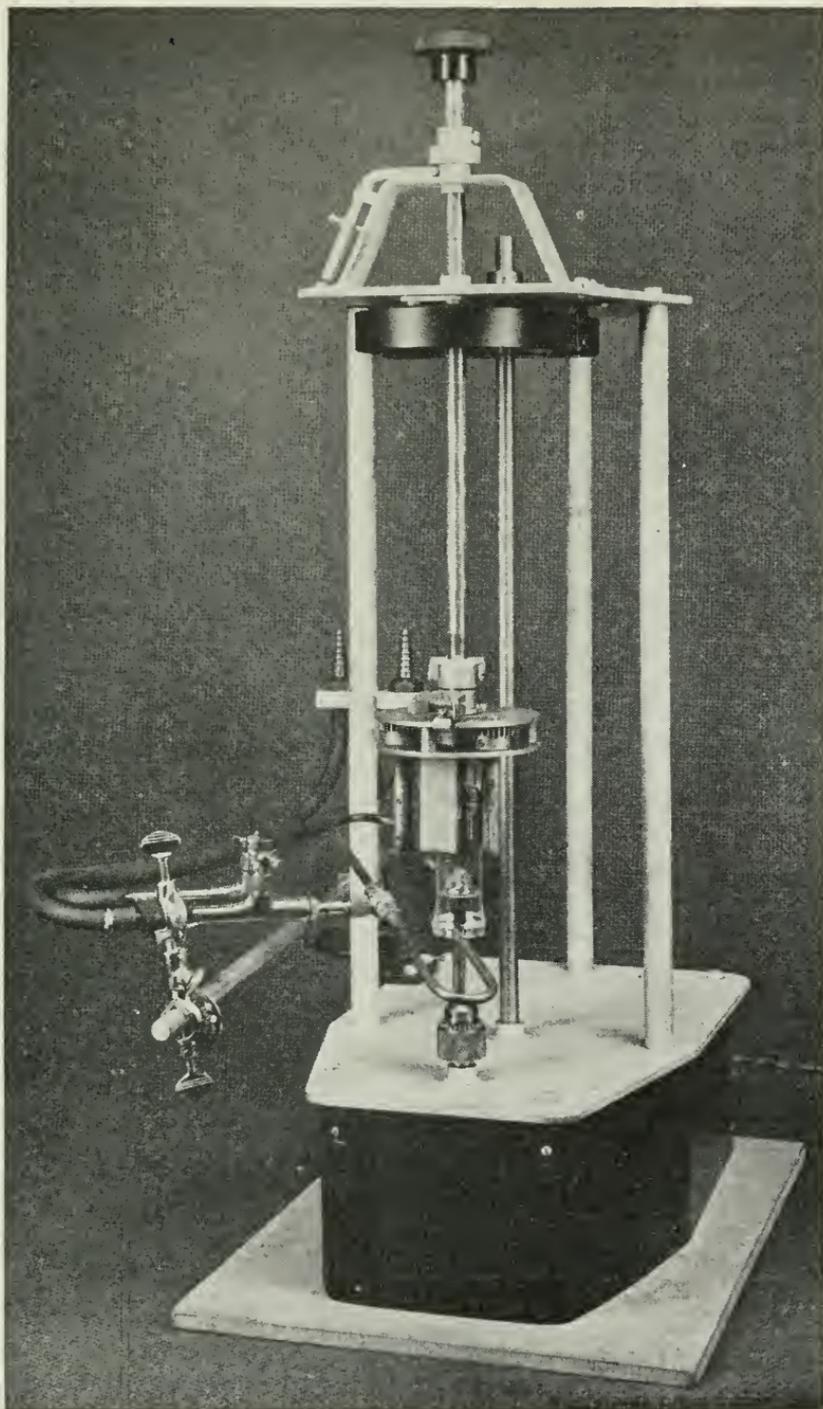


Fig. 13.12—Glass-sealing machine.

13.48 *Stabilization of Crystal Units*

Despite the close dimensional tolerances applying to the manufacture of the individual crystal plates, the exact frequencies are rarely realized in the mounted crystals. To bring the crystal to frequency it is therefore necessary to grind off minute layers from the lengthwise or widthwise edges of the plate depending on the mode of vibration. This adjustment, which causes superficial disruption of the quartz areas affected, results in unstable operation of the unit with respect to frequency and resistance. Unless the crystal plate is properly treated after these operations, considerable drift in these characteristics will take place, particularly during the initial service life of the unit. To alleviate this condition, the crystal units are first rough-adjusted to the approximate frequency and then heat-aged in an oven which subjects the units to several heating and cooling cycles between 240°F and 75°F. The units are then mounted in their cages as previously described and fine-adjusted after which they are again aged. This operation also tends to drive off any moisture which might be troublesome. With this type of accelerated aging the crystals are stabilized to the point where changes in performance can be detected only by the use of the most precise measuring equipment over long periods of time. Crystals so stabilized may generally be depended upon, at any one temperature, maintaining their frequency indefinitely within two or three parts per million provided they are not subjected to excess voltage.

13.49 *Cleaning of Crystal Units*

Throughout the manufacturing process it is essential that every precaution be taken to keep the crystal plate and associated parts absolutely free from contamination and dirt. The rigorous cleaning necessary before the spotting operation has already been discussed. In all the subsequent operations care must be taken to prevent the plates coming in contact with substances that might tend to cause corrosion. Any particles of foreign matter that may have accumulated on the plate or wires before rough and fine adjustments should be carefully washed off. Otherwise the performance and life of the crystal may be adversely affected. A suitable method for cleaning crystal units before sealing consists of washing and rinsing in chemically pure carbon-tetrachloride or other suitable solvent to remove grease, followed by washing and rinsing in hot distilled water at about 150°F. To facilitate removal of unwanted substances, the parts should be scrubbed gently with a soft brush or agitated in the solution during this process. The use of pure alcohol (95%) in addition to carbon-tetrachloride is also good for this process, but is not essential. The cleansed crystal units should be carefully dried out and protected from further contamination prior to the sealing operation.

13.5 CONCLUSION

In ending I should like to acknowledge the aid and useful suggestions given me by Mr. C. E. Lane and my other associates in preparing this article. I should also like to reiterate the fact that the status of the art as described was reached after many years of pioneering development by many engineers. In some cases the names of individuals associated with specific contributions of a major nature have been mentioned.

CHAPTER XIV

Effects of Manufacturing Deviations on Crystal Units for Filters

By A. R. D'HEEDENE

14.1 THE EFFECT OF DEVIATIONS IN THE CHARACTERISTICS OF CRYSTAL UNITS ON FILTER PERFORMANCE

THIS chapter emphasizes primarily the need for close control in the manufacture of crystal units for use in filters. The first telephone use of crystal units in the commercial manufacture of filters was made by the Western Electric Company in about 1936. To make such commercial manufacture practical, it was necessary to establish accurate design information and allowable manufacturing tolerances. The quantitative data collected for this purpose provided the chief source of material for this chapter. While the data is quite extensive, it will be observed that there are still some factors which must be treated qualitatively.

While filter crystal units are like oscillator crystal units in that they must have low internal dissipation and a close control of resonant frequency, they are different in that many additional characteristics of the filter crystal units must also be controlled accurately. Two typical illustrations will demonstrate how characteristics other than resonant frequency and Q may react on filter performance.

The first characteristic considered is the slope of the reactance with frequency curve in the vicinity of the series resonant frequency. This slope is sometimes referred to as the impedance level of the crystal unit. A convenient measure is the inductance of the equivalent electrical circuit. When this inductance departs from its nominal value, the performance of the filter using the crystal unit may undergo appreciable change. This is particularly true of filters in which the schematic contains a lattice or some other type of bridge circuit with crystal units contained in all the bridge arms. For example, in Fig. 14.1 the solid curve illustrates the transmission characteristic obtained from a lattice-type crystal filter, in which both the series branches and the diagonal branches contain two balanced crystal units. High loss results from a close impedance balance between the branches of the lattice. When the inductance of any of the crystal units departs from its nominal value, the bridge balance is disturbed and the transmission characteristic of the filter is changed. The two dotted curves of Fig. 14.1 illustrates the characteristics that result when the inductance

values of the crystal unit in either branch depart from their nominal values by about one per cent. A negative departure in one branch results in about the same effect on performance as a positive departure in the other branch. The difference between the two curves shown on Fig. 14.1 is that one assumes a positive departure and the other a negative departure for the inductance of a branch.

Due to the close impedance balance which is required for these filters, the effect of small departures in resonant frequency will produce rather large variations in the transmission characteristic. For example, departures of about 10 cycles per second in the crystal units of either branch will produce variations in discrimination of about the same type and magnitude as those

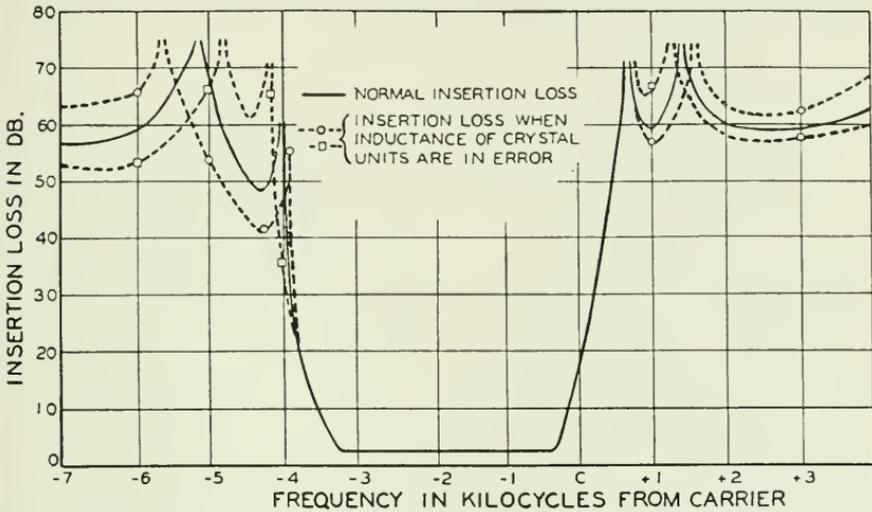


Fig. 14.1.—The insertion loss characteristic of a crystal band-pass filter as affected by deviations in the inductance of the crystal units.

illustrated in Fig. 14.1 for departures in inductance. On the other hand, if the crystal units of both branches exhibit equal departures the entire transmission characteristic will be shifted by the frequency departure of the crystal units, and there will be no loss in discrimination.

Another way in which deviations in the properties of crystal units may react on filter performance is illustrated by the schematic and curves shown in Fig. 14.2. The schematic is the equivalent electrical circuit of a narrow band filter, using two balanced quartz crystal units. The filter is designed to provide a passed band of about 10 cycles per second with distortion of less than 0.2 db. The insertion loss characteristics show that the desired transmission can be obtained for various magnitudes of effective resistance as long as the resistances in the series and diagonal branches are equal.

However, if the effective resistance in one branch is twice as large as that in the other branch a highly distorted characteristic results as shown by the middle curve of Fig. 14.2.

Both of these illustrations show that filter performance is degraded rapidly if the crystal units of the lattice have characteristics which depart from their nominal values by different extents for the two branches. A similar effect is produced when the temperature coefficient of resonant frequency for the crystal units in one branch differs from the temperature coefficient of the units in the other branch. Deviations occurring in a single unit may also affect filter performance. Such deviations include the presence of unwanted resonances of even weak amplitude, inadequate insulation resistance between the metallized coatings or unbalance between the halves of plates on which the coating has been divided.

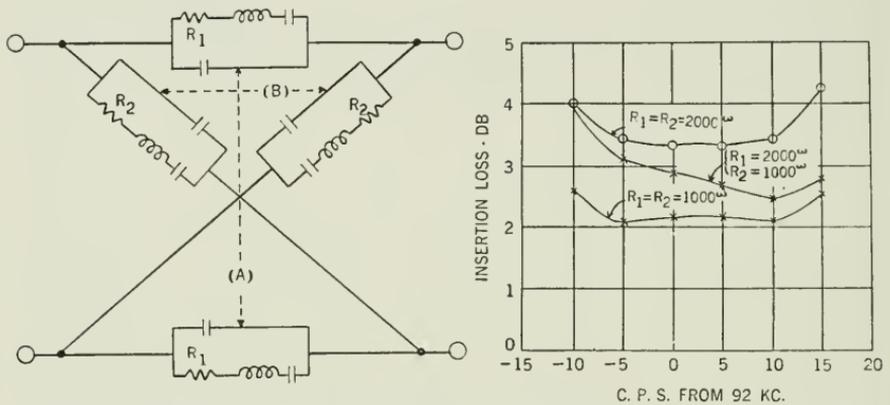


Fig. 14.2.—Effect of deviation in the effective resistance of crystal units on the distortion characteristic of a crystal filter.

The importance of controlling the electrical characteristics of the crystal units is indicated from the above considerations. It is pertinent to correlate deviations in the mechanical properties of the crystal unit with the deviations in electrical characteristics. This is the subject of the succeeding sections. Consideration is restricted to the plates commonly used in filters, that is, X-cut plates, vibrating in extensional or flexural modes, and GT-cut plates.

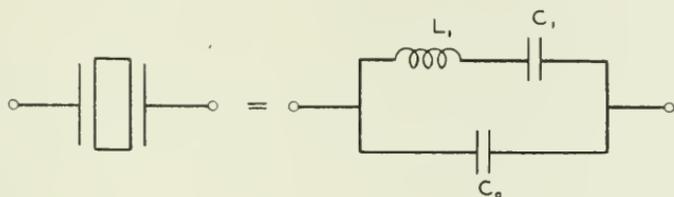
14.2 THE EFFECT OF DEVIATIONS THAT OCCUR IN THE MANUFACTURE OF QUARTZ PLATES

Quartz is an anisotropic material. Accordingly, plates cut from a quartz crystal exhibit elastic and piezo-electric properties which depend on the orientation of the plates with respect to the principal axes of the crystal.

For that reason, any deviation in the orientation of the plates from nominal will affect the electrical characteristics of the crystal units. In addition, these characteristics are affected by imperfections in the plates due to deviations in linear dimensions, to the presence of flaws, or to the condition of the surface of the plates. The effects of these deviations differ for various cuts of crystal plates, for plates of various shapes and for the various modes of vibration. In the following paragraphs, each type of deviation will be considered in turn and data will be presented to show its effect on the characteristics of crystal units using the various types of plates.

14.21 DEVIATIONS IN THE ANGLE OF ORIENTATION

Accurate information is available on the effect of deviation in angle of orientation on the characteristics of X-cut plates vibrating in the exten-



$$C_0 = \frac{K \omega \ell}{4 \pi t} \times \frac{1}{9 \times 10^{11}} \text{ FARADS}$$

$$f_R = \frac{1}{2 \ell \sqrt{\rho s'_{22}}} \text{ CYCLES PER SECOND}$$

$$L_1 = \frac{\rho s'_{22}{}^2 t \ell}{8 d_{12}'{}^2 \omega} \times 9 \times 10^{11} \text{ HENRIES}$$

$$C_1 = \frac{8 d_{12}'{}^2 \omega \ell}{\pi^2 s'_{22} t} \times \frac{1}{9 \times 10^{11}} \text{ FARADS}$$

Fig. 14.3.—Equivalent electrical circuit of piezoelectric crystal.

sional mode. The relation between the electrical characteristics of this type of vibration and the properties of the quartz are shown in Fig. 14.3. This information, with minor changes, is reproduced from a preceding publication.¹ In Fig. 14.3: ℓ, w and t are the length, width and thickness respectively of the plate; K is the dielectric constant; ρ is the density; d'_{12} is the piezo-electric constant; and s'_{22} is the modulus of compliance (inverse of Young's modulus). All these individual quantities are expressed in electrostatic units. The quantities which depend on the orientation of the plates are the piezo-electric constant and the modulus of compliance. The symbols for these quantities usually are primed when they are used for a generalized orientation. When unprimed, the symbols designate quantities measured along the principal axes. For X-cut plates, deviations of the plane of the major surface from the YZ plane have relatively

¹“Electrical Wave Filters Employing Crystals with Normal and Divided Electrodes”, W. P. Mason and R. A. Sykes, *B. S. T. J.*, April 1940, page 222.

small effect, while variations in the angle of rotation about the X-axis have a relatively large effect on these quantities.

Mason has shown² how the magnitudes of the piezo-electric constants and the moduli of compliance for any angle of rotation may be derived from their magnitudes along the principal axes of quartz. Using these equations and the magnitudes for the principal axes tabulated in a recent paper³ by Mason, d'_{12} and s'_{22} have been calculated as a function of the angle of rotation of the plates about the X-axis. In turn, the frequency and inductance constants have been calculated as a function of the angle of rotation, using the relations shown in Fig. 14.3. Figure 14.4 is a plot of the frequency and inductance constants as a function of the angle of rotation for angles between about -70° and $+70^\circ$. It shows how the inductance and resonant

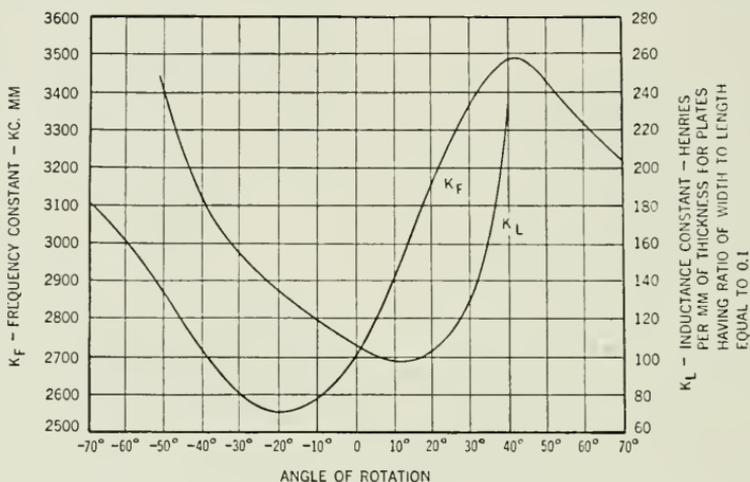


Fig. 14.4.—Frequency and inductance constants of X-cut quartz plates as a function of their angle of rotation around the X-axis.

frequency of these plates will change with deviations in the angle of rotation. The frequency constant used is the product of the resonant frequency in kilocycles and the length of the plate in millimeters. The inductance constant used is the inductance per millimeter of thickness of a plate having a width-to-length ratio equal to 0.1.

The change of inductance and resonant frequency with deviations in angle of rotation about the X-axis is of particular interest for the angles of rotation most commonly used, that is for -18.5° and for $+5^\circ$. The calculations indicate that for an X-cut plate rotated -18.5° , a deviation of $\pm 1^\circ$ in the angle of rotation will change the inductance by $\mp 1.2\%$

² "Electrical Wave Filters Employing Quartz Crystals as Elements", W. P. Mason, *B. S. T. J.*, July 1934, equations on page 451; also in Chapter VI.

³ "Quartz Crystal Applications," W. P. Mason, *B. S. T. J.*, July 1943.

respectively, and the resonant frequency by $+0.05\%$ and -0.02% , respectively. For an X-cut plate rotated $+5^\circ$, a deviation of $\pm 1^\circ$ will change the inductance -0.9% , and $+0.6\%$, respectively, and the resonant frequency by $\pm 0.7\%$.

Deviations in angle of rotation will, in general, affect temperature coefficient. The effect is illustrated by Fig. 1.19 of Chapter I,³ which, for X-cut plates, shows the relation between temperature coefficient and angle of rotation. This curve shows that the temperature coefficient is practically zero for an angle of rotation of $+5^\circ$. For that reason this particular cut is used whenever a low-temperature coefficient is desired. The curve also shows that at this point the slope of temperature coefficient as a function of angle of rotation is zero. Hence, for the $+5^\circ$ X-cut plate, which is most important from the standpoint of temperature coefficient, there will be little change due to a deviation in the angle of rotation.

In GT-cut plates the effect of deviation in the angle of orientation must be considered in combination with deviations in linear dimensions. Mason shows⁴ that for an angle of rotation of $+51$ degrees 7.5 minutes and a width-to-length ratio of $.859$, a temperature coefficient close to zero may be obtained from -25° C to $+75^\circ$ C. He also has shown that this temperature coefficient varies with both the angle of rotation and the width-to-length ratio. Because of this, it has been found possible to compensate for small deviations in the angle of rotation by adjusting the linear dimensions. The net effect of a deviation in angle of rotation, after it has been so compensated, is to raise (or lower) the temperature region of zero temperature coefficient by 11° C for each 10 -minute increase (or decrease) in angle of rotation. In GT-cut plates, the width dimension directly controls the primary resonance. For this reason, it is preferable to adjust temperature coefficient by varying the length dimension rather than the width. The crystal plates are cut larger than desired. The frequency of resonance and the temperature coefficient are then adjusted simultaneously by grinding either the width or length dimension as required. Experimental work carried on by L. F. Willey of the Laboratories shows that an increase of 1.0 per cent in the width-to-length ratio results in an increase in temperature coefficient of approximately $+1.35$ parts per million per degree C. In his experimental work Willey used the ratio of the secondary to the primary frequency as a convenient measure of the width-to-length ratio. The inductance constant for GT plates, which is about 17 henries per millimeter of thickness, will increase by less than 1% for deviations in any of the angles of rotation of as much as 30 minutes. However, the inductance may depart appreciably from nominal due to the adjustment of width and length dimensions.

⁴ "New Quartz-Crystal Plate, GT, Produces Constant Frequency Over Wide Temperature Range", W. P. Mason, *Proc. I. R. E.*, May, 1940, page 220.

14.22 DEVIATIONS IN LINEAR DIMENSIONS

In the case of X-cut plates, the length dimension is used to control the location of their resonant frequency. The length is lapped to its final dimension after all other processes have been completed, so that this dimension will be such as to compensate for the effect of any other deviations that may have occurred. The sensitivity of this adjustment depends on the mode of vibration. For plates vibrating in their extensional mode, the resonant frequency is inversely proportional to the length, as shown in Fig. 14.3, while for plates vibrating in their flexural mode, the resonant frequency is inversely proportional to the square of the length. The amount of the adjustment required depends on the magnitude of the frequency errors that may have been introduced due to deviations in the width or in the angular orientation of the plates or due to still other causes. The magnitude of such frequency errors, in turn, depends to a considerable degree on the angle of rotation of the plates. For example, it was shown in Section 14.21 that a deviation in the angle of rotation of a $+5^\circ$ plate changes its resonant frequency more than ten times as much as a similar deviation in a -18.5° plate. It must be noted that the adjustment of length compensates for frequency errors only and that errors in inductance or temperature coefficient may be increased by such adjustment.

Deviations in the thickness dimension principally affect the impedance level of the plates. As shown by Fig. 14.3, the inductance is directly and the capacity inversely proportional to the thickness. In the case of GT-cut plates the thickness dimension is important also because it controls the location of the most prominent unwanted resonances, which arise from vibrations in thickness flexure. However, plates are designed to avoid critical thicknesses and small deviations from the nominal thickness will not usually result in plates having unwanted resonances.

Deviations in the width dimension affect the equivalent electrical characteristics appreciably. The effect of deviations in width on the frequency of X-cut plates vibrating in their extensional mode can be deduced from the curves of Fig. 14.5. The curves show that this effect is more pronounced for larger values of the width-to-length ratio where coupling with the width extensional mode becomes appreciable. For a -18.5° plate with a width-to-length ratio of 0.8 an increase of 1% in the width dimension will decrease the frequency by about .04%. For a $+5^\circ$ plate with a width-to-length ratio of 0.4 an increase of 1% in the width dimension will also decrease the frequency by about .04%, but for a ratio of 0.6 the decrease in frequency will amount to 0.13%.

Similar information is available for crystals vibrating in width flexure from the measurements published by Harrison⁵ and the calculations pub-

⁵ "Piezo-Electric Resonance and Oscillatory Phenomena with Flexural Vibration in Quartz Plates", J. R. Harrison, *Proc. I. R. E.*, Dec. 1927.

lished by Mason.⁶ This information shows that for small ratios of axes the resonant frequency will be directly proportional to the width dimension. However, as the ratio is increased to 0.5, a change of 1% in the width dimension will change the frequency by only 0.5%.

The effect of the width dimension on the inductance of the plates frequently is important. Fig. 14.6 illustrates the relation between inductance and the ratio of axes. From these curves the effects of deviations in width can be deduced. For the two longitudinal plates, the inductance is almost inversely proportional to width. For the flexure plate, the decrease of

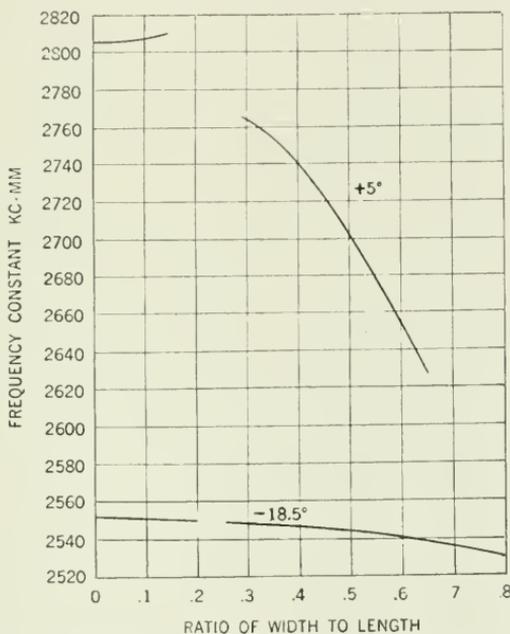


Fig. 14.5.—Frequency constant of the longitudinal mode of X-cut quartz plates as a function of their ratio of width to length.

inductance with increase in width is much more rapid. With a ratio of axes of 0.6 the inductance decreases about as a square power, while with a ratio of 0.1 the decrease is about as the third power.

The width dimension of the +5° plate has an appreciable effect on the temperature coefficient of the plate. Mason has shown³ that while the temperature coefficient is zero for a long narrow bar, it increases quite rapidly as the width dimension increases, due to coupling between the face shear and the longitudinal modes. In the case of an -18.5° plate, coupling with other modes is relatively weak. Hence its temperature co-

⁶ "Motion of a Bar Vibrating in Flexure Including the Effects of Rotary and Lateral Inertia", W. P. Mason, *Jour. Acous. Soc. America*, April, 1935, pages 246-249.

efficient, which is about 25 parts per million per degree C, does not change appreciably with changes in width. For a $+5^\circ$ plate vibrating in its flexure mode, Fig. 14.7 illustrates measurements made on the variation of temperature coefficient with ratio of axes. For all these X-cut crystals, it may be observed that deviations of 1% in the width dimension will not

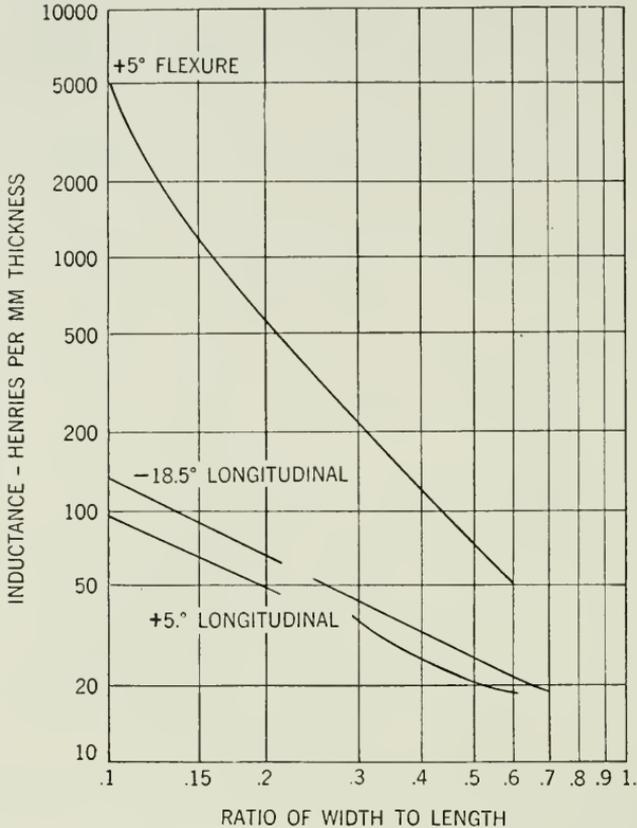


Fig. 14.6.—Inductance of the crystal units used in filters as a function of the cuts of the plates and their ratio of width to length.

change the temperature coefficient by more than 5%. Such changes are usually negligible.

14.23 INTERNAL DEFECTS

Internal defects in the quartz plates may have a large effect on their electrical characteristics. These defects vary so widely in type, size and concentration that it is impossible to predict the effects quantitatively. General comments regarding the results that may be expected for various

defects are described in Chapter IV⁷. The conclusions drawn there for oscillator plates are also applicable to filter plates. These are: (1) Evidence that a particular defect is permissible in a given type of plate does not prove that a similar defect is permissible in some other type of plate, and (2) proof that a particular defect is permissible in a given type of plate can be obtained only by a statistical study.

Some qualitative statements can be made regarding the effect of mechanical flaws. Cracks result in instability of resonant frequency and effective resistance and must be avoided. The effect of inclusions or chips depends

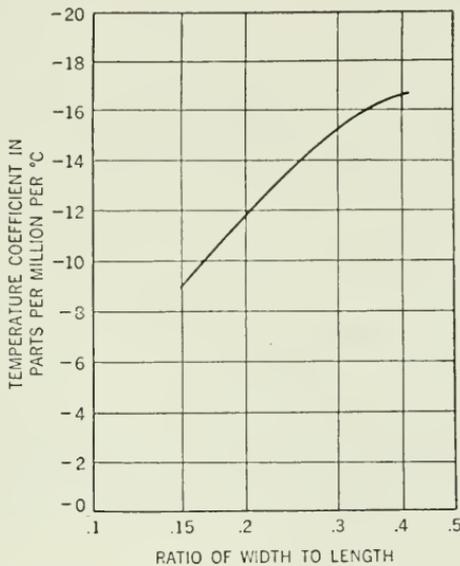


Fig. 14.7.—Temperature coefficient for $+5^\circ$ flexural crystal units as a function of the ratio of width to length of the plates.

on the size of these defects relative to the size of the finished plates and also on their location in the plate.

Twinning in quartz may be either of the optical type (Brazil twin) or of the electrical type (Dauphiné twin)⁸. The effect of these two types of twinning on the performance of oscillator crystal units has been described thoroughly in Chapter V⁹.

When optical twinning is present, the plate will exhibit the same elastic properties throughout, but the two portions of the plate will tend to expand

⁷ "Raw Quartz, its Defects and Inspection", G. W. Willard, *B. S. T. J.*, October 1943, pp. 338-361.

⁸ "The Properties of Silica", R. B. Sosman, Chapter XII.

⁹ "Use of the Etch Technique for Determining Orientation and Twinning in Quartz Crystals", G. W. Willard, *B. S. T. J.*, January 1944, pp. 11-51.

and contract in opposite phase. Hence there is little change in frequency constant or temperature coefficient, but there will be a large change in inductance. The change in inductance can be estimated roughly by comparing the twinned plate with an untwinned plate in which activity has been reduced by removing electrical charge from part of the surface. The area of the surface from which this charge is removed would be twice the twinned area and located at about the same position in the plate.

It is believed that a small amount of electrical twinning is more serious than a similar amount of optical twinning, because the twinned areas are of opposite angular sense. Each of the two areas has a different modulus of compliance and the effective modulus of the plate has a value intermediate between the two different values of modulus. Therefore, the frequency constant of the plate will be intermediate between that of the desired cut and its electrical twin. For a small amount of twinning, the direction and rate of change of frequency can be estimated from the comparison shown on Table I between the standard filter cuts and their electrical twins.

TABLE I

Filter Plate	Frequency Constant—kc. m.m.	Electrical Twin	Frequency Constant—kc. m.m.
-18.5°	2560	+18.5°	3120
+5°	2815	-5°	2650
+51.1° (GT)	3280	-51.1°	2610

This verifies the experimentally observed fact that for -18.5° X-cut plates, twinning increases the frequency, while for +5° X-cut and GT plates, twinning decreases the frequency. Even for small amounts of twinning the inductance will increase rapidly for plates of any orientation. When the amount of twinning becomes large, the equivalent inductance approaches infinity. That is, the crystal will not be set in motion by an applied voltage.

The quantitative effect of twinning (probably electrical) has been measured on one set of plates by R. M. Jensen. Figure 14.8 includes a photograph of the plates used, illustrating the extent of the twinning in each. All of the plates are -18.5° X-cut plates, having the dimensions 30.88 x 10.56 x .86 mm. The tabulation below the photograph compares the inductance and resonant frequency measured for each of the plates with the one, designated AN-3, which shows the least effect of twinning. While there is a good correlation between the amount of twinning in the plates and their change in electrical performances, it is not practical to estimate accurately the effect of a given amount of twinning. For this reason, crystal plates having any twinning should not be used for crystal units for filters.

14.24 ETCHING

The surface condition of the quartz plates also has some effect on crystal characteristics. This surface condition is determined in large part by the lapping operation used to obtain final dimensions. As described in Chapter



Plate Designation	Percentage Increase over Values Measured for AN-3	
	Inductance	Resonant Frequency
AN-1	+22.12	+ .50
AN-2	+27.20	+ .59
AN-3	0	0
AN-4	+3.03	+ .04
AN-5	+3.12	+ .01
AN-6	+276.54	+5.01
AN-7	+1.58	- .03
AN-8	+14.21	+ .37
AN-9	+738.00	+6.63
AN-10	+32.44	+ .83

Fig. 14.8.—Effect of various degrees of twinning on the performance of -18.5° X-cut quartz crystal plates.

XIII¹⁰, the plates are given a final lap with 400 or 600-mesh carborundum. This, in turn, is followed by an etching bath which removes foreign particles. A short etch, about eight minutes in 47% hydrofluoric acid, has been found adequate to ensure firm adherence of the metal coating to the quartz. On

¹⁰ "The Mounting and Fabrication of Plated Quartz Crystal Units," R. M. C. Greenidge, this issue of the *B. S. T. J.*

the other hand, the use of a relatively long etch, 30 minutes or more, is desirable when a high Q is desired. The long etch also results in an improved stability of the resonant frequency as a function of current. This will be discussed in a subsequent paragraph. A disadvantage of a long etch is the difficulty of controlling the etching process within close tolerances. The variations in rate of removing material may be sufficient to affect the uniformity of the linear dimensions of the plates.

These factors indicate that etching is an important process in preparing crystal plates. A close control must be maintained on the strength of the acid, the uniformity with which the surfaces of the plates are exposed and the duration of the exposure.

14.3 THE EFFECTS OF DEVIATIONS DURING FABRICATION OF WIRE-SUPPORTED UNIT

As described in Chapter XIII¹⁰, two types of mountings have been developed for supporting crystal plates, the Pressure Type and the Wire-Supported Type. The wire-supported type of mounting is the more recent development and has resulted in crystal units which have a much higher degree of stability and can be reproduced within much closer tolerances than the units using the pressure type of mounting. Since this chapter is concerned chiefly with the problem of obtaining a high degree of precision in crystal units, the discussion is restricted to the wire-supported type of mounting.

14.31 SILVER SPOTTING

For the wire-supported type of mounting the first operation is to bake small silver spots on the surface of the crystal plates. In the application of these silver spots to the crystal plates three factors are of importance in their effect on the characteristics of the plate, namely, the size of the spot, its location, and the firing temperature. Since in all crystal designs to date the silver spots are applied at or near the nodal line of the crystal plate the principal effect of the spots is to increase the stiffness of the plate, so slightly increasing the frequency of resonance. Variations of an appreciable magnitude in either the amount of silver paste used (that is, the size of the spot) or in the location of the spot with respect to the nodal line will change the resonant frequency of the plate. Such changes could be corrected later, when the plates are adjusted for resonant frequency, as long as the length is increased sufficiently to allow such adjustment. However, if the length be increased sufficiently to allow for extreme cases, average adjusting time will be increased materially, while if the allowance is insufficient some of the plates may be unusable. For this reason, close control of the size and location of the silver spots is well justified.

In baking the silver spots, care must be taken to prevent "heat" twinning. If the temperature of a quartz plate is raised above the inversion point (573°C) and then is reduced again, the plate will be electrically twinned.⁸ The firing temperature of the silver paste currently used for the spots is not many degrees below this inversion point. Hence, the firing temperature may easily become so high as to result in twinned plates. In addition, it has been observed that the twinning may occur at a considerably lower temperature if the plate is subjected to large thermal stress. For this reason, care must be taken to heat the plates uniformly during the baking operation.

14.32 DIVISION OF COATING

The next operation is to evaporate a coating of silver on the surface of the quartz plates. The plates must be thoroughly cleaned before this coating is applied in order to ensure firm adherence of the coating. Poor adherence may cause the coating to peel off the plate, changing all of the electrical characteristics of the plate. In many cases the coating must also be divided.¹¹ Two methods are in general use for dividing the silver coating on crystal plates, namely, an abrasive method and an electrical stylus method. In general, the abrasive method of dividing the coating is superior to the electric stylus for all cases requiring a simple straight line division, but it has not been found practical for complicated divisions such as are desirable for harmonic longitudinal plates and flexure plates.

In using the abrasive method for dividing the coating only two factors are likely to change the characteristics of the crystal plate, these being the location and the width of the dividing line. Deviations in the location of the dividing line from the lengthwise center line for a longitudinal plate will affect the capacity and inductance balance between the two halves of the plate. Deviations in the width of a properly centered dividing line will cause changes in the inductance of the plate since for a given plate the inductance is a function of the ratio of the plated area to the total area of the plate. So, for a wide crystal plate deviations in the width of the dividing line will be negligible while for narrow plates these deviations can cause an appreciable change in the inductance of the plates.

When the electric stylus is used for dividing the coating, the location and the width of the dividing line again will affect the performance of the plates. In addition, varying amounts of twinning will occur along the division line apparently due to instantaneous high temperature gradients introduced by burning of the silver at the point of contact of the stylus. In measure-

¹ Loc. cit.

⁸ Loc. cit.

¹¹ "Crystal Channel Filters for the Cable Carrier System", C. E. Lane, *B. S. T. J.*, January 1938, pp. 125-136.

ments made on a group of -18.5° X-cut crystals on which the coating has been divided carefully with an electric stylus, the increase in the inductance of the plates ranged from 1.4% to 2.6%. Any twinning resulting from the dividing operation will also change the resonant frequency of the plates.

14.33 SOLDERING OF WIRES TO PLATES

The next process, that of soldering the supporting wires to the crystal plate may have considerable effect on the performance of the unit. The deviations which may be introduced depend on the amount of solder used, the location of the solder button with respect to the nodal line of the plate,

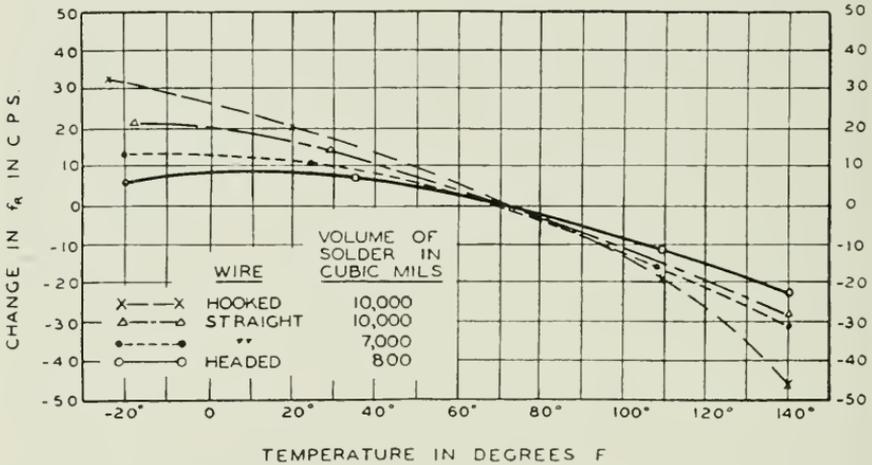


Fig. 14.9.—Change of resonant frequency with temperature of $+5^\circ$ X-cut quartz crystal plates. The curves show that when the volume of solder used for joining the plates to the supporting wires is appreciable compared to the volume of the plates, the frequency-temperature coefficient characteristic is affected by the volume of solder.

the shape of the solder button, and the possible twinning of the plate during the soldering operation.

The amount of solder used in forming the joint of the wire to the plate becomes extremely important when the plate is small. For example, Fig. 14.9 illustrates the changes in the frequency-temperature characteristic resulting from the use of varying amounts of solder for a particular size of plate. The units on which these measurements were made used X-cut $+5^\circ$ plates of 16 mm x 6 mm x 0.5 mm. The types of wire referred to in the figure, that is, hooked, straight and headed, were described in Chapter XIII¹⁰. The frequency-temperature characteristic expected on the basis of measurements made on crystal units using larger plates is approximated

¹⁰ Loc. cit.

closely by the solid curve. This solid curve actually was obtained from measurements made on crystal units using plates supported with headed wires and using a very small amount of solder. The other curves indicate that the temperature coefficient may be increased appreciably due to the presence of a larger amount of solder. Further, when the larger amounts of solder are used, the characteristics depend on the exact amount of the solder, so that the characteristics represented by the dashed curves are hard to reproduce.

The amount of solder used in this operation also affects the Q of the crystal unit and its resonant frequency. Measurements using several crystal plates of relatively small sizes have shown improvements in Q of as much as 25 per cent when headed wires are used over that obtained with other wires using larger amounts of solder.

Variations in the consistency of the solder joint will, of course, affect the adherence of the supporting wire to the plate. A poor joint will result in a high effective resistance for the crystal unit and will generally cause instability both in resistance and in resonance frequency.

In soldering the supporting wire to the crystal plate two methods have been used for melting the solder; namely, the soldering iron, and the hot-air blast. With either method, lack of sufficient control can seriously change the electrical characteristics of the plate due to twinning. It has been observed that this twinning occurs when there is a large temperature gradient in the quartz, even at temperatures well below the inversion point. Experimental work by G. W. Willard has indicated that it may occur even when the temperature of the soldering iron is as low as 300°C. To avoid such twinning during the soldering operation, it has been found desirable to raise the temperature of the entire plate to just below the melting point of the solder.

Twinning, when it occurs, will affect the crystal plate by causing an increase in inductance, a change in the resonant frequency, increased effective resistance, and a change in the temperature coefficient, as stated previously. Also, in crystals with divided plating there will be an inductance unbalance between the two halves of the crystal plate set up due to unequal amounts of twinning. Several measurements made, using GT plates at 160 kc, showed that twinning during the soldering operation decreased the resonant frequency in a range from 200 to 100 cps and the temperature coefficient of the units ranged from 2 to 6 times that of units using untwinned crystal plates.

14.34 EFFECTS DUE TO WIRE RESONANCE

As described in Chapter VIII¹², the characteristics of crystal units may be changed due to vibrations set up in the supporting wires. When any

¹² "Methods of Mounting and Holding Crystals", R. A. Sykes, *B. S. T. J.*, April 1944.

one of the wires is not located exactly on a node of the plate, the plate will set the wire into vibration. For certain critical lengths of the wire, it will offer considerable resistance to this motion and there will be a rapid increase in effective resistance and some change in resonant frequency of the crystal plate.

The effect of wire vibration can be described in terms of its electrical analogy. The vibrating wire, clamped at its far end, may be considered a rather special electrical transmission line open-circuited at its far end. When viewed from the crystal plate the impedance changes rapidly with

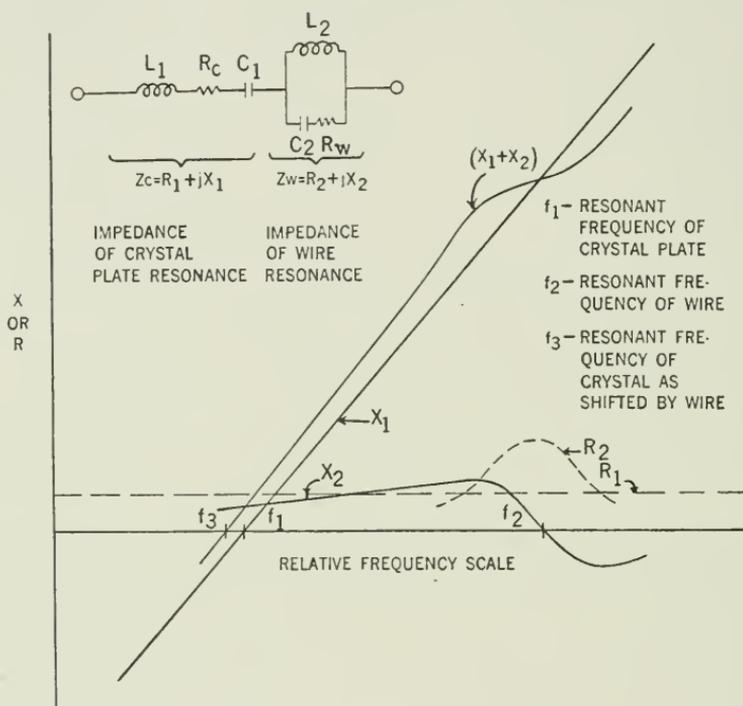


Fig. 14.10.—Effect of wire resonance on the resonant frequency of a crystal unit.

frequency in a succession of pronounced resonances and anti-resonances. In the vicinity of an anti-resonance the electrical equivalent of the vibrating wire may be approximated by a coil and condenser in parallel as shown by L_2 and C_2 of Fig. 14.10. This acts in series with the mechanical resonance of the quartz plate, represented by L_1 and C_1 . The impedance curves illustrate the effect of the wire resonance on the crystal impedance. R_1 , the equivalent resistance of the crystal plate, is constant for frequencies in the vicinity of resonance. X_1 , the equivalent reactance of the crystal plate, increases rapidly as the frequency departs from resonance. R_2 and X_2 , the equivalent resistance and reactance of the wire resonance, are typical of an

anti-resonant electrical network. The curve labeled $(X_1 + X_2)$ shows the effect of the wire resonance on the response of the crystal plate. It may be observed that the apparent resonance has been reduced by a small frequency decrement. The amount of frequency shift and the increase in effective resistance depend on the Q of the wire resonance, its frequency location compared with the resonance of the crystal plate, the mass of the wire relative to that of the quartz plate, and the distance from the node to the point at which the wire is actually fastened to the plate.

The slope of the frequency-reactance characteristic corresponding to the mechanical resonance of the quartz plate is very steep and the effect of the wire resonance will be noticed only when an anti-resonant frequency of the wire is close to the resonant frequency of the plate. The changes in resonant frequency and effective resistance due to wire resonance have been measured for some filter crystal units and the measurements are tabulated in Table II.

TABLE II
EFFECT OF WIRE VIBRATIONS ON THE RESISTANCE OF A QUARTZ CRYSTAL PLATE

Crystal Type	Mode of Vibration for Crystal	Resonant Frequency	Crystal Mass in Grams	Distance of Wire from Nodal Line	Maximum Frequency Shift CPS	Maximum Increase in Resistance
+5° X-Cut	Flexural	12 kc	.51	(N) .060"	±2.0	250%
+5° X-Cut	Longitudinal	164 kc	.142	(N) .012"	±30	640%
-18° X-Cut	Longitudinal	335 kc	.075	(M) .002"	±90	360%
-18° X-Cut	Longitudinal	552 kc	.068	(N) 0.0"	±75	1100%
5th Harmonic GT	Longitudinal	164 kc	.98	(N) .011"	±12	370%

(N) Specified Dimension.

(M) Measured Dimension.

The relation between the length of a wire and the frequencies at which it will resonate in flexural modes is expressed by the following equation:

$$l = m \sqrt{\frac{vd}{8\pi f}}$$

where v is the velocity of sound in the wire

d is the diameter of the wire

l is the length of the wire

f is the frequency of wire resonance in cycles per second

m is a number that depends on the manner in which the ends of the wire can move.

At a particular frequency and for wire of a particular material and diameter there is a series of critical wire lengths which must be avoided. The critical lengths are those which cause the wire to present a high impedance to the motion of the plate. This high impedance may be considered, from the electrical point of view, as corresponding to an anti-resonance of the wire.

The critical lengths are defined by the series of numbers $m \doteq (n + \frac{1}{2}) \pi$ where n takes the values 1, 2, 3, etc. and apply to successive modes of a bar clamped at both ends. Beyond the first mode, the critical wire lengths are spaced at equal intervals, corresponding to increments of m each equal to π .

There is also a series of wire lengths which will present minimum impedance to the motion of the plate. These may be considered as corresponding to a resonance of the wire. These minima of impedance are obtained for lengths of wire defined by the series of numbers $m \doteq (n - \frac{1}{4}) \pi$. They apply to a bar which is clamped at one end and, while free to vibrate at the other end, is constrained to a slope perpendicular to the plate.

In selecting a desirable length for the supporting wire, it is not essential that this length be such as to cause the wire to present minimum impedance to motion of the plate. As a matter of fact, since the wire is of relatively low characteristic impedance a small departure from the critical length is sufficient to avoid trouble from wire resonance. In order to allow for as wide a manufacturing tolerance as possible the supporting wire is usually designed to have a length half-way between two successive critical lengths. For a 6.3-mil phosphor-bronze wire, the spacing between successive critical lengths ranges from about 58 mils at 100 kc to about 15 mils at 1000 kc. Hence, even at 100 kc the length of the supporting wire must be controlled within a tolerance of about 20 mils.

These supporting wires are formed to have definite bends along their length and the location of these bends varies slightly from one wire to another. In addition, the wires are terminated by solder at both ends. Because of these complications it is impractical to meet such close tolerances on the effective length of the wires. Furthermore, a wire that does have a suitable effective length at room temperatures may exhibit sufficient change of properties with variations in temperature so that it becomes of critical length at some other operating temperature.

Much of the difficulty due to wire resonance is avoided by use of a solder ball on the supporting wire, as described in Chapters VIII and XIII. The solder ball is located near the quartz plate. Since it serves as a clamp at that point, it makes the supporting wire short. By locating and forming the solder ball accurately, the length of the supporting wire is controlled within a close tolerance. Further, since the wire is shortened by use of the ball it is less affected by changes in temperature. Experience at about 500 kc indicates that a tolerance of about 10 mils in locating the solder ball is practical and has provided satisfactory operation between -40 C and $+85$ C.

14.4 NEED FOR CLEANLINESS AND LOW RELATIVE HUMIDITY

One of the most serious difficulties encountered in manufacturing quartz crystal plates is that of assuring sufficient cleanliness. Even minute par-

ticles of foreign matter will introduce appreciable changes in crystal performance.

Usually, the presence of foreign matter will act to load the crystal and will reduce the resonant frequency but there are also instances where the added matter tends to stiffen the plate and increase its frequency. The latter has been observed to occur as the result of the deposit of a thin film of rosin on the surface of the plate. In the presence of foreign matter on the surface of the plates, the performance will be unstable with time and temperature even after the plate is sealed into a container. Also, erratic variations are observed as the plate is shifted from a normal atmosphere to a container which is evacuated or filled with dry air. Experience has shown

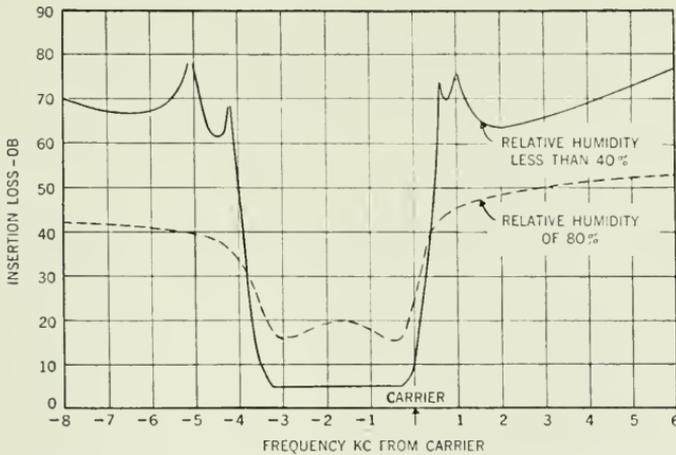


Fig. 14.11.—Effect of humidity on the discrimination of channel crystal filters. To prevent decrease of discrimination with increase of relative humidity, the crystal units must be hermetically sealed.

that elaborate precautions for insuring cleanliness are justified by the time saved in the adjusting processes.

The need for cleanliness is closely related to the effect of humidity on the insulation resistance of crystal units. As used in filters, crystal units must provide extremely high impedances at their anti-resonant frequencies. These impedances may be as high as 100 megohms. With clean crystal plates in relatively dry atmospheres, such insulation resistance can be maintained up to 1000 kc. However, even a trace of salts or other types of contamination will make the insulation resistance highly sensitive to moisture in the adjacent air. While it is relatively difficult to measure insulation resistance at high carrier frequencies, the effect of the reduced insulation due to moisture is evident on inspection of the discrimination characteristics of the filters. For example, Fig. 14.11 illustrates the result of high relative

humidity on the transmission characteristic of a typical crystal filter. It may be observed that the discrimination almost disappears for a relative humidity of 80%. These measurements were made on a filter containing well cleaned crystal plates. It will be found frequently that an unsatisfactory discrimination characteristic is produced by considerably lower values of relative humidity when the plates are not so clean. Experience has shown that it is impractical to let the relative humidity surrounding the crystal plate exceed 40% for satisfactory filter performance. When a high degree of accuracy is required, the plates are assembled

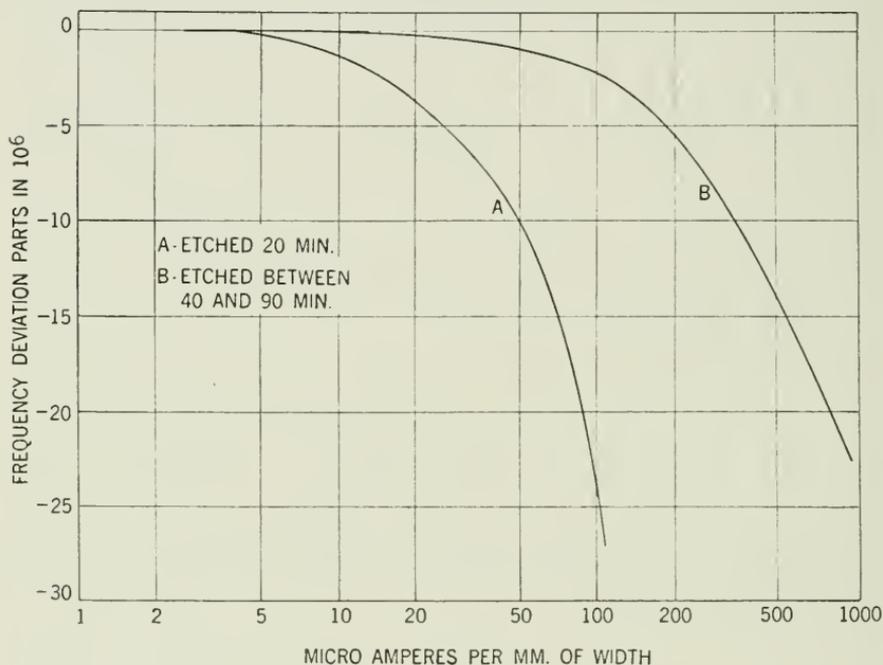


Fig. 14.12.—Change of resonant frequency of GT-cut plates due to increase of transmitted current.

in a unit which is either evacuated or filled with air at a relative humidity of less than 5%.

14.5 EFFECT OF CURRENT LEVEL

Crystal units will undergo change in effective resistance and in frequency of resonance as the current transmitted is increased. Some change might be expected due to the heating of the plate by the dissipative loss associated with the transmission of current. However, the effects are not identical with those obtained with a change in ambient temperature. Appreciable

changes have been observed even when using GT-cut plates adjusted to zero temperature coefficient as shown, for example, in the curves of Fig. 14.12. Also, it has been observed that after a plate has been driven hard and the transmitted current then reduced, the original resonant frequency is restored only after a considerable time interval. The data of Fig. 14.12 provides a rough correlation between stability and current levels. For example, if the stability desired for a crystal unit using a GT type plate be in the order of one part per million, the circuit design should be such as to keep the current level in the plate below about 10 microamperes per millimeter of width.

The parameter used in these paragraphs for measuring current levels is the current per unit of width. This appears to be useful as a common basis for comparing various plates of any one cut and mode of vibration. Theoretically, in the case of a plate vibrating longitudinally, the current, I , per unit of width, w , is directly proportional to the elongation per unit of length, y_y , as shown by following equation:

$$I/w = K y_y$$

where K is a constant which depends on the cut of plate and mode.

Figure 14.12 also illustrates the importance of the surface condition of the plates. Curve A is the average frequency-current characteristic for a group of crystal units using plates etched for twenty minutes in 47% hydrofluoric acid and curve B the average characteristic for a group of crystal units using plates etched for over forty minutes but less than ninety minutes. Evidently, crystal units using plates which have been etched for a long period exhibit a frequency-current characteristic which is appreciably more constant than those using plates etched for a shorter period.

Mathematical Analysis of Random Noise

By S. O. RICE

INTRODUCTION

THIS paper deals with the mathematical analysis of noise obtained by passing random noise through physical devices. The random noise considered is that which arises from shot effect in vacuum tubes or from thermal agitation of electrons in resistors. Our main interest is in the statistical properties of such noise and we leave to one side many physical results of which Nyquist's law may be given as an example.¹

About half of the work given here is believed to be new, the bulk of the new results appearing in Parts III and IV. In order to provide a suitable introduction to these results and also to bring out their relation to the work of others, this paper is written as an exposition of the subject indicated in the title.

When a broad band of random noise is applied to some physical device, such as an electrical network, the statistical properties of the output are often of interest. For example, when the noise is due to shot effect, its mean and standard deviations are given by Campbell's theorem (Part I) when the physical device is linear. Additional information of this sort is given by the (auto) correlation function which is a rough measure of the dependence of values of the output separated by a fixed time interval.

The paper consists of four main parts. The first part is concerned with shot effect. The shot effect is important not only in its own right but also because it is a typical source of noise. The Fourier series representation of a noise current, which is used extensively in the following parts, may be obtained from the relatively simple concepts inherent in the shot effect.

The second part is devoted principally to the fundamental result that the power spectrum of a noise current is the Fourier transform of its correlation function. This result is used again and again in Parts III and IV.

A rather thorough discussion of the statistics of random noise currents is given in Part III. Probability distributions associated with the maxima of the current and the maxima of its envelope are developed. Formulas for the expected number of zeros and maxima per second are given, and a start is made towards obtaining the probability distribution of the zeros.

When a noise voltage or a noise voltage plus a signal is applied to a non-

¹ An account of this field is given by E. B. Moullin, "Spontaneous Fluctuations of Voltage," Oxford (1938).

linear device, such as a square-law or linear rectifier, the output will also contain noise. The methods which are available for computing the amount of noise and its spectral distribution are discussed in Part IV.

ACKNOWLEDGEMENT

I wish to thank my friends for many helpful suggestions and discussions regarding the subject of this paper. Although it has been convenient to acknowledge some of this assistance in the text, I appreciate no less sincerely the considerable amount which is not mentioned. In particular, I am indebted to Miss Darville for computing the curves in Parts III and IV.

SUMMARY OF RESULTS

Before proceeding to the main body of the paper, we shall state some of the principal results. It is hoped that this summary will give the casual reader an over-all view of the material covered and at the same time guide the reader who is interested in obtaining some particular item of information to those portions of the paper which may possibly contain it.

Part I—Shot Effect

Shot effect noise results from the superposition of a great number of disturbances which occur at random. A large class of noise generators produce noise in this way.

Suppose that the arrival of an electron at the anode of the vacuum tube at time $t = 0$ produces an effect $F(t)$ at some point in the output circuit. If the output circuit is such that the effects of the various electrons add linearly, the total effect at time t due to all the electrons is

$$I(t) = \sum_{k=-\infty}^{+\infty} F(t - t_k) \tag{1.2-1}$$

where the k^{th} electron arrives at t_k and the series is assumed to converge. Although the terminology suggests that $I(t)$ is a current, and it will be spoken of as a noise current, it may be any quantity expressible in the form (1.2-1).

1. Campbell's theorem: The average value of $I(t)$ is

$$\overline{I(t)} = \nu \int_{-\infty}^{+\infty} F(t) dt \tag{1.2-2}$$

and the mean square value of the fluctuation about this average is

$$\text{ave. } [I(t) - \overline{I(t)}]^2 = \nu \int_{-\infty}^{+\infty} F^2(t) dt \tag{1.2-3}$$

where ν is the average number of electrons arriving per second at the anode. In this expression the electrons are supposed to arrive independently and at random. $\nu e^{-\nu t} dt$ is the probability that the length of the interval between two successive arrivals lies between t and $t + dt$.

2. Generalization of Campbell's theorem. Campbell's theorem gives information about the average value and the standard deviation of the probability distribution of $I(t)$. A generalization of the theorem gives information about the third and higher order moments. Let

$$I(t) = \sum_{-\infty}^{+\infty} a_k F(t - t_k) \quad (1.5-1)$$

where $F(t)$ and t_k are of the same nature as those in (1.2-1) and $\dots a_1, a_2, \dots a_k, \dots$ are independent random variables all having the same distribution. Then the n^{th} semi-invariant of the probability density $P(I)$ of $I = I(t)$ is

$$\lambda_n = \nu \bar{a}^n \int_{-\infty}^{+\infty} [F(t)]^n dt \quad (1.5-2)$$

The semi-invariants are defined as the coefficients in the expansion of the characteristic function $f(u)$:

$$\log_e f(u) = \sum_{n=1}^{\infty} \frac{\lambda_n}{n!} (iu)^n \quad (1.5-3)$$

where

$$f(u) = \text{ave. } e^{iIu} = \int_{-\infty}^{+\infty} P(I) e^{iIu} dI$$

The moments may be computed from the λ 's.

3. As $\nu \rightarrow \infty$ the probability density $P(I)$ of the shot effect current approaches a normal law. The way it is approached is given by

$$P(I) \sim \sigma^{-1} \varphi^{(0)}(x) - \frac{\lambda_3 \sigma^{-4}}{3!} \varphi^{(3)}(x) + \left[\frac{\lambda_4 \sigma^{-5}}{4!} \varphi^{(4)}(x) + \frac{\lambda_3^2 \sigma^{-7}}{72} \varphi^{(6)}(x) \right] + \dots \quad (1.6-3)$$

where the λ 's are given by (1.5-2) and

$$\sigma^2 = \lambda_2 \quad x = \frac{I - \bar{I}}{\sigma} \quad \varphi^{(n)}(x) = \frac{1}{\sqrt{2\pi}} \frac{d^n}{dx^n} e^{-x^2/2}$$

Since the λ 's are of the order of ν , σ is of the order of $\nu^{1/2}$ and the orders of σ^{-1} , $\lambda_3 \sigma^{-4}$, $\lambda_4 \sigma^{-5}$ and $\lambda_3^2 \sigma^{-7}$ are $\nu^{-1/2}$, ν^{-1} , $\nu^{-3/2}$ and $\nu^{-3/2}$ respectively. A

possible use of this result is to determine whether a noise due to random independent events occurring at the rate of ν per second may be regarded as "random noise" in the sense of this work.

4. When $I(t)$, as given by (1.5-1), is analyzed as a Fourier series over an interval of length T a set of Fourier coefficients is obtained. By taking many different intervals, all of length T , many sets of coefficients are obtained. If ν is sufficiently large these coefficients tend to be distributed normally and independently. A discussion of this is given in section 1.7.

Part II—Power Spectra and Correlation Functions

1. Suppose we have a curve, such as an oscillogram of a noise current, which extends from $t = 0$ to $t = \infty$. Let this curve be denoted by $I(t)$. The correlation function of $I(t)$ is $\psi(\tau)$ which is defined as

$$\psi(\tau) = \text{Limit}_{T \rightarrow \infty} \frac{1}{T} \int_0^T I(t)I(t + \tau) dt \tag{2.1-4}$$

where the limit is assumed to exist. This function is closely connected with another function, the power spectrum, $w(f)$, of $I(t)$. $I(t)$ may be regarded as composed of many sinusoidal components. If $I(t)$ were a noise current and if it were to flow through a resistance of one ohm the average power dissipated by those components whose frequencies lie between f and $f + df$ would be $w(f) df$.

The relation between $w(f)$ and $\psi(\tau)$ is

$$w(f) = 4 \int_0^\infty \psi(\tau) \cos 2\pi f\tau d\tau \tag{2.1-5}$$

$$\psi(\tau) = \int_0^\infty w(f) \cos 2\pi f\tau df \tag{2.1-6}$$

When $I(t)$ has no d.c. or periodic components,

$$w(f) = \text{Limit}_{T \rightarrow \infty} \frac{2 |S(f)|^2}{T} \tag{2.1-3}$$

where

$$S(f) = \int_0^T I(t)e^{-2\pi ift} dt.$$

The correlation function for

$$I(t) = A + C \cos (2\pi f_0 t - \varphi)$$

is

$$\psi(\tau) = A^2 + \frac{C^2}{2} \cos 2\pi f_0 \tau \tag{2.2-3}$$

These results are discussed in sections 2.1 to 2.4 inclusive.

2. So far we have supposed $I(t)$ to be some definite function for which a curve may be drawn. Now consider $I(t)$ to be given by a mathematical expression into which, besides t , a number of parameters enter. $w(f)$ and $\psi(\tau)$ are now obtained by averaging the integrals over the possible values of the parameters. This is discussed in section 2.5.

3. The correlation function for the shot effect current of (1.2-1) is

$$\psi(\tau) = \nu \int_{-\infty}^{+\infty} F(t)F(t + \tau) dt + \left[\nu \int_{-\infty}^{+\infty} F(t) dt \right]^2 \quad (2.6-2)$$

The distributed portion of the power spectrum is

$$w_1(f) = 2\nu |s(f)|^2$$

where

$$s(f) = \int_{-\infty}^{+\infty} F(t)e^{-2\pi ift} dt \quad (2.6-5)$$

The complete power spectrum has in addition to $w_1(f)$ an impulse function representing the d.c. component $\bar{I}(t)$.

In the formulas above for the shot effect it was assumed that the expected number, ν , of electrons per second did not vary with time. A case in which ν does vary with time is briefly discussed near the end of Section 2.6.

4. Random telegraph signal. Let $I(t)$ be equal to either a or $-a$ so that it is of the form of a flat top wave, and let the lengths of the tops and bottoms be distributed independently and exponentially. The correlation function and power spectrum of I are

$$\psi(\tau) = a^2 e^{-2\mu|\tau|} \quad (2.7-4)$$

$$w(f) = \frac{2a^2 \mu}{\pi^2 f^2 + \mu^2} \quad (2.7-5)$$

where μ is the expected number of changes of sign per second.

Another type of random telegraph signal may be formed as follows: Divide the time scale into intervals of equal length h . In an interval selected at random the value of $I(t)$ is independent of the value in the other intervals and is equally likely to be $+a$ or $-a$. The correlation function of $I(t)$ is zero for $|\tau| > h$ and is

$$a^2 \left(1 - \frac{|\tau|}{h} \right)$$

for $0 \leq |\tau| < h$ and the power spectrum is

$$w(f) = 2h \left(\frac{a \sin \pi fh}{\pi fh} \right)^2 \quad (2.7-9)$$

5. There are two representations of a random noise current which are especially useful. The first one is

$$I(t) = \sum_{n=1}^N (a_n \cos \omega_n t + b_n \sin \omega_n t) \tag{2.8-1}$$

where a_n and b_n are independent random variables which are distributed normally about zero with the standard deviation $\sqrt{w(f_n)\Delta f}$ and where

$$\omega_n = 2\pi f_n, \quad f_n = n\Delta f$$

The second one is

$$I(t) = \sum_{n=1}^N c_n \cos (\omega_n t - \varphi_n) \tag{2.8-6}$$

where φ_n is a random phase angle distributed uniformly over the range $(0, 2\pi)$ and

$$c_n = [2w(f_n)\Delta f]^{1/2}$$

At an appropriate point in the analysis N and Δf are made to approach infinity and zero, respectively, in such a manner that the entire frequency band is covered by the summations (which then become integrations).

6. The normal distribution in several variables and the central limit theorem are discussed in sections 2.9 and 2.10.

Part III—Statistical Properties of Noise Current

1. The noise current is distributed normally. This has already been discussed in section 1.6 for the shot-effect. It is discussed again in section 3.1 using the concepts introduced in Part II, and the assumption, used throughout Part III, that the average value of the noise current $I(t)$ is zero. The probability that $I(t)$ lies between I and $I + dI$ is

$$\frac{dI}{\sqrt{2\pi\psi_0}} e^{-I^2/2\psi_0} \tag{3.1-3}$$

where ψ_0 is the value of the correlation function, $\psi(\tau)$, of $I(t)$ at $\tau = 0$

$$\psi_0 = \psi(0) = \int_0^\infty w(f) df, \tag{3.1-2}$$

$w(f)$ being the power spectrum of $I(t)$. ψ_0 is the mean square value of $I(t)$, i.e., the r.m.s. value of $I(t)$ is $\psi_0^{1/2}$.

The characteristic function (ch. f.) of this distribution is

$$\text{ave. } e^{iuI(t)} = \exp - \frac{\psi_0}{2} u^2 \tag{3.1-6}$$

2. The probability that $I(t)$ lies between I_1 and $I_1 + dI$, and $I(t + \tau)$ lies between I_2 and $I_2 + dI_2$ when t is chosen at random is

$$[\psi_0^2 - \psi_\tau^2]^{-1/2} \frac{dI_1 dI_2}{2\pi} \exp \left[\frac{-\psi_0 I_1^2 - \psi_0 I_2^2 + 2\psi_\tau I_1 I_2}{2(\psi_0^2 - \psi_\tau^2)} \right] \quad (3.2-4)$$

where ψ_τ is the correlation function $\psi(\tau)$ of $I(t)$:

$$\psi(\tau) = \int_0^\infty w(f) \cos 2\pi f\tau df \quad (3.2-3)$$

The ch. f. for this distribution is

$$\text{ave. } e^{iuI(t)+ivI(t+\tau)} = \exp \left[-\frac{\psi_0}{2} (u^2 + v^2) - \psi_\tau uv \right] \quad (3.2-7)$$

3. The expected number of zeros per second of $I(t)$ is

$$\frac{1}{\pi} \left[-\frac{\psi''(0)}{\psi(0)} \right]^{1/2} = 2 \left[\frac{\int_0^\infty f^2 w(f) df}{\int_0^\infty w(f) df} \right]^{1/2} \quad (3.3-11)$$

assuming convergence of the integrals. The primes denote differentiation with respect to τ :

$$\psi''(\tau) = \frac{d^2}{d\tau^2} \psi(\tau).$$

For an ideal band-pass filter whose pass band extends from f_a to f_b the expected number of zeros per second is

$$2 \left[\frac{1}{3} \frac{f_b^3 - f_a^3}{f_b - f_a} \right]^{1/2} \quad (3.3-12)$$

When f_a is zero this becomes $1.155 f_b$ and when f_a is very nearly equal to f_b it approaches $f_b + f_a$.

4. The problem of determining the distribution function for the length of the interval between two successive zeros of $I(t)$ seems to be quite difficult. In section 3.4 some related results are given which lead, in some circumstances, to approximations to the distribution. For example, for an ideal narrow band-pass filter the probability that the distance between two successive zeros lies between τ and $\tau + d\tau$ is approximately

$$\frac{d\tau}{2} \frac{a}{[1 + a^2(\tau - \tau_1)^2]^{3/2}}$$

where

$$a = \sqrt{3} \frac{(f_b + f_a)^2}{f_b - f_a}, \quad \tau_1 = \frac{1}{f_b + f_a}$$

f_b and f_a being the upper and lower cut-off frequencies.

5. In section 3.5 several multiple integrals which occur in the work of Part III are discussed.

6. The distribution of the maxima of $I(t)$ is discussed in section 3.6. The expected number of maxima per second is

$$\frac{1}{2\pi} \left[\frac{\psi_0^{(4)}}{\psi_0'^2} \right]^{1/2} = \left[\frac{\int_0^\infty f^4 \omega(f) df}{\int_0^\infty f^2 \omega(f) df} \right]^{1/2} \tag{3.6-6}$$

For a band-pass filter the expected number of maxima per second is

$$\left[\frac{3 f_b^5 - f_a^5}{5 f_b^3 - f_a^3} \right]^{1/2} \tag{3.6-7}$$

For a low-pass filter where $f_a = 0$ this number is $0.775 f_b$.

The expected number of maxima per second lying above the line $I(t) = I_1$ is approximately, when I_1 is large,

$$e^{-I_1^2/2\psi_0} \times \frac{1}{2} [\text{the expected number of zeros of } I \text{ per second}] \tag{3.6-11}$$

where ψ_0 is the mean square value of $I(t)$.

For a low-pass filter the probability that a maximum chosen at random from the universe of maxima lies between I and $I + dI$ is approximately, when I is large,

$$\frac{\sqrt{5}}{3} y e^{-y^2/2} \frac{dI}{\psi_0^{1/2}} \tag{3.6-9}$$

where

$$y = \frac{I}{\psi_0^{1/2}}$$

7. When we pass noise through a relatively narrow band-pass filter one of the most noticeable features of an oscillogram of the output current is its fluctuating envelope. In sections 3.7 and 3.8 some statistical properties of this envelope, denoted by R or $R(t)$, are derived.

The probability that the envelope lies between R and $R + dR$ is

$$\frac{R}{\psi_0} e^{-R^2/2\psi_0} dR \tag{3.7-10}$$

where ψ_0 is the mean square value of $I(t)$. The probability that $R(t)$ lies between R_1 and $R_1 + dR_1$ and at the same time $R(t + \tau)$ lies between R_2 and $R_2 + dR_2$ when t is chosen at random is obtained by multiplying (3.7-13) by $dR_1 dR_2$. For an ideal band-pass filter, the expected number of maxima of the envelope in one second is

$$.64110(f_b - f_a) \quad (3.8-15)$$

When R is large, say $y > 2.5$ where

$$y = \frac{R}{\psi_0^{1/2}}, \quad \psi_0^{1/2} = \text{r.m.s. value of } I(t),$$

the probability that a maximum of the envelope, selected at random from the universe of such maxima, lies between R and $R + dR$ is approximately

$$1.13(y^2 - 1)e^{-y^2/2} \frac{dR}{\psi_0^{1/2}}$$

A curve for the corresponding probability density is shown for the range $0 \leq y \leq 4$. Curves which compare the distribution function of the maxima of R with other distribution functions of the same type are also given.

8. In section 3.9 some information is given regarding the statistical behavior of the random variable:

$$E = \int_{t_1}^{t_1 + T} I^2(t) dt \quad (3.9-1)$$

where t_1 is chosen at random and $I(t)$ is a noise current with the power spectrum $w(f)$ and the correlation function $\psi(\tau)$. The average value m_T of E is $T\psi_0$ and its standard deviation σ_T is given by (3.9-9). For a relatively narrow band-pass filter

$$\frac{\sigma_T}{m_T} \sim \frac{1}{\sqrt{T(f_b - f_a)}}$$

when $T(f_b - f_a) \gg 1$. This follows from equation (3.9-10). An expression which is believed to approximate the distribution of E is given by (3.9-20).

9. In section 3.10 the distribution of a noise current plus one or more sinusoidal currents is discussed. For example, if I consists of two sine waves plus noise:

$$I = P \cos pt + Q \cos qt + I_N, \quad (3.10-20)$$

where p and q are incommensurable and the r.m.s. value of the noise current I_N is $\psi_0^{1/2}$, the probability density of the envelope R is

$$R \int_0^\infty r J_0(Rr) J_0(Pr) J_0(Qr) e^{-\psi_0 r^2/2} dr \quad (3.10-21)$$

where $J_0(\)$ is a Bessel function.

Curves showing the probability density and distribution function of R , when $Q = 0$, for various ratios of P /r.m.s. I_N are given.

10. In section 3.11 it is pointed out that the representations (2.8-1) and (2.8-6) of the noise current as the sum of a great number of sinusoidal components are not the only ones which may be used in deriving the results given in the preceding sections of Part III. The shot effect representation

$$I(t) = \sum_{-\infty}^{+\infty} F(t - t_k)$$

studied in Part I may also be used.

Part IV—Noise Through Non-Linear Devices

1. Suppose that the power spectrum of the voltage V applied to the square-law device

$$I = \alpha V^2 \tag{4.1-1}$$

is confined to a relatively narrow band. The total low-frequency output current I_{ℓ} may be expressed as the sum

$$I_{\ell} = I_{dc} + I_{\ell f} \tag{4.1-2}$$

where I_{dc} is the d.c. component and $I_{\ell f}$ is the variable component. When none of the low-frequency band is eliminated (by audio frequency filters)

$$I_{\ell} = \frac{\alpha R^2}{2} \tag{4.1-6}$$

where R is the envelope of V . If V is of the form

$$V = V_N + P \cos pt + Q \cos qt, \tag{4.1-4}$$

where V_N is a noise voltage whose mean square value is ψ_0 , then

$$I_{dc} = \alpha \left(\psi_0 + \frac{P^2}{2} + \frac{Q^2}{2} \right)$$

$$\overline{I_{\ell f}^2} = \alpha^2 \left[\psi_0^2 + P^2 \psi_0 + Q^2 \psi_0 + \frac{P^2 Q^2}{2} \right] \tag{4.1-16}$$

2. If instead of a square-law device we have a linear rectifier,

$$I = \begin{cases} 0 & V < 0 \\ \alpha V, & V > 0 \end{cases} \tag{4.2-1}$$

the total low-frequency output is

$$I_{\ell} = \frac{\alpha R}{\pi} \tag{4.2-2}$$

When V is a sine wave plus noise, $V_N + P \cos pt$,

$$I_{dc} = \alpha \left(\frac{\psi_0}{2\pi} \right)^{1/2} {}_1F_1\left(-\frac{1}{2}; 1; -x\right) \quad (4.2-3)$$

$$\overline{I_{i\ell}^2} = \frac{\alpha^2}{\pi^2} (P^2 + 2\psi_0) \quad (4.2-6)$$

where ${}_1F_1$ is a hypergeometric function and

$$x = \frac{P^2}{2\psi_0} = \frac{\text{Ave. sine wave power}}{\text{Ave. noise power}} \quad (4.2-4)$$

When x is large

$$\overline{I_{i\ell}^2} \sim \frac{\alpha^2 \psi_0}{\pi^2} \left[1 - \frac{1}{4x} \dots \right] \quad (4.2-7)$$

If V consists of two sine waves plus noise, I_{dc} consists of a hypergeometric function of two variables. The equations running from (4.2-9) to (4.2-15) are concerned with this case. About the only simple equation is

$$\overline{I_{i\ell}^2} = \frac{\alpha^2}{\pi^2} [2\psi_0 + P^2 + Q^2] \quad (4.2-14)$$

3. The expressions (4.1-6) and (4.2-2) for $I_{i\ell}$ in terms of the envelope R of V , namely

$$\frac{\alpha R^2}{2} \quad \text{and} \quad \frac{\alpha R}{\pi},$$

are special cases of a more general result

$$I_{i\ell} = A_0(R) = \frac{1}{2\pi} \int_C F(iu) J_0(uR) du. \quad (4.3-11)$$

In this expression $J_0(uR)$ is a Bessel function. The path of integration C and the function $F(iu)$ are chosen so that the relation between I and V may be expressed as

$$I = \frac{1}{2\pi} \int_C F(iu) e^{i^v u} du. \quad (4A-1)$$

A table giving $F(iu)$ and C for a number of common non-linear devices is shown in Appendix 4A.

If this relation is used to study the biased linear rectifier.

$$I = \begin{cases} 0, & V < B \\ V - B, & V > B \end{cases}$$

for the case in which V is $V_N + P \cos pt$, we find

$$I_{dc} \sim -\frac{B}{2} + \frac{P}{\pi} + \frac{B^2 + \psi_0}{2\pi P}$$

$$\overline{I_{\ell}^2} \sim \frac{P^2 - B^2}{\pi^2 P^2} \psi_0 \tag{4.3-17}$$

when $P \gg |B|$, $P^2 \gg \psi_0$ where ψ_0 is the mean square value of V_N .

4. When V is confined to a relatively narrow band and there are no audio-frequency filters, the probability density and all the associated statistical properties of I_{ℓ} may be obtained by expressing I_{ℓ} as a function of the envelope R of V and then using the probability density of R . When V is $V_N + P \cos pt + Q \cos qt$ this probability density is given by the integral, (3.10-21) (which is the integral containing three Bessel functions stated in the above summary of Part III). When V consists of three sine waves plus noise there are four J_0 's in the integrand, and so on. Expressions for $\overline{R^n}$ when R has the above distribution are given by equations (3.10-25) and (3.10-27).

When audio-frequency filters remove part of the low-frequency band the statistical properties, except the mean square value, of the resulting current are hard to compute. In section 4.3 it is shown that as the output band is chosen narrower and narrower, the statistical properties of the output current approach those of a random noise current.

5. The sections in Part IV from 4.4 onward are concerned with the problem: Given a non-linear device and an input voltage consisting of noise alone or of a signal plus noise. What is the power spectrum of the output? A survey of the methods available for the solution of this problem is given in section 4.4.

6. When a noise voltage V_N with the power spectrum $w(f)$ is applied to the square-law device

$$I = \alpha V^2 \tag{4.1-1}$$

the power spectrum of the output current I is, when $f \neq 0$,

$$W(f) = \alpha^2 \int_{-\infty}^{+\infty} w(x)w(f-x) dx \tag{4.5-5}$$

where $w(-x)$ is defined to equal $w(x)$. The power spectrum of I when V is either $P \cos pt + V_N$ or

$$Q(1 + k \cos pt) \cos qt + V_N$$

is considered in the portion of section 4.5 containing equations (4.5-10) to (4.5-17).

7. A method discovered independently by Van Vleck and North shows that the correlation function $\Psi(\tau)$ of the output current for an unbiased linear rectifier is

$$\Psi(\tau) = \frac{\psi_\tau}{4} + \frac{\psi_0}{2} {}_2F_1 \left[-\frac{1}{2}, -\frac{1}{2}; \frac{1}{2}; \frac{\psi_\tau^2}{\psi_0^2} \right] \quad (4.7-6)$$

where the input voltage is V_N . The correlation function $\psi(\tau)$ of V_N is denoted by ψ_τ and the mean square value of V_N is ψ_0 . The power spectrum $W(f)$ of I may be obtained from

$$W(f) = 4 \int_0^\infty \Psi(\tau) \cos 2\pi f\tau \, d\tau \quad (4.6-1)$$

by expanding the hypergeometric function and integrating termwise using

$$G_n(f) = \int_0^\infty \psi_\tau^n \cos 2\pi f\tau \, d\tau. \quad (4C-1)$$

Appendix 4C is devoted to the problem of evaluating the integral for $G_n(f)$.

8. Another method of obtaining the correlation function $\psi(\tau)$ of I , termed the "characteristic function method," is explained in section 4.8. It is illustrated in section 4.9 where formulas for $\Psi(\tau)$ and $W(f)$ are developed when the voltage $P \cos pt + V_N$ is applied to a general non-linear device.

9. Several miscellaneous results are given in section 4.10. The characteristic function method is used to obtain the correlation function for a square-law device. The general formulas of section 4.9 are applied to the case of a ν^{th} law rectifier when the input noise spectrum has a normal law distribution. Some remarks are also made concerning the audio-frequency output of a linear rectifier when the input voltage V is

$$Q(1 + r \cos pt) \cos qt + V_N.$$

10. A discussion of the hypergeometric function ${}_1F_1(a; c; x)$, which often occurs in problems concerning a sine wave plus noise, is given in Appendix 4B.

PART I

THE SHOT EFFECT

The shot effect in vacuum tubes is a typical example of noise. It is due to fluctuations in the intensity of the stream of electrons flowing from the cathode to the anode. Here we analyze a simplified form of the shot effect.

1.1 THE PROBABILITY OF EXACTLY K ELECTRONS ARRIVING AT THE ANODE IN TIME T

The fluctuations in the electron stream are supposed to be random. We shall treat this randomness as follows. We count the number of electrons flowing in a long interval of time T measured in seconds. Suppose there are K_1 . Repeating this counting process for many intervals all of length T gives a set of numbers $K_2, K_3 \cdots K_M$ where M is the total number of intervals. The average number ν , of electrons per second is defined as

$$\nu = \lim_{M \rightarrow \infty} \frac{K_1 + K_2 + \cdots + K_M}{MT} \tag{1.1-1}$$

where we assume that this limit exists. As M is increased with T being held fixed some of the K 's will have the same value. In fact, as M increases the number of K 's having any particular value will tend to increase. This of course is based on the assumption that the electron stream is a steady flow upon which random fluctuations are superposed. The probability of getting K electrons in a given trial is defined as

$$p(K) = \lim_{M \rightarrow \infty} \frac{\text{Number of trials giving exactly } K \text{ electrons}}{M} \tag{1.1-2}$$

Of course $p(K)$ also depends upon T . We assume that the randomness of the electron stream is such that the probability that an electron will arrive at the anode in the interval $(t, t + \Delta t)$ is $\nu \Delta t$ where Δt is such that $\nu \Delta t \ll 1$, and that this probability is independent of what has happened before time t or will happen after time $t + \Delta t$.

This assumption is sufficient to determine the expression for $p(K)$ which is

$$p(K) = \frac{(\nu T)^K}{K!} e^{-\nu T} \tag{1.1-3}$$

This is the "law of small probabilities" given by Poisson. One method of derivation sometimes used can be readily illustrated for the case $K = 0$. Thus, divide the interval, $(0, T)$ into M intervals each of length $\Delta t = \frac{T}{M}$. Δt is taken so small that $\nu \Delta t$ is much less than unity. (This is the "small probability" that an electron will arrive in the interval Δt). The probability that an electron will not arrive in the first sub-interval is $(1 - \nu \Delta t)$. The probability that one will not arrive in either the first or the second sub-interval is $(1 - \nu \Delta t)^2$. The probability that an electron will not arrive in any of the M intervals is $(1 - \nu \Delta t)^M$. Replacing M by $T/\Delta t$ and letting $\Delta t \rightarrow 0$ gives

$$p(0) = e^{-\nu T}$$

The expressions for $p(1), p(2), \dots, p(K)$ may be derived in a somewhat similar fashion.

1.2 STATEMENT OF CAMPBELL'S THEOREM

Suppose that the arrival of an electron at the anode at time $t = 0$ produces an effect $F(t)$ at some point in the output circuit. If the output circuit is such that the effects of the various electrons add linearly, the total effect at time t due to all the electrons is

$$I(t) = \sum_{k=-\infty}^{+\infty} F(t - t_k) \quad (1.2-1)$$

where the k^{th} electron arrives at t_k and the series is assumed to converge.

Campbell's theorem² states that the average value of $I(t)$ is

$$\overline{I(t)} = \nu \int_{-\infty}^{+\infty} F(t) dt \quad (1.2-2)$$

and the mean square value of the fluctuation about this average is

$$\overline{(I(t) - \overline{I(t)})^2} = \nu \int_{-\infty}^{+\infty} F^2(t) dt \quad (1.2-3)$$

where ν is the average number of electrons arriving per second.

The statement of the theorem is not precise until we define what we mean by "average". From the form of the equations the reader might be tempted to think of a time average; e.g. the value

$$\text{Lim}_{T \rightarrow \infty} \frac{1}{T} \int_0^T I(t) dt \quad (1.2-4)$$

However, in the proof of the theorem the average is generally taken over a great many intervals of length T with t held constant. The process is somewhat similar to that employed in (1.1) and in order to make it clear we take the case of $\overline{I(t)}$ for illustration. We observe $I(t)$ for many, say M , intervals each of length T where T is large in comparison with the interval over which the effect $F(t)$ of the arrival of a single electron is appreciable. Let ${}_n I(t')$ be the value of $I(t)$, t' seconds after the beginning of the n^{th} interval. t' is equal to t plus a constant depending upon the beginning time of the interval. We put the subscript in front because we wish to reserve the usual place for another subscript later on. The value of $\overline{I(t')}$ is then defined as

$$\overline{I(t')} = \text{Limit}_{M \rightarrow \infty} \frac{1}{M} [{}_1 I(t') + {}_2 I(t') + \dots + {}_M I(t')] \quad (1.2-5)$$

and this limit is assumed to exist. The mean square value of the fluctuation of $I(t')$ is defined in much the same way.

² *Proc. Camb. Phil. Soc.* 15 (1909), 117-136, 310-328. Our proof is similar to one given by J. M. Whittaker, *Proc. Camb. Phil. Soc.* 33 (1937), 451-458.

Actually, as the equations (1.2-2) and (1.2-3) of Campbell's theorem show, these averages and all the similar averages encountered later turn out to be independent of the time. When this is true and when the M intervals in (1.2-5) are taken consecutively the time average (1.2-4) and the average (1.2-5) become the same. To show this we multiply both sides of (1.2-5) by dt' and integrate from 0 to T :

$$\begin{aligned} \overline{I(t')} &= \text{Limit}_{M \rightarrow \infty} \frac{1}{MT} \sum_{m=1}^M \int_0^T {}_m I(t') dt' \\ &= \text{Limit}_{M \rightarrow \infty} \frac{1}{MT} \int_0^{MT} I(t) dt \end{aligned} \tag{1.2-6}$$

and this is the same as the time average (1.2-4) if the latter limit exists.

1.3 PROOF OF CAMPBELL'S THEOREM

Consider the case in which exactly K electrons arrive at the anode in an interval of length T . Before the interval starts, we think of these K electrons as fated to arrive in the interval $(0, T)$ but any particular electron is just as likely to arrive at one time as any other time. We shall number these fated electrons from one to K for purposes of identification but it is to be emphasized that the numbering has nothing to do with the order of arrival. Thus, if t_k be the time of arrival of electron number k , the probability that t_k lies in the interval $(t, t + dt)$ is dt/T .

We take T to be very large compared with the range of values of t for which $F(t)$ is appreciably different from zero. In physical applications such a range usually exists and we shall call it Δ even though it is not very definite. Then, when exactly K electrons arrive in the interval $(0, T)$ the effect is approximately

$$I_K(t) = \sum_{k=1}^K F(t - t_k) \tag{1.3-1}$$

the degree of approximation being very good over all of the interval except within Δ of the end points.

Suppose we examine a large number M of intervals of length T . The number having exactly K arrivals will be, to a first approximation $M p(K)$ where $p(K)$ is given by (1.1-3). For a fixed value of t and for each interval having K arrivals, $I_K(t)$ will have a definite value. As $M \rightarrow \infty$, the average value of the $I_K(t)$'s, obtained by averaging over the intervals, is

$$\begin{aligned} \overline{I_K(t)} &= \int_0^T \frac{dt_1}{T} \cdots \int_0^T \frac{dt_K}{T} \sum_{k=1}^K F(t - t_k) \\ &= \sum_{k=1}^K \int_0^T \frac{dt_k}{T} F(t - t_k) \end{aligned} \tag{1.3-2}$$

and if $\Delta < t < T - \Delta$, we have effectively

$$\overline{I_K(t)} = \frac{K}{T} \int_{-\infty}^{+\infty} F(t) dt \quad (1.3-3)$$

If we now average $I(t)$ over all of the M intervals instead of only over those having K arrivals, we get, as $M \rightarrow \infty$,

$$\begin{aligned} \overline{I(t)} &= \sum_{K=0}^{\infty} p(K) \overline{I_K(t)} \\ &= \sum_{K=0}^{\infty} \frac{K}{T} \frac{(\nu T)^K}{K!} e^{-\nu T} \int_{-\infty}^{+\infty} F(t) dt \\ &= \nu \int_{-\infty}^{+\infty} F(t) dt \end{aligned} \quad (1.3-4)$$

and this proves the first part of the theorem. We have used this rather elaborate proof to prove the relatively simple (1.3-4) in order to illustrate a method which may be used to prove more complicated results. Of course, (1.3-4) could be established by noting that the integral is the average value of the effect produced by one arrival, the average being taken over one second, and that ν is the average number of arrivals per second.

In order to prove the second part, (1.2-3) of Campbell's theorem we first compute $\overline{I^2(t)}$ and use

$$\begin{aligned} \overline{(I(t) - \overline{I(t)})^2} &= \overline{I^2(t)} - 2 \overline{I(t)\overline{I(t)}} + \overline{I(t)}^2 \\ &= \overline{I^2(t)} - \overline{I(t)}^2 \end{aligned} \quad (1.3-5)$$

From the definition (1.3-1) of $I_K(t)$,

$$I_K^2(t) = \sum_{k=1}^K \sum_{m=1}^K F(t - t_k) F(t - t_m)$$

Averaging this over all values of t_1, t_2, \dots, t_K with t held fixed as in (1.3-2),

$$\overline{I_K^2(t)} = \sum_{k=1}^K \sum_{m=1}^K \int_0^T \frac{dt_1}{T} \dots \int_0^T \frac{dt_K}{T} F(t - t_k) F(t - t_m)$$

The multiple integral has two different values. If $k = m$ its value is

$$\int_0^T I^2(t - t_k) \frac{dt_k}{T}$$

and if $k \neq m$ its value is

$$\int_0^T F(t - t_k) \frac{dt_k}{T} \int_0^T F(t - t_m) \frac{dt_m}{T}$$

Counting up the number of terms in the double sum shows that there are K of them having the first value and $K^2 - K$ having the second value. Hence, if $\Delta < t < T - \Delta$ we have

$$\overline{I_K^2(t)} = \frac{K}{T} \int_{-\infty}^{+\infty} F^2(t) dt + \frac{K(K-1)}{T^2} \left[\int_{-\infty}^{+\infty} F(t) dt \right]^2$$

Averaging over all the intervals instead of only those having K arrivals gives

$$\begin{aligned} \overline{I^2(t)} &= \sum_{K=0}^{\infty} p(K) \overline{I_K^2(t)} \\ &= \nu \int_{-\infty}^{+\infty} F^2(t) dt + \overline{I(t)}^2 \end{aligned}$$

where the summation with respect to K is performed as in (1.3-4), and after summation the value (1.3-4) for $\overline{I(t)}$ is used. Comparison with (1.3-5) establishes the second part of Campbell's theorem.

1.4 THE DISTRIBUTION OF $I(t)$

When certain conditions are satisfied the proportion of time which $I(t)$ spends in the range $I, I + dI$ is $P(I)dI$ where, as $\nu \rightarrow \infty$, the probability density $P(I)$ approaches

$$\frac{1}{\sigma_I \sqrt{2\pi}} e^{-(I-\bar{I})^2/2\sigma_I^2} \tag{1.4-1}$$

where \bar{I} is the average of $I(t)$ given by (1.2-2) and the square of the standard deviation σ_I , i.e. the variance of $I(t)$, is given by (1.2-3). This normal distribution is the one which would be expected by virtue of the "central limit theorem" in probability. This states that, under suitable conditions, the distribution of the sum of a large number of random variables tends toward a normal distribution whose variance is the sum of the variances of the individual variables. Similarly the average of the normal distribution is the sum of the averages of the individual variables.

So far, we have been speaking of the limiting form of the probability density $P(I)$. It is possible to write down an explicit expression for $P(I)$, which, however, is quite involved. From this expression the limiting form may be obtained. We now obtain this expression. In line with the discussion given of Campbell's theorem, we seek the probability density $P(I)$ of the values of $I(t)$ observed at t seconds from the beginning of each of a large number, M , of intervals, each of length T .

Probability that $I(t)$ lies in range $(I, I + dI)$

$$= \sum_{K=0}^{\infty} (\text{Probability of exactly } K \text{ arrivals}) \times \\ (\text{Probability that if there are exactly } \\ K \text{ arrivals, } I_K(t) \text{ lies in } (I, I + dI)).$$

Denoting the last probability in the summation by $P_K(I)dI$, using notation introduced earlier, and cancelling out the factor dI gives

$$P(I) = \sum_{K=0}^{\infty} p(K)P_K(I) \quad (1.4-2)$$

We shall compute $P_K(I)$ by the method of "characteristic functions"³ from the definition

$$I_K(t) = \sum_{k=1}^K F(t - t_k) \quad (1.3-1)$$

of $I_K(t)$. The method will be used in its simplest form: the probability that the sum

$$x_1 + x_2 + \cdots + x_K$$

of K independent random variables lies between X and $X + dX$ is

$$dX \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-ixu} \prod_{k=1}^K (\text{average value of } e^{ix_k u}) du \quad (1.4-3)$$

The average value of $e^{ix_k u}$, i.e., the characteristic function of the distribution of x_k , is obtained by averaging over the values of x_k . Although this is the simplest form of the method it is also the least general in that the integral does not converge for some important cases. The distribution which gives a probability of $\frac{1}{2}$ that $x_k = -1$ and $\frac{1}{2}$ that $x_k = +1$ is an example of such a case. However, we may still use (1.4-3) formally in such cases by employing the relation

$$\int_{-\infty}^{+\infty} e^{-iau} du = 2\pi\delta(a) \quad (1.4-4)$$

where $\delta(a)$ is zero except at $a = 0$ where it is infinite and its integral from $a = -\epsilon$ to $a = +\epsilon$ is unity where $\epsilon > 0$.

When we identify x_k with $F(t - t_k)$ we see that the average value of $e^{ix_k u}$ is

$$\frac{1}{T} \int_0^T \exp [iuF(t - t_k)] dt_k$$

³ The essentials of this method are due to Laplace. A few remarks on its history are given by E. C. Molina, *Bull. Amer. Math. Soc.*, 36 (1930), pp. 369-392. An account of the method may be found in any one of several texts on probability theory. We mention "Random Variables and Probability Distributions," by H. Cramér, Camb. Tract in Math. and Math. Phys. No. 36 (1937), Chap. IV. Also "Introduction to Mathematical Probability," by J. V. Uspensky, McGraw-Hill (1937), pages 240, 264, and 271-278.

All of the K characteristic functions are the same and hence, from (1.4-3), $P_K(I)dI$ is

$$dI \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-iIu} \left(\frac{1}{T} \int_0^T \exp [iuF(t - \tau)] d\tau \right)^K du$$

Although in deriving this relation we have taken $K > 0$, it also holds for $K = 0$ (provided we use (1.4-4)). In this case $P_0(I) = \delta(I)$, because $I = 0$ when no electrons arrive.

Inserting our expression for $P_K(I)$ and the expression (1.1-3) for $p(K)$ in (1.4-2) and performing the summation gives

$$P(I) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \exp \left(-iIu - \nu T + \nu \int_0^T \exp [iuF(t - \tau)] d\tau \right) du \quad (1.4-5)$$

The first exponential may be simplified somewhat. Using

$$\nu T = \nu \int_0^T d\tau$$

permits us to write

$$-\nu T + \nu \int_0^T \exp [iuF(t - \tau)] d\tau = \nu \int_0^T (\exp [iuF(t - \tau)] - 1) d\tau$$

Suppose that $\Delta < t < T - \Delta$ where Δ is the range discussed in connection with equation (1.3-1). Taking $|F(t - \tau)| = 0$ for $|t - \tau| > \Delta$ then enables us to write the last expression as

$$\nu \int_{-\infty}^{+\infty} [e^{iuF(t)} - 1] dt \quad (1.4-6)$$

Placing this in (1.4-5) yields the required expression for $P(I)$:

$$P(I) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \exp \left(-iIu + \nu \int_{-\infty}^{+\infty} [e^{iuF(t)} - 1] dt \right) du \quad (1.4-7)$$

An idea of the conditions under which the normal law (1.4-1) is approached may be obtained from (1.4-7) by expanding (1.4-6) in powers of u and determining when the terms involving u^3 and higher powers of u may be neglected. This is taken up for a slightly more general form of current in section 1.6.

1.5 EXTENSION OF CAMPBELL'S THEOREM

In section 1.2 we have stated Campbell's theorem. Here we shall give an extension of it. In place of the expression (1.2-1) for the $I(t)$ of the shot effect we shall deal with the current

$$I(t) = \sum_{k=-\infty}^{+\infty} a_k F(t - t_k) \quad (1.5-1)$$

where $F(t)$ is the same sort of function as before and where $\dots a_1, a_2, \dots a_k, \dots$ are independent random variables all having the same distribution. It is assumed that all of the moments \bar{a}^n exist, and that the events occur at random

The extension states that the n th semi-invariant of the probability density $P(I)$ of I , where I is given by (1.5-1), is

$$\lambda_n = \nu \bar{a}^n \int_{-\infty}^{+\infty} [F(t)]^n dt \quad (1.5-2)$$

where ν is the expected number of events per second. The semi-invariants of a distribution are defined as the coefficients in the expansion

$$\log_e (\text{ave. } e^{iu}) = \sum_{n=1}^N \frac{\lambda_n}{n!} (iu)^n + o(u^N) \quad (1.5-3)$$

i.e. as the coefficients in the expansion of the logarithm of the characteristic function. The λ 's are related to the moments of the distribution. Thus if m_1, m_2, \dots denote the first, second \dots moments about zero we have

$$\text{ave. } e^{iu} = 1 + \sum_{n=1}^N \frac{m_n}{n!} (iu)^n + o(u^N)$$

By combining this relation with the one defining the λ 's it may be shown that

$$\begin{aligned} \bar{I} &= m_1 = \lambda_1 \\ \bar{I}^2 &= m_2 = \lambda_2 + \lambda_1 m_1 \\ \bar{I}^3 &= m_3 = \lambda_3 + 2\lambda_2 m_1 + \lambda_1 m_2 \\ &\dots \end{aligned}$$

It follows that $\lambda_1 = \bar{I}$ and $\lambda_2 = \text{ave. } (I - \bar{I})^2$. Hence (1.5-2) yields the original statement of Campbell's theorem when we set n equal to one and two and also take all the a 's to be unity.

The extension follows almost at once from the generalization of expression (1.4-7) for the probability density $P(I)$. By proceeding as in section 1.4 and identifying x_k with $a_k F(t - t_k)$ we see that

$$\text{ave. } e^{ix_k u} = \frac{1}{T} \int_{-\infty}^{+\infty} q(a) da \int_0^T \exp [iuaF(t - t_k)] dt_k$$

where $q(a)$ is the probability density function for the a 's. It turns out that the probability density $P(I)$ of I as defined by (1.5-1) is

$$P(I) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \exp \left(-iIu + \nu \int_{-\infty}^{+\infty} q(a) da \int_{-\infty}^{+\infty} [e^{iuaF(t)} - 1] dt \right) du \quad (1.5-4)$$

The logarithm of the characteristic function of $P(I)$ is, from (1.5-4),

$$\begin{aligned} \nu \int_{-\infty}^{+\infty} q(a) da \int_{-\infty}^{+\infty} [e^{iuaF(t)} - 1] dt \\ = \sum_{n=1}^{\infty} \frac{(iu)^n}{n!} \nu \int_{-\infty}^{+\infty} q(a) da a^n \int_{-\infty}^{+\infty} F^n(t) dt \end{aligned}$$

Comparison with the series (1.5-3) defining the semi-invariants gives the extension of Campbell's theorem stated by (1.5-2).

Other extensions of Campbell's theorem may be made. For example, suppose in the expression (1.5-1) for $I(t)$ that $t_1, t_2, \dots, t_k, \dots$ while still random variables, are no longer necessarily distributed according to the laws assumed above. Suppose now that the probability density $p(x)$ is given where x is the interval between two successive events:

$$t_2 = t_1 + x_1 \quad (1.5-5)$$

$$t_3 = t_2 + x_2 = t_1 + x_1 + x_2$$

and so on. For the case treated above

$$p(x) = \nu e^{-\nu x}. \quad (1.5-6)$$

We assume that the expected number of events per second is still ν .

Also we take the special, but important, case for which

$$F(t) = 0, \quad t < 0 \quad (1.5-7)$$

$$F(t) = e^{-\alpha t}, \quad t > 0.$$

For a very long interval extending from $t = t_1$ to $t = T + t_1$ inside of which there are exactly K events we have, if t is not near the ends of the interval,

$$\begin{aligned} I(t) &= a_1 F(t - t_1) + a_2 F(t - t_1 - x_1) + \dots \\ &\quad + a_{K+1} F(t - t_1 - x_1 - \dots - x_K) \\ &= a_1 F(t') + a_2 F(t' - x_1) + \dots + a_{K+1} F(t' - x_1 - \dots - x_K) \end{aligned}$$

$$I^2(t) = a_1^2 F^2(t') + a_2^2 F^2(t' - x_1) + \cdots + a_{K+1}^2 F^2(t' - x_1 \cdots - x_K) \\ + 2a_1 a_2 F(t') F(t' - x_1) + \cdots + 2a_1 a_{K+1} F(t') F(t' - x_1 \cdots - x_K) \\ + 2a_2 a_3 F(t' - x_1) F(t' - x_1 - x_2) + \cdots + \cdots$$

where $t' = t - t_1$. If we integrate $I^2(t)$ over the entire interval $0 < t' < T$ and drop the primes we get approximately

$$\int_0^T I^2(t) dt = (a_1^2 + \cdots + a_{K+1}^2) \varphi(0) \\ + 2a_1 a_2 \varphi(x_1) + 2a_1 a_3 \varphi(x_1 + x_2) + \cdots + 2a_1 a_{K+1} \varphi(x_1 + \cdots + x_K) \\ + 2a_2 a_3 \varphi(x_2) + \cdots + \cdots + 2a_K a_{K+1} \varphi(x_K)$$

where

$$\varphi(x) = \int_{-\infty}^{+\infty} F(t) F(t - x) dx$$

When we divide both sides by T and consider K and T to be very large,

$$\frac{K}{T} \frac{a_1^2 + \cdots + a_{K+1}^2}{K} \varphi(0) \approx \overline{va^2} \varphi(0)$$

$$\frac{1}{T} [a_1 a_2 \varphi(x_1) + a_2 a_3 \varphi(x_2) + \cdots + a_K a_{K+1} \varphi(x_K)] = \frac{K}{T} \text{average } a_k a_{k+1} \varphi(x_k)$$

$$\approx \overline{va^2} \int_0^{\infty} \varphi(x) p(x) dx$$

$$\frac{1}{T} [a_1 a_3 \varphi(x_1 + x_2) + \cdots] = \frac{K-1}{T} \text{ave. } a_k a_{k+3} \varphi(x_k + x_{k+1})$$

$$\approx \overline{va^2} \int_0^{\infty} dx_1 \int_0^{\infty} dx_2 p(x_1) p(x_2) \varphi(x_1 + x_2)$$

Consequently

$$\overline{I^2(t)} = \text{Lim}_{T \rightarrow \infty} \frac{1}{T} \int_0^T I^2(t) dt \\ = \overline{va^2} \varphi(0) + 2\overline{va^2} \left[\int_0^{\infty} p(x) \varphi(x) dx \right. \\ \left. + \int_0^{\infty} dx_1 \int_0^{\infty} dx_2 p(x_1) p(x_2) \varphi(x_1 + x_2) + \cdots \right]$$

For our special exponential form (1.5-7) for $F(t)$,

$$\varphi(x) = \frac{e^{-\alpha x}}{2\alpha}$$

and the multiple integrals occurring in the expression for $\overline{I^2(t)}$ may be written in terms of powers of

$$q = \int_0^\infty p(x)e^{-\alpha x} dx \tag{1.5-8}$$

Thus

$$2\alpha\overline{I^2(t)} = v\bar{a}^2 + 2\bar{a}^2v \frac{q}{1-q}$$

and since

$$\overline{I(t)} = v\bar{a} \int_{-\infty}^{+\infty} F(t) dt = v\bar{a}/\alpha$$

we have

$$\overline{I^2(t)} - \overline{I(t)}^2 = \frac{v\bar{a}^2}{2\alpha} + \left(\frac{v\bar{a}}{\alpha}\right)^2 \left[\frac{\alpha q}{v(1-q)} - 1 \right] \tag{1.5-9}$$

Equations (1.5-8) and (1.5-9) give us an extension of Campbell's theorem subject to the restrictions discussed in connection with equations (1.5-5) and (1.5-7). Other generalizations have been made⁴ but we shall leave the subject here. The reader may find it interesting to verify that (1.5-9) gives the correct answer when $p(x)$ is given by (1.5-6), and also to investigate the case when the events are spaced equally.

1.6 APPROACH OF DISTRIBUTION OF I TO A NORMAL LAW

In section 1.5 we saw that the probability density $P(I)$ of the noise current I may be expressed formally as

$$P(I) + \frac{1}{2\pi} \int_{-\infty}^{+\infty} \exp \left[-iIu + \sum_{n=1}^{\infty} (iu)^n \lambda_n/n! \right] du \tag{1.6-1}$$

where λ_n is the n th semi-invariant given by (1.5-2). By setting

$$\begin{aligned} \lambda_2 &= \sigma^2 \\ x &= \frac{I - \lambda_1}{\sigma} = \frac{I - \overline{I}}{\sigma} \end{aligned} \tag{1.6-2}$$

⁴ See E. N. Rowland, *Proc. Camb. Phil. Soc.* 32 (1936), 580-597. He extends the theorem to the case where there are two functions instead of a single one, which we here denote by $I(t)$. According to a review in the *Zentralblatt für Math.*, 19, p. 224, Khintchine in the *Bull. Acad. Sci. URSS, sér. Math.* Nr. 3 (1938), 313-322, has continued and made precise the earlier work of Rowland.

expanding

$$\exp \sum_{n=3}^{\infty} (iu)^n \lambda_n / n!$$

as a power series in u , integrating termwise using

$$\frac{1}{2\pi} \int_{-\infty}^{+\infty} (iu\sigma)^n \exp \left[-iu\sigma x - \frac{u^2 \sigma^2}{2} \right] du = (-)^n \sigma^{-1} \varphi^{(n)}(x),$$

$$\varphi^{(n)}(x) = \frac{1}{\sqrt{2\pi}} \frac{d^n}{dx^n} e^{-x^2/2}$$

and finally collecting terms according to their order in powers of $\nu^{-1/2}$, gives

$$P(I) \sim \sigma^{-1} \varphi^{(0)}(x) - \frac{\lambda_3 \sigma^{-4}}{3!} \varphi^{(3)}(x) + \left[\frac{\lambda_4 \sigma^{-5}}{4!} \varphi^{(4)}(x) + \frac{\lambda_3^2 \sigma^{-7}}{72} \varphi^{(6)}(x) \right] + \dots \quad (1.6-3)$$

The first term is $O(\nu^{-1/2})$, the second term is $O(\nu^{-1})$, and the term within brackets is $O(\nu^{-3/2})$. This is Edgeworth's series.⁵ The first term gives the normal distribution and the remaining terms show how this distribution is approached as $\nu \rightarrow \infty$.

1.7 THE FOURIER COMPONENTS OF $I(t)$

In some analytical work noise current is represented as

$$I(t) = \frac{a_0}{2} + \sum_{n=1}^N \left(a_n \cos \frac{2\pi n t}{T} + b_n \sin \frac{2\pi n t}{T} \right) \quad (1.7-1)$$

where at a suitable place in the work T and N are allowed to become infinite. The coefficients a_n and b_n , $1 \leq n \leq N$, are regarded as independent random variables distributed about zero according to a normal law.

It appears that the association of (1.7-1) with a sequence of disturbances occurring at random goes back many years. Rayleigh⁶ and Gouy suggested that black-body radiation and white light might both be regarded as sequences of irregularly distributed impulses.

Einstein⁷ and von Laue have discussed the normal distribution of the coefficients in (1.7-1) when it is used to represent black-body radiation, this radiation being the resultant produced by a great many independent os-

⁵ See, for example, pp. 86-87, in "Random Variables and Probability Distributions" by H. Cramér, *Cambridge Tract No. 36* (1937).

⁶ *Phil. Mag. Ser. 5*, Vol. 27 (1889) pp. 460-469.

⁷ A. Einstein and L. Hopf, *Ann. d. Physik* 33 (1910) pp. 1095-1115.

M. V. Laue, *Ann. d. Physik* 47 (1915) pp. 853-878.

A. Einstein, *Ann. d. Physik* 47 (1915) pp. 879-885.

M. V. Laue, *Ann. d. Physik* 48 (1915) pp. 668-680.

I am indebted to Prof. Goudsmit for these references.

cillators. Some argument arose as to whether the coefficients in (1.7-1) were statistically independent or not. It was finally decided that they are independent.

The shot effect current has been represented in this way by Schottky.⁸ The Fourier series representation has been discussed by H. Nyquist⁹ and also by Goudsmit and Weiss. Remarks made by A. Schuster¹⁰ are equivalent to the statement that a_n and b_n are distributed normally.

In view of this wealth of information on the subject it may appear superfluous to say anything about it. However, for the sake of completeness, we shall outline the thoughts which lead to (1.7-1).

In line with our usual approach to the shot effect, we suppose that exactly K electrons arrive during the interval $(0, T)$, so that the noise current for the interval is

$$I_K(t) = \sum_{k=1}^K F(t - t_k) \tag{1.7-2}$$

The coefficients in the Fourier series expansion of $I_K(t)$ over the interval $(0, T)$ are a_{nK} and b_{nK} where

$$\begin{aligned} a_{nK} - ib_{nK} &= \frac{2}{T} \sum_{k=1}^K \int_0^T F(t - t_k) \exp \left[-i \frac{2\pi n t}{T} \right] dt \\ &= \frac{2}{T} \sum_{k=1}^K \int_{-\infty}^{+\infty} F(t) \exp \left[-i \frac{2\pi n}{T} (t + t_k) \right] dt \\ &= R_n e^{-i\varphi_n} \sum_{k=1}^K e^{-in\theta_k} \end{aligned} \tag{1.7-3}$$

In this expression

$$\theta_k = \frac{2\pi t_k}{T} \tag{1.7-4}$$

$$R_n e^{-i\varphi_n} = C_n - iS_n = \frac{2}{T} \int_{-\infty}^{+\infty} F(t) e^{-i2\pi n t/T} dt$$

In the earlier sections the arrival times t_1, t_2, \dots, t_K were regarded as K independent random variable each distributed uniformly over the interval $(0, T)$. Hence the θ_k 's may be regarded as random variables distributed uniformly over the interval 0 to 2π .

Incidentally, it will be noted that in (1.7-3) there occurs the sum of K randomly oriented unit vectors. When K becomes very large, as it does

⁸ *Ann. d. Physik*, 57 (1918) pp. 541-567.

⁹ Unpublished Memorandum, "Fluctuations in Vacuum Tube Noise and the Like," March 17, 1932.

¹⁰ Investigation of Hidden Periodicities, *Terrestrial Magnetism*, 3 (1898), pp. 13-41. See especially propositions 1 and 2 on page 26 of Schuster's paper.

when $\nu \rightarrow \infty$, it is known that the real and imaginary parts of this sum are random variables, which tend to become independent and normally distributed about zero. This suggests the manner in which the normal distribution of the coefficients arises. Averaging over the θ_k 's in (1.7-3) gives when $n > 0$

$$\bar{a}_{nK} = \bar{b}_{nK} = 0 \quad (1.7-5)$$

Some further algebra gives

$$\overline{a_{nK}^2} = \overline{b_{nK}^2} = \frac{K}{2} R_n^2 \quad (1.7-6)$$

$$\overline{a_{nK} b_{nK}} = \overline{a_{nK} a_{mK}} = \overline{b_{nK} b_{mK}} = 0$$

where $n \neq m$ and $n, m > 0$.

So far, we have been considering the case of exactly K arrivals in our interval of length T . Now we pass to the general case of any number of arrivals by making use of formulas analogous to

$$\overline{a_n^2} = \sum_{K=0}^{\infty} p(K) \overline{a_{nK}^2} \quad (1.7-7)$$

as has been done in section 1.3. Thus, for $n > 0$,

$$\bar{a}_n = \bar{b}_n = 0$$

$$\overline{a_n^2} = \overline{b_n^2} = \frac{\nu T}{2} R_n^2 = \sigma_n^2 \quad (1.7-8)$$

$$\overline{a_n b_n} = \overline{a_n a_m} = \overline{b_n b_m} = 0, \quad n \neq m$$

In the second line we have used σ_n to denote the standard deviation of a_n and b_n . We may put the expression for σ_n^2 in a somewhat different form by writing

$$f_n = \frac{n}{T} = n\Delta f, \quad \Delta f = \frac{1}{T} \quad (1.7-9)$$

where f_n is the frequency of the n th component. Using (1.7-4),

$$\sigma_n^2 = 2\nu\Delta f \left| \int_{-\infty}^{+\infty} F(t) e^{-i2\pi f_n t} dt \right|^2 \quad (1.7-10)$$

Thus, σ_n^2 is proportional to ν/T .

The probability density function $P(a_1, \dots, a_N, b_1, \dots, b_N)$ for the $2N$ coefficients, $a_1, \dots, a_N, b_1, \dots, b_N$ may be derived in much the same fashion as was the probability density of the noise current in section 1.4. Here N

is arbitrary but fixed. The expression analogous to (1.4-5) is the $2N$ fold integral

$$P(a_1, \dots, b_N) = (2\pi)^{-2N} \int_{-\infty}^{+\infty} du_1 \dots \int_{-\infty}^{+\infty} dv_N \quad (1.7-11)$$

$$\exp [-i(a_1 u_1 + \dots + b_N v_N) - \nu T + \nu T E]$$

where

$$E = \frac{1}{2\pi} \int_0^{2\pi} d\theta \exp \left[i \sum_{n=1}^N (u_n C_n + v_n S_n) \cos n\theta + (v_n C_n - u_n S_n) \sin n\theta \right] \quad (1.7-12)$$

in which $C_n - iS_n$ is defined as the Fourier transform (1.7-4) of $F(t)$.

The next step is to show that (1.7-11) approaches a normal law in $2N$ dimensions as $\nu \rightarrow \infty$. This appears to be quite involved. It will be noted that the integrand in the integral defining E is composed of N factors of the form

$$\exp [i\rho_n \cos (n\theta - \psi_n)]$$

$$= J_0(\rho_n) + 2i \cos (n\theta - \psi_n) J_1(\rho_n) - 2 \cos (2n\theta - 2\psi_n) J_2(\rho_n) + \dots$$

where

$$\rho_n^2 = (u_n^2 + v_n^2)(C_n^2 + S_n^2) = \frac{2}{\nu T} \sigma_n^2 (u_n^2 + v_n^2).$$

As ν becomes large, it turns out that the integral (1.7-11) for the probability density obtains most of its contributions from small values of u and v . By substituting the product of the Bessel function series in the integral for E and integrating we find

$$E = \prod_{n=1}^N J_0(\rho_n) + A + B + C$$

where A is the sum of products such as

$$-2i \cos (\psi_{k+l} - \psi_k - \psi_l) J_1(\rho_k) J_1(\rho_l) J_1(\rho_{k+l}) \text{ times } N - 3 J_0\text{'s}$$

in which $0 < k \leq l$ and $2 \leq k + l \leq N$. Similarly B is the sum of products of the form

$$-2i \cos (\psi_{2k} - 2\psi_k) J_1(\rho_{2k}) J_2(\rho_k) \text{ times } N - 2 J_0\text{'s}$$

C consists of terms which give fourth and higher powers in u and v . There are roughly $N^2/4$ terms of form A and $N/2$ terms of form B .

Expanding the Bessel functions, neglecting all powers above the third and

proceeding as in section 1.4, will give us the normal distribution plus the first correction term. It is rather a messy affair. An idea of how it looks may be obtained by taking the special case in which $F(t)$ is an even function of t and neglecting terms of type B . Then

$$P(a_1, \dots, a_N, b_1, \dots, b_N) = (1 + \eta) \prod_{n=1}^N \frac{e^{-(x_n^2 + y_n^2)/2}}{2\pi\sigma_n^2} \quad (1.7-12)$$

where

$$x_n = \frac{a_n}{\sigma_n}, \quad y_n = \frac{b_n}{\sigma_n}$$

$$\eta = (2\nu T)^{-1/2} \sum_{k,l} [x_{k+l}(x_k x_l - y_k y_l) + 2 y_{k+l} y_k y_l] \quad (1.7-13)$$

and the summation extends over $2 \leq k + l \leq N$ with $k \leq l$.

It is seen that if T and N are held constant, the correction term η approaches zero as ν becomes very large. A very rough idea of the magnitude of η may be obtained by assuming that unity is a representative value of the x 's and y 's. Further assuming that there are N^2 terms in the summation each one of which may be positive or negative suggests that magnitude of the sum is of the order of N . Hence we might expect to find that η is of the order of $N(2\nu T)^{-1/2}$.

PART II

POWER SPECTRA AND CORRELATION FUNCTIONS

2.0 INTRODUCTION

A theory for analyzing functions of time, t , which do not die down and which remain finite as t approaches infinity has gradually been developed over the last sixty years. A few words of its history together with an extensive bibliography are given by N. Wiener in his paper on "Generalized Harmonic Analysis".¹¹ G. Gouy, Lord Rayleigh and A. Schuster were led to study this problem in their investigations of such things as white light and noise. Schuster¹² invented the "periodogram" method of analysis which has as its object the discovery of any periodicities hidden in a continuous curve representing meteorological or economic data.

¹¹ *Acta Math.*, Vol. 55, pp. 117-258 (1930). See also "Harmonic Analysis of Irregular Motion," *Jour. Math. and Phys.* 5 (1926) pp. 99-189.

¹² The periodogram was first introduced by Schuster in reference 10 cited in Section 1.7. He later modified its definition in the *Trans. Camb. Phil. Soc.* 18 (1903), pp. 107-135, and still later redefined it in "The Periodogram and its Optical Analogy," *Proc. Roy. Soc., London, Ser. A*, 77 (1906) pp. 136-140. In its final form the periodogram is equivalent to $\frac{1}{2}w(f)$, where $w(f)$ is the power spectrum defined in Section 2.1, plotted as a function of the period $T = (2\pi f)^{-1}$.

The correlation function, which turns out to be a very useful tool, apparently was introduced by G. I. Taylor.¹³ Recently it has been used by quite a few writers¹⁴ in the mathematical theory of turbulence.

In section 2.1 the power spectrum and correlation function of a specific function, such as one given by a curve extending to $t = \infty$, are defined by equations (2.1-3) and (2.1-4) respectively. That they are related by the Fourier inversion formulae (2.1-5) and (2.1-6) is merely stated; the discussion of the method of proof being delayed until sections 2.3 and 2.4. In section 2.3 a discussion based on Fourier series is given and in section 2.4 a parallel treatment starting with Parseval's integral theorem is set forth. The results as given in section 2.1 have to be supplemented when the function being analyzed contains a d.c. or periodic components. This is taken up in section 2.2.

The first four sections deal with the analysis of a specific function of t . However, most of the applications are made to functions which behave as though they are more or less random in character. In the mathematical analysis this randomness is introduced by assuming the function of t to be also a function of suitable parameters, and then letting these parameters be random variables. This question is taken up in section 2.5. In section 2.6 the results of 2.5 are applied to determine the average power spectrum and the average correlation function of the shot effect current. The same thing is done in 2.7 for a flat top wave, the tops (and bottoms) being of random length. The case in which the intervals are of equal length but the sign of the wave is random is also discussed in 2.7. The representation of the noise current as a trigonometrical series with random variable coefficients is taken up in 2.8. The last two sections 2.9 and 2.10 are devoted to probability theory. The normal law and the central limit theorem, respectively, are discussed.

2.1 SOME RESULTS OF GENERALIZED HARMONIC ANALYSIS

We shall first state the results which we need, and then show that they are plausible by methods which are heuristic rather than rigorous. Suppose that $I(t)$ is one of the functions mentioned above. We may think of it as being specified by a curve extending from $t = -\infty$ to $t = \infty$. $I(t)$ may be regarded as composed of a great number of sinusoidal components whose frequencies range from 0 to $+\infty$. It does not necessarily have to be a noise current, but if we think of it as such, then, in flowing through a resistance of one ohm it will dissipate a certain average amount of power, say ρ watts.

¹³ Diffusion by Continuous Movements, *Proc. Lond. Math. Soc.*, Ser. 2, 20, pp. 196-212 (1920).

¹⁴ See the text "Modern Developments in Fluid Dynamics" edited by S. Goldstein, Oxford (1938).

That portion of ρ arising from the components having frequencies between f and $f + df$ will be denoted by $w(f)df$, and consequently

$$\rho = \int_0^{\infty} w(f)df \quad (2.1-1)$$

Since $w(f)$ is the spectrum of the average power we shall call it the "power spectrum" of $I(t)$. It has the dimensions of energy and on this account is frequently called the "energy-frequency spectrum" of $I(t)$. A mathematical formulation of this discussion leads to a clear cut definition of $w(f)$.

Let $\Phi(t)$ be a function of t , which is zero outside the interval $0 \leq t \leq T$ and is equal to $I(t)$ inside the interval. Its spectrum $S(f)$ is given by

$$S(f) = \int_0^T I(t)e^{-2\pi ift} dt \quad (2.1-2)$$

The spectrum of the power, $w(f)$, is defined as

$$w(f) = \text{Limit}_{T \rightarrow \infty} \frac{2|S(f)|^2}{T} \quad (2.1-3)$$

where we consider only values of $f > 0$ and assume that this limit exists. This is substantially the definition of $w(f)$ given by J. R. Carson¹⁵ and is useful when $I(t)$ has no periodic terms and no d.c. component. In the latter case (2.1-3) must either be supplemented by additional definitions or else a somewhat different method of approach used. These questions will be discussed in section 2.2.

The correlation function $\psi(\tau)$ of $I(t)$ is defined by the limit

$$\psi(\tau) = \text{Limit}_{T \rightarrow \infty} \frac{1}{T} \int_0^T I(t)I(t + \tau) dt \quad (2.1-4)$$

which is assumed to exist. $\psi(\tau)$ is closely related to the correlation coefficients used in statistical theory to measure the correlation of two random variables. In the present case the value of $I(t)$ at time t is one variable and its value at a different time $t + \tau$ is the other variable.

The spectrum of the power $w(f)$ and the correlation function $\psi(\tau)$ are related by the equations

$$w(f) = 4 \int_0^{\infty} \psi(\tau) \cos 2\pi f\tau d\tau \quad (2.1-5)$$

$$\psi(\tau) = \int_0^{\infty} w(f) \cos 2\pi f\tau df \quad (2.1-6)$$

¹⁵ "The Statistical Energy-Frequency Spectrum of Random Disturbances," *B.S.T.J.*, Vol. 10, pp. 374-381 (1931).

It is seen that $\psi(\tau)$ is an even function of τ and that

$$\psi(0) = \rho \tag{2.1-7}$$

When either $\psi(\tau)$ or $w(f)$ is known the other may be obtained provided the corresponding integral converges.

2.2 POWER SPECTRUM FOR D.C. AND PERIODIC COMPONENTS

As mentioned in section 2.1, when $I(t)$ has a d.c. or a periodic component the limit in the definition (2.1-3) for $w(f)$ does not exist for f equal to zero or to the frequency of the periodic component. Perhaps the most satisfactory method of overcoming this difficulty, from the mathematical point of view, is to deal with the integral of the power spectrum.¹⁶

$$\int_0^f w(g) dg \tag{2.2-1}$$

instead of with $w(f)$ itself.

The definition (2.1-4) for $\psi(\tau)$ still holds. If, for example,

$$I(t) = A + C \cos(2\pi f_0 t - \varphi) \tag{2.2-2}$$

$\psi(\tau)$ as given by (2.1-4) is

$$\psi(\tau) = A^2 + \frac{C^2}{2} \cos 2\pi f_0 \tau \tag{2.2-3}$$

The inversion formulas (2.1-5) and (2.1-6) give

$$\begin{aligned} \int_0^f w(g) dg &= \frac{2}{\pi} \int_0^\infty \psi(\tau) \frac{\sin 2\pi f \tau}{\tau} d\tau \\ \psi(\tau) &= \int_0^\infty \cos 2\pi f \tau d \left[\int_0^f w(g) dg \right] \end{aligned} \tag{2.2-4}$$

¹⁶ This is done by Wiener,¹¹ loc. cit., and by G. W. Kenrick, "The Analysis of Irregular Motions with Applications to the Energy Frequency Spectrum of Static and of Telegraph Signals," *Phil. Mag.*, Ser. 7, Vol. 7, pp. 176-196 (Jan. 1929). Kenrick appears to be one of the first to apply, to noise problems, the correlation function method of computing the power spectrum (one of his problems is discussed in Sec. 2.7). He bases his work on results due to Wiener. Khintchine, in "Korrelationstheorie der stationären stochastischen Prozesse," *Math. Annalen*, 109 (1934), pp. 604-615, proves the following theorem: A necessary and sufficient condition that a function $R(t)$ may be the correlation function of a continuous, stationary, stochastic process is that $R(t)$ may be expressed as

$$R(t) = \int_{-\infty}^{+\infty} \cos tx dF(x)$$

where $F(x)$ is a certain distribution function. This expression for $R(t)$ is essentially the second of equations (2.2-4). Khintchine's work has been extended by H. Cramér, "On the theory of stationary random processes," *Ann. of Math.*, Ser. 2, Vol. 41 (1941), pp. 215-230. However, Khintchine and Cramér appear to be interested primarily in questions of existence, representation, etc., and do not stress the concept of the power spectrum.

where the last integral is to be regarded as a Stieltjes' integral. When the expression (2.2-3) for $\psi(\tau)$ is placed in the first formula of (2.2-4) we get

$$\int_0^f w(g) dg = \begin{cases} A^2 & \text{when } 0 < f < f_0 \\ A^2 + \frac{C^2}{2}, & \text{" } f > f_0 \end{cases} \quad (2.2-5)$$

When this expression is used in the second formula of (2.2-4), the increments of the differential are seen to be A^2 at $f = 0$ and $C^2/2$ at $f = f_0$. The resulting expression for $\psi(\tau)$ agrees with the original one.

Here we desire to use a less rigorous, but more convenient, method of dealing with periodic components. By examining the integral of $w(f)$ as given by (2.2-5) we are led to write

$$w(f) = 2A^2\delta(f) + \frac{C^2}{2}\delta(f - f_0) \quad (2.2-6)$$

where $\delta(x)$ is an even unit impulse function so that if $\epsilon > 0$

$$\int_0^\epsilon \delta(x) dx = \frac{1}{2} \int_{-\epsilon}^\epsilon \delta(x) dx = \frac{1}{2} \quad (2.2-7)$$

and $\delta(x) = 0$ except at $x = 0$, where it is infinite. This enables us to use the simpler inversion formulas of section 2.1. The second of these, (2.1-6), is immediately seen to give the correct expression for $\psi(\tau)$. The first one, (2.1-5), gives the correct expression for $w(f)$ provided we interpret the integrals as follows:

$$\begin{aligned} \int_0^\infty \cos 2\pi f\tau d\tau &= \frac{1}{2}\delta(f) \\ \int_0^\infty \cos 2\pi f_0\tau \cos 2\pi f\tau d\tau &= \frac{1}{4}\delta(f - f_0) \end{aligned} \quad (2.2-8)$$

It is not hard to show that these are in agreement with the fundamental interpretation

$$\int_{-\infty}^{+\infty} e^{-i2\pi f t} dt = \int_{-\infty}^{+\infty} e^{i2\pi f t} dt = \delta(f) \quad (2.2-9)$$

which in its turn follows from a formal application of the Fourier integral formula and

$$\int_{-\infty}^{+\infty} \delta(f)e^{i2\pi f t} df = \int_{-\infty}^{+\infty} \delta(f)e^{-i2\pi f t} df = 1 \quad (2.2-10)$$

We must remember that $f_0 > 0$ and $f \geq 0$ in (2.2-8) so that $\delta(f + f_0) = 0$ for $f \geq 0$.

The definition (2.1-3) for $w(f)$ gives the continuous part of the power spectrum. In order to get the part due to the d.c. and periodic components, which is exemplified by the expression (2.2-6) for $w(f)$ involving the δ functions, we must supplement (2.1-3) by adding terms of the type

$$2A^2\delta(f) + \frac{C^2}{2}\delta(f - f_0) = \left[\text{Limit}_{T \rightarrow \infty} \frac{2|S(0)|^2}{T^2} \right] \delta(f) + \left[\text{Limit}_{T \rightarrow \infty} \frac{2|S(f_0)|^2}{T^2} \right] \delta(f - f_0) \tag{2.2-11}$$

The correctness of this expression may be verified by calculating $S(f)$ for the $I(t)$ of this section given by (2.2-2), and actually carrying out the limiting process.

2.3 DISCUSSION OF RESULTS OF SECTION ONE—FOURIER SERIES

The fact that the spectrum of power $w(f)$ and the correlation function $\psi(\tau)$ are related by Fourier inversion formulas is closely connected with Parseval's theorems for Fourier series and integrals. In this section we shall not use Parseval's theorems explicitly. We start with Fourier's series and use the concept of each component dissipating its share of energy independently of the behavior of the other components.

Let that portion of $I(t)$ which lies in the interval $0 \leq t < T$ be expanded in the Fourier series

$$I(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left(a_n \cos \frac{2\pi nt}{T} + b_n \sin \frac{2\pi nt}{T} \right) \tag{2.3-1}$$

where

$$a_n = \frac{2}{T} \int_0^T I(t) \cos \frac{2\pi nt}{T} dt$$

$$b_n = \frac{2}{T} \int_0^T I(t) \sin \frac{2\pi nt}{T} dt \tag{2.3-2}$$

Then for the interval $-\tau \leq t < T - \tau$,

$$I(t + \tau) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left(a_n \cos \frac{2\pi n(t + \tau)}{T} + b_n \sin \frac{2\pi n(t + \tau)}{T} \right) \tag{2.3-3}$$

Multiplying the series for $I(t)$ and $I(t + \tau)$ together and integrating with respect to t gives, after some reduction,

$$\frac{1}{T} \int_0^T I(t)I(t + \tau) dt = \frac{a_0^2}{4} + \sum_{n=1}^{\infty} \frac{1}{2} (a_n^2 + b_n^2) \cos \frac{2\pi n}{T} \tau + O\left(\frac{\tau I^2}{T}\right) \tag{2.3-4}$$

where the last term represents correction terms which must be added because the series (2.3-3) does not represent $I(t + \tau)$ in the interval $(T - \tau, T)$ when $\tau > 0$, or in the interval $(0, -\tau)$ if $\tau < 0$.

If $I(t)$ were a current and if it were to flow through one ohm for the interval $(0, T)$, each component would dissipate a certain average amount of power. The average power dissipated by the component of frequency $f_n = n/T$ cycles per second would be, from the Fourier series and elementary principles,

$$\begin{aligned} \frac{1}{2} (a_n^2 + b_n^2) \text{ watts,} & \quad n \neq 0 \\ \frac{a_0^2}{4} \text{ watts,} & \quad n = 0 \end{aligned} \tag{2.3-5}$$

The band width associated with the n th component is the difference in frequency between the $n + 1$ th and n th components:

$$f_{n+1} - f_n = \frac{n+1}{T} - \frac{n}{T} = \frac{1}{T} \text{ cps}$$

Hence if the average power in the band $f, f + df$ is defined as $w(f)df$, the average power in the band $f_{n+1} - f_n$ is

$$w(f_n)(f_{n+1} - f_n) = w\left(\frac{n}{T}\right) \frac{1}{T}$$

and, from (2.3-5), this is given by

$$\begin{aligned} w\left(\frac{n}{T}\right) \frac{1}{T} &= \frac{1}{2} (a_n^2 + b_n^2), & n \neq 0 \\ w(0) \frac{1}{T} &= \frac{a_0^2}{4}, & n = 0 \end{aligned} \tag{2.3-6}$$

When the coefficients in (2.3-4) are replaced by their expressions in terms of $w(f)$ we get

$$\begin{aligned} \frac{1}{T} \int_0^T I(t)I(t + \tau) dt + O\left(\frac{\tau I^2}{T}\right) & \\ &= \frac{1}{T} \sum_{n=0}^{\infty} w\left(\frac{n}{T}\right) \cos \frac{2\pi n\tau}{T} \\ &= \int_0^{\infty} w\left(\frac{n}{T}\right) \cos \frac{2\pi n\tau}{T} \frac{dn}{T} \\ &= \int_0^{\infty} w(f) \cos 2\pi f\tau df \end{aligned} \tag{2.3-7}$$

where we have assumed T so large and $w(f)$ of such a nature that the summation may be replaced by integration.

If I remains finite, then as $T \rightarrow \infty$ with τ held fixed, the correction term on the left becomes negligibly small and we have, upon using the definitions (2.1-4) for the correlation function $\psi(\tau)$, the second of the fundamental inversion formulas (2.1-6). The first inversion formula may be obtained from this at once by using Fourier's double integral for $w(f)$.

Incidentally, the relation (2.3-6) between $w(f)$ and the coefficients a_n and b_n is in agreement with the definition (2.1-3) for $w(f)$ as a limit involving $|S(f)|^2$. From the expressions (2.3-2) for a_n and b_n , the spectrum $S(f_n)$ given by (2.1-2) is

$$S(f_n) = \frac{T}{2} (a_n - ib_n)$$

Then, from (2.1-3) $w(f_n)$ is given by the limit, as $T \rightarrow \infty$, of

$$\begin{aligned} \frac{2}{T} |S(f_n)|^2 &= \frac{2}{T} \cdot \frac{T^2}{4} (a_n^2 + b_n^2) \\ &= \frac{T}{2} (a_n^2 + b_n^2) \end{aligned}$$

and this is the expression for $w\left(\frac{n}{T}\right)$ given by (2.3-6).

2.4 DISCUSSION OF RESULTS OF SECTION ONE—PARSEVAL'S THEOREM

The use of Parseval's theorem¹⁷ enables us to derive the results of section 2.1 more directly than the method of the preceding section. This theorem states that

$$\int_{-\infty}^{+\infty} F_1(f)F_2(f) df = \int_{-\infty}^{+\infty} G_1(t)G_2(-t) dt \tag{2.4-1}$$

where F_1, G_1 and F_2, G_2 are Fourier mates related by

$$\begin{aligned} F(f) &= \int_{-\infty}^{+\infty} G(t)e^{-i2\pi ft} dt \\ G(t) &= \int_{-\infty}^{+\infty} F(f)e^{i2\pi ft} df \end{aligned} \tag{2.4-2}$$

It may be proved in a formal manner by replacing the F_1 on the left of (2.4-1) by its expression as an integral involving $G_1(t)$. Interchanging the

¹⁷ E. C. Titchmarsh, Introduction to the Theory of Fourier Integrals, Oxford (1937).

order of integration and using the second of (2.4-2) to replace F_2 by G_2 gives the right hand side.

We now set $G_1(t)$ and $G_2(t)$ equal to zero except for intervals of length T . These intervals and the corresponding values of G_1 and G_2 are

$$G_1(t) = I(t), \quad 0 < t < T \quad (2.4-3)$$

$$G_2(t) = I(-t + \tau), \quad \tau - T < t < \tau$$

From (2.4-3) it follows that $F_1(f)$ is the spectrum $S(f)$ of $I(t)$ given by equation (2.1-2). Since $I(t)$ is real it follows from the first of equations (2.4-2) that

$$S(-f) = S^*(f), \quad (2.4-4)$$

where the star denotes conjugate complex, and hence that $|S(f)|^2$ is an even function of f .

The first of equations (2.4-2) also gives

$$\begin{aligned} F_2(f) &= \int_{\tau-T}^{\tau} I(-t + \tau) e^{-i2\pi f t} dt \\ &= \int_0^T I(t) e^{i2\pi f(t-\tau)} dt \\ &= S^*(f) e^{-i2\pi f \tau} \end{aligned} \quad (2.4-5)$$

When these G 's and F 's are placed in (2.4-1) we obtain

$$\int_{-\infty}^{+\infty} |S(f)|^2 e^{-2\pi f \tau} df = \int_0^{T-\tau} I(t) I(t + \tau) dt \quad (2.4-6)$$

where we have made use of the fact that $G_2(-t)$ is zero except in the interval $-\tau < t < T - \tau$ and have assumed $\tau > 0$. If $\tau < 0$ the limits of integration on the right would be $-\tau$ and T .

Since $|S(f)|^2$ is an even function of f we may write (2.4-6) as

$$\frac{1}{T} \int_0^T I(t) I(t + \tau) dt + O\left(\frac{\tau I^2}{T}\right) = \int_0^{\infty} \frac{2|S(f)|^2}{T} \cos 2\pi f \tau df \quad (2.4-7)$$

If we now define the correlation function $\psi(\tau)$ as the limit, as $T \rightarrow \infty$, of the left hand side and define $w(f)$ as the function

$$w(f) = \text{Limit}_{T \rightarrow \infty} \frac{2|S(f)|^2}{T}, \quad f > 0 \quad (2.1-3)$$

we obtain the second, (2.1-6), of the fundamental inversion formulas. As before, the first may be obtained from Fourier's integral theorem.

In order to obtain the interpretation of $w(f)df$ as the average power dissipated in one ohm by those components of $I(t)$ which lie in the band $f, f + df$, we set $\tau = 0$ in (2.4-7):

$$\text{Limit}_{T \rightarrow \infty} \frac{1}{T} \int_0^T I^2(t) dt = \int_0^\infty w(f) df \tag{2.4-8}$$

The expression on the left is certainly the total average power which would be dissipated in one ohm and the right hand side represents a summation over all frequencies extending from 0 to ∞ . It is natural therefore to interpret $w(f)df$ as the power due to the components in $f, f + df$.

The preceding sections have dealt with the power spectrum $w(f)$ and correlation function $\psi(\tau)$ of a very general type of function. It will be noted that a knowledge of $w(f)$ does not enable us to determine the original function. In obtaining $w(f)$, as may be seen from the definition (2.1-3) or from (2.3-6), the information carried by the phase angles of the various components of $I(t)$ has been dropped out. In fact, as we may see from the Fourier series representation (2.3-1) of $I(t)$ and from (2.3-6), it is possible to obtain an infinite number of different functions all of which have the same $w(f)$, and hence the same $\psi(\tau)$. All we have to do is to assign different sets of values to the phase angles of the various components, thereby keeping $a_n^2 + b_n^2$ constant.

2.5 HARMONIC ANALYSIS FOR RANDOM FUNCTIONS

In many applications of the theory discussed in the foregoing sections $I(t)$ is a function of t which has a certain amount of randomness associated with it. For example $I(t)$ may be a curve representing the price of wheat over a long period of years, a component of air velocity behind a grid placed in a wind tunnel, or, of primary interest here, a noise current.

In some mathematical work this randomness is introduced by considering $I(t)$ to involve a number of parameters, and then taking the parameters to be random variables. Thus, in the shot effect the arrival times t_1, t_2, \dots, t_K of the electrons were taken to be the parameters and each was assumed to be uniformly distributed over an interval $(0, T)$.

For any particular set of values of the parameters, $I(t)$ has a definite power spectrum $w(f)$ and correlation function $\psi(\tau)$. However, now the principal interest is not in these particular functions, but in functions which give the average values of $w(f)$ and $\psi(\tau)$ for fixed f and τ . These functions are obtained by averaging $w(f)$ and $\psi(\tau)$ over the ranges of the parameters, using, of course, the distribution functions of the parameters.

By averaging both sides of the appropriate equations in sections 2.1 and

2.2 it is seen that our fundamental inversion formulae (2.1-5) and (2.1-6) are unchanged. Thus,

$$\bar{w}(f) = 4 \int_0^{\infty} \bar{\psi}(\tau) \cos 2\pi f\tau \, d\tau \quad (2.5-1)$$

$$\bar{\psi}(\tau) = \int_0^{\infty} \bar{w}(f) \cos 2\pi f\tau \, df \quad (2.5-2)$$

where the bars indicate averages taken over the parameters with f or τ held constant.

The definitions of \bar{w} and $\bar{\psi}$ appearing in these equations are likewise obtained from (2.1-3) and (2.1-4)

$$\bar{w}(f) = \text{Limit}_{T \rightarrow \infty} \frac{2 \overline{|S(f)|^2}}{T} \quad (2.5-3)$$

and

$$\bar{\psi}(\tau) = \text{Limit}_{T \rightarrow \infty} \frac{1}{T} \int_0^T \overline{I(t)I(t+\tau)} \, dt \quad (2.5-4)$$

The values of t and τ are held fixed while averaging over the parameters. In (2.5-3) $S(f)$ is regarded as a function of the parameters obtained from $I(t)$ by

$$S(f) = \int_0^T I(t)e^{-2\pi ift} \, dt \quad (2.1-2)$$

Similar expressions may be obtained for the average power spectrum for d.c. and periodic components. All we need to do is to average the expression (2.2-11)

Sometimes the average value of the product $I(t)I(t+\tau)$ in the definition (2.5-4) of $\bar{\psi}(\tau)$ is independent of the time T . This enables us to perform the integration at once and obtain

$$\bar{\psi}(\tau) = \overline{I(t)I(t+\tau)} \quad (2.5-5)$$

This introduces a considerable simplification and it appears that the simplest method of computing $\bar{w}(f)$ for an $I(t)$ of this sort is first to compute $\bar{\psi}(\tau)$, and then use the inversion formula (2.5-1).

2.6 FIRST EXAMPLE—THE SHOT EFFECT

We first compute the average on the right of (2.5-5). By using the method of averaging employed many times in part I, we have

$$\overline{I(t)I(t+\tau)} = \sum_{K=0}^{\infty} p(K) \overline{I_K(t)I_K(t+\tau)} \quad (2.6-1)$$

where $p(K)$ is the probability of exactly K electrons arriving in the interval $(0, T)$,

$$p(K) = \frac{(\nu T)^K}{K!} e^{-\nu T} \tag{1.1-3}$$

and

$$I_K(t) = \sum_{k=1}^K F(t - t_k) \tag{1.3-1}$$

Multiplying $I_K(t)$ and $I_K(t + \tau)$ together and averaging t_1, t_2, \dots, t_K over their ranges gives

$$\overline{I_K(t)I_K(t + \tau)} = \sum_{k=1}^K \sum_{m=1}^K \int_0^T \frac{dt_1}{T} \cdots \int_0^T \frac{dt_K}{T} F(t - t_k)F(t + \tau - t_m)$$

This is similar to the expression for $\overline{I_K^2(t)}$ which was used in section 1.3 to prove Campbell's theorem and may be treated in much the same way. Thus, if t and $t + \tau$ lie between Δ and $T - \Delta$, the expression above becomes

$$\frac{K}{T} \int_{-\infty}^{+\infty} F(t)F(t + \tau) dt + \frac{K(K - 1)}{T^2} \left[\int_{-\infty}^{+\infty} F(t) dt \right]^2$$

When this is placed in (2.6-1) and the summation performed we obtain an expression independent of T . Consequently we may use (2.5-5) and get

$$\bar{\psi}(\tau) = \nu \int_{-\infty}^{+\infty} F(t)F(t + \tau) dt + \overline{I(t)}^2 \tag{2.6-2}$$

where we have used the expression for the average current

$$\overline{I(t)} = \nu \int_{-\infty}^{+\infty} F(t) dt \tag{1.3-4}$$

In order to compute $\bar{w}(f)$ from $\bar{\psi}(\tau)$ it is convenient to make use of the fact that $\psi(\tau)$ is always an even function of τ and hence (2.5-1) may also be written as

$$\bar{w}(f) = 2 \int_{-\infty}^{+\infty} \bar{\psi}(\tau) \cos 2\pi f\tau d\tau \tag{2.6-3}$$

Then

$$\begin{aligned} \bar{w}(f) &= 2\nu \int_{-\infty}^{+\infty} dt F(t) \int_{-\infty}^{+\infty} d\tau F(t + \tau) \cos 2\pi f\tau \\ &\quad + 2 \int_{-\infty}^{+\infty} \frac{\overline{I(t)}^2}{T} \cos 2\pi f\tau d\tau \end{aligned}$$

$$\begin{aligned}
&= 2\nu \text{ Real Part of } \int_{-\infty}^{+\infty} dt F(t)e^{-2\pi ift} \int_{-\infty}^{+\infty} dt' F(t')e^{2\pi ift'} \\
&\quad + 2\overline{I(t)}^2 \int_{-\infty}^{+\infty} e^{i2\pi f\tau} d\tau \\
&= 2\nu |s(f)|^2 + 2\overline{I(t)}^2 \delta(f)
\end{aligned} \tag{2.6-4}$$

In going from the first equation to the second we have written $t' = t + \tau$ and have considered $\cos 2\pi f\tau$ to be the real part of the corresponding exponential. In going from the second equation to the third we have set

$$s(f) = \int_{-\infty}^{+\infty} F(t)e^{-2\pi ift} dt \tag{2.6-5}$$

and have used

$$\int_{-\infty}^{+\infty} e^{i2\pi ft} dt = \delta(f) \tag{2.2-9}$$

The term in $\bar{w}(f)$ involving $\delta(f)$ represents the average power which would be dissipated by the d.c. component of $I(t)$ in flowing through one ohm. It is in agreement with the concept that the average power in the band $0 \leq f < \epsilon$, $\epsilon > 0$ but very small, is

$$\begin{aligned}
\int_0^\epsilon \bar{w}(f) dt &= 2\overline{I(t)}^2 \int_0^\epsilon \delta(f) df \\
&= \overline{I(t)}^2
\end{aligned} \tag{2.6-6}$$

The expression (2.6-4) for $\bar{w}(f)$ may also be obtained from the definition (2.5-3) for $\bar{w}(f)$ plus the additional term due to the d.c. component obtained by averaging the expressions (2.2-11). We leave this as an exercise for the reader. He will find it interesting to study the steps in Carson's¹⁵ paper leading up to equation (8). Carson's $R(\omega)$ is related to our $\bar{w}(f)$ by

$$\bar{w}(f) = 2\pi R(\omega)$$

and his $f(i\omega)$ is equal to our $s(f)$.

Integrating both sides of (2.6-4) with respect to f from 0 to ∞ and using

$$\bar{I}^2 = \int_0^\infty \bar{w}(f) df$$

gives the result

$$\bar{I}^2 - \overline{I}^2 = 2\nu \int_0^{+\infty} |s(f)|^2 df \tag{2.6-7}$$

¹⁵ Loc. cit.

This may be obtained immediately from Campbell's theorem by applying Parseval's theorem.

As an example of the use of these formulas we derive the power spectrum of the voltage across a resistance R when a current consisting of a great number of very short pulses per second flows through R . Let $F(t - t_k)$ be the voltage produced by the pulse occurring at time t_k . Then

$$F(t) = R\varphi(t)$$

where $\varphi(t)$ is the current in the pulse. We confine our interest to relatively low frequencies such that we may make the approximation

$$\begin{aligned} s(f) &= \int_{-\infty}^{+\infty} R\varphi(t)e^{-2\pi ift} dt \\ &\approx R \int_{-\infty}^{+\infty} \varphi(t) dt = Rq \end{aligned}$$

where q is the charge carried through the resistance by one pulse. From (2.6-4) it follows that for these low frequencies the continuous portion of the power spectrum for the voltage is constant and equal to

$$\bar{w}(f) = 2\nu R^2 q^2 = 2\bar{I}R^2 q \tag{2.6-8}$$

where $\bar{I} = \nu q$ is the average current flowing through R . This result is often used in connection with the shot effect in diodes.

In the study of the shot effect it was assumed that the probability of an event (electron arriving at the anode) happening in dt was νdt where ν is the expected number of events per second. This probability is independent of the time t . Sometimes we wish to introduce dependency on time.¹⁸ As an example, consider a long interval extending from 0 to T . Let the probability of an event happening in $t, t + dt$ be $\bar{K}p(t)dt$ where \bar{K} is the average number of events during T and $p(t)$ is a given function of t such that

$$\int_0^T p(t) dt = 1$$

For the shot effect $p(t) = 1/T$.

What is the probability that exactly K events happen in T ? As in the case of the shot effect, section 1.1, we may divide $(0, T)$ into N intervals each of length Δt so that $N\Delta t = T$. The probability of no event happening in the first Δt is

$$1 - \bar{K}p\left(\frac{\Delta t}{2}\right)\Delta t$$

¹⁸ A careful discussion of this subject is given by Hurwitz and Kac in "Statistical Analysis of Certain Types of Random Functions." I understand that this paper will soon appear in the *Annals of Math. Statistics*.

The product of N such probabilities is, as $N \rightarrow \infty$, $\Delta t \rightarrow 0$;

$$\exp \left[-\bar{K} \int_0^T p(t) dt \right] = e^{-\bar{K}}$$

This is the probability that exactly 0 events happen in T . In the same way we are led to the expression

$$\frac{\bar{K}^K}{K!} e^{-\bar{K}} \quad (2.6-9)$$

for the probability that exactly K events happen in T .

When we consider many intervals $(0, T)$ we obtain many values of K and also many values of I measured t seconds from the beginning of each interval. These values of I define the distribution of I at time t . By proceeding as in section 1.4 we find that the probability density of I is

$$P(I, t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} du \exp \left[-iuI + \bar{K} \int_0^T p(x) (e^{iuF(t-x)} - 1) dx \right]$$

The corresponding average and variance is

$$\begin{aligned} \bar{I} &= \bar{K} \int_0^T p(x) F(t-x) dx \\ \overline{(I - \bar{I})^2} &= \bar{K} \int_0^T p(x) F^2(t-x) dx \end{aligned} \quad (2.6-10)$$

If $S(f)$ is given by (2.1-2) and $s(f)$ by (2.6-5) (assuming the duration of $F(t)$ short in comparison with T) the average value of $|S(f)|^2$ may be obtained by putting (1.3-1) in (2.1-2) to get

$$S_K(f) = s(f) \sum_1^K e^{-2\pi i f t_k}$$

Expressing $S_K(f)$ $S_K^*(f)$, where the star denotes conjugate complex, as a double sum and averaging over the t_k 's, using $p(t)$, and then averaging over the K 's gives

$$\overline{|S(f)|^2} = \bar{K} |s(f)|^2 \left[1 + \bar{K} \left| \int_0^T p(x) e^{-2\pi i f x} dx \right|^2 \right] \quad (2.6-11)$$

This may be used to compute the power spectrum from (2.5-3) provided $p(x)$ is not periodic. If $p(x)$ is periodic then the method of section 2.2 should be used at the harmonic frequencies. If the fluctuations of $p(t)$ are slow in comparison with the fluctuations of $F(t)$ the second term within the brackets of (2.6-11) may generally be neglected since there are no values of

f which make both it and $s(f)$ large at the same time. On the other hand, if both $p(t)$ and $F(t)$ fluctuate at about the same rate this term must be considered.

2.7 SECOND EXAMPLE—RANDOM TELEGRAPH SIGNAL¹⁶

Let $I(t)$ be equal to either a or $-a$ so that it is of the form of a flat top wave. Let the intervals between changes of sign, i.e. the lengths of the tops and bottoms, be distributed exponentially. We are led to this distribution by assuming that, if on the average there are μ changes of sign per second, the probability of a change of sign in $t, t + dt$ is μdt and is independent of what happens outside the interval $t, t + dt$. From the same sort of reasoning as employed in section 1.1 for the shot effect we see that the probability of obtaining exactly K changes of sign in the interval $(0, T)$ is

$$p(K) = \frac{(\mu T)^K}{K!} e^{-\mu T} \tag{2.7-1}$$

We consider the average value of the product $I(t)I(t + \tau)$. This product is a^2 if the two I 's are of the same sign and is $-a^2$ if they are of opposite sign. In the first case there are an even number, including zero, of changes of sign in the interval $(t, t + \tau)$, and in the second case there are an odd number of changes of sign. Thus

$$\begin{aligned} &\text{Average value of } I(t)I(t + \tau) && (2.7-2) \\ &= a^2 \times \text{probability of an even number of} \\ &\quad \text{changes of sign in } t, t + \tau \\ &- a^2 \times \text{probability of an odd number of} \\ &\quad \text{changes of sign in } t, t + \tau \end{aligned}$$

The length of the interval under consideration is $|t + \tau - t| = |\tau|$ seconds. Since, by assumption, the probability of a change of sign in an elementary interval of length Δt is independent of what happens outside that interval, it follows that the same is true of any interval irrespective of when it starts. Hence the probabilities in (2.7-2) are independent of t and may be obtained from (2.7-1) by setting $T = |\tau|$. Then (2.7-2) becomes, assuming $\tau > 0$ for the moment,

$$\begin{aligned} \overline{I(t)I(t + \tau)} &= a^2[p(0) + p(2) + p(4) + \dots] \\ &\quad - a^2[p(1) + p(3) + p(5) + \dots] \\ &= a^2 e^{-\mu\tau} \left[1 - \frac{\mu\tau}{1!} + \frac{(\mu\tau)^2}{2!} - \dots \right] \\ &= a^2 e^{-2\mu\tau} \tag{2.7-3} \end{aligned}$$

¹⁶ Kenrick, cited in Section 2.2.

From (2.5-5), this gives the correlation function for $I(t)$

$$\bar{\psi}(\tau) = a^2 e^{-2\mu|\tau|} \quad (2.7-4)$$

The corresponding power spectrum is, from (2.5-1),

$$\begin{aligned} \bar{w}(f) &= 4a^2 \int_0^{\infty} e^{-2\mu\tau} \cos 2\pi f\tau \, d\tau \\ &= \frac{2a^2 \mu}{\pi^2 f^2 + \mu^2} \end{aligned} \quad (2.7-5)$$

Correlation functions and power spectra of this type occur quite frequently. In particular, they are of use in the study of turbulence in hydrodynamics. We may also obtain them from our shot effect expressions if we disregard the d.c. component. All we have to do is to assume that the effect $F(t)$ of an electron arriving at the anode at time $t = 0$ is zero for $t < 0$, and that $F(t)$ decays exponentially with time after jumping to its maximum value at $t = 0$. This may be verified by substituting the value

$$F(t) = 2a \sqrt{\frac{\mu}{v}} e^{-2\mu t}, \quad t > 0 \quad (2.7-6)$$

for $F(t)$ in the expressions (2.6-2) and (2.6-4) (after using 2.6-5) for the correlation function and energy spectrum of the shot effect.

The power spectrum of the current flowing through an inductance and a resistance in series in response to a very wide band thermal noise voltage is also of the form (2.7-5).

Incidentally, this gives us an example of two quite different $I(t)$'s, one a flat top wave and the other a shot effect current, which have the same correlation functions and power spectra, aside from the d.c. component.

There is another type of random telegraph signal which is interesting to analyze. The time scale is divided into intervals of equal length h . In an interval selected at random the value of $I(t)$ is independent of the values in the other intervals, and is equally likely to be $+a$ or $-a$. We could construct such a wave by flipping a penny. If heads turned up we would set $I(t) = a$ in $0 < t < h$. If tails were obtained we would set $I(t) = -a$ in this interval. Flipping again would give either $+a$ or $-a$ for the second interval $h < t < 2h$, and so on. This gives us one wave. A great many waves may be constructed in this way and we denote averages over these waves, with t held constant, by bars.

We ask for the average value of $I(t)I(t + \tau)$, assuming $\tau > 0$. First we note that if $\tau > h$ the currents correspond to different intervals for all

values of t . Since the values in these intervals are independent we have

$$\overline{I(t)I(t + \tau)} = \overline{I(t)} \overline{I(t + \tau)} = 0$$

for all values of t when $\tau > h$.

To obtain the average when $\tau < h$ we consider t to lie in the first interval $0 < t < h$. Since all the intervals are the same from a statistical point of view we lose no generality in doing this. If $t + \tau < h$, i.e., $t < h - \tau$, both currents lie in the first interval and

$$\overline{I(t)I(t + \tau)} = a^2$$

If $t > h - \tau$ the current $I(t + \tau)$ corresponds to the second interval and hence the average value is zero.

We now return to (2.5-4). The integral there extends from 0 to T . When $\tau > h$, the integrand is zero and hence

$$\bar{\psi}(\tau) = 0, \quad \tau > h \tag{2.7-7}$$

When $\tau < h$, our investigation of the interval $0 < t < h$ enables us to write down the portion of the integral extending from 0 to h :

$$\begin{aligned} \int_0^h I(t)I(t + \tau) dt &= \int_0^{h-\tau} a^2 dt + \int_{h-\tau}^h 0 dt \\ &= a^2(h - \tau) \end{aligned}$$

Over the interval of integration $(0, T)$ we have T/h such intervals each contributing the same amount. Hence, from (2.5-4),

$$\begin{aligned} \bar{\psi}(\tau) &= \text{Limit}_{T \rightarrow \infty} \frac{a^2}{T} \cdot \frac{T}{h} (h - \tau) \\ &= a^2 \left(1 - \frac{\tau}{h} \right), \quad 0 \leq \tau < h \end{aligned} \tag{2.7-8}$$

The power spectrum of this type of telegraph wave is thus

$$\begin{aligned} \bar{w}(f) &= 4a^2 \int_0^h \left(1 - \frac{\tau}{h} \right) \cos 2\pi f \tau d\tau \\ &= 2h \left(\frac{a \sin \pi fh}{\pi fh} \right)^2 \end{aligned} \tag{2.7-9}$$

This is seen to have the same general behavior as $\bar{w}(f)$ for the first type of telegraph signal given by (2.7-5), when we relate the average number, μ , of changes of sign per second to the interval length h by $\mu h = 1$.

2.8 REPRESENTATION OF NOISE CURRENT

In section 1.8 the Fourier series representation of the shot effect current was discussed. This suggests the representation*

$$I(t) = \sum_{n=1}^N (a_n \cos \omega_n t + b_n \sin \omega_n t) \quad (2.8-1)$$

where

$$\omega_n = 2\pi f_n, \quad f_n = n\Delta f \quad (2.8-2)$$

a_n and b_n are taken to be independent random variables which are distributed normally about zero with the standard deviation $\sqrt{w(f_n)\Delta f}$. $w(f)$ is the power spectrum of the noise current, i.e., $w(f) df$ is the average power which would be dissipated by those components of $I(t)$ which lie in the frequency range $f, f + df$ if they were to flow through a resistance of one ohm.

The expression for the standard deviation of a_n and b_n is obtained when we notice that Δf is the width of the frequency band associated with the n th component. Hence $w(f_n)\Delta f$ is the average energy which would be dissipated if the current

$$a_n \cos \omega_n t + b_n \sin \omega_n t$$

were to flow through a resistance of one ohm, this average being taken over all possible values of a_n and b_n . Thus

$$w(f_n)\Delta f = a_n^2 \cos^2 \omega_n t + \overline{2a_n b_n \cos \omega_n t \sin \omega_n t} + \overline{b_n^2 \sin^2 \omega_n t} = \overline{a_n^2} = \overline{b_n^2} \quad (2.8-3)$$

The last two steps follow from the independence of a_n and b_n and the identity of their distributions. It will be observed that $w(f)$, as used with the representation (2.8-1), is the same sort of average as was denoted in section 2.5 by $\bar{w}(f)$. However, $w(f)$ is often given to us in order to specify the spectrum of a given noise.

For example, suppose we are interested in the output of a certain filter when a source of thermal noise is applied to the input. Let $A(f)$ be the absolute value of the ratio of the output current to the input current when a steady sinusoidal voltage of frequency f is applied to the input. Then

$$w(f) = cA^2(f) \quad (2.8-4)$$

* As mentioned in section 1.7 this sort of representation was used by Einstein and Hopf for radiation. Shottky (1918) used (2.8-1), apparently without explicitly taking the coefficients to be normally distributed. Nyquist (1932) derived the normal distribution from the shot effect.

If W is the average power dissipated in one ohm by $I(t)$,

$$\begin{aligned} W &= \text{Limit}_{T \rightarrow \infty} \frac{1}{T} \int_0^T I^2(t) dt = \int_0^\infty w(f) df \\ &= c \int_0^\infty A^2(f) df \end{aligned} \tag{2.8-5}$$

which is an equation to determine c when W and $A(f)$ are known.

In using the representation (2.8-1) to investigate the statistical properties of $I(t)$ we first find the corresponding statistical properties of the summation on the right when the a 's and b 's are regarded as random variables distributed as mentioned above and t is regarded as fixed. In general, the time t disappears in this procedure just as it did in (2.8-3). We then let $N \rightarrow \infty$ and $\Delta f \rightarrow 0$ so that the summations may be replaced by integrations. Finally, the frequency range is extended to cover all frequencies from 0 to ∞ .

The usual way of looking at the representation (2.8-1) is to suppose that we have an oscillogram of $I(t)$ extending from $t = 0$ to $t = \infty$. This oscillogram may be cut up into strips of length T . A Fourier analysis of $I(t)$ for each strip will give a set of coefficients. These coefficients will vary from strip to strip. Our representation ($T\Delta f = 1$) assumes that this variation is governed by a normal distribution. Our process for finding statistical properties by regarding the a 's and b 's as random variables while t is kept fixed corresponds to examining the noise current at a great many instants. Corresponding to each strip there is an instant, and this instant occurs at t (this is the t in (2.8-1)) seconds from the beginning of the strip. This is somewhat like examining the noise current at a great number of instants selected at random.

Although (2.8-1) is the representation which is suggested by the shot effect and similar phenomena, it is not the only representation, nor is it always the most convenient. Another representation which leads to the same results when the limits are taken is¹⁹

$$I(t) = \sum_{n=1}^N c_n \cos(\omega_n t - \varphi_n) \tag{2.8-6}$$

where $\varphi_1, \varphi_2, \dots, \varphi_N$ are angles distributed at random over the range $(0, 2\pi)$ and

$$c_n = [2w(f_n)\Delta f]^{1/2}, \quad \omega_n = 2\pi f_n, \quad f_n = n\Delta f \tag{2.8-7}$$

¹⁹ This representation has often been used by W. R. Bennett in unpublished memoranda written in the 1930's.

In this representation $I(t)$ is regarded as the sum of a number of sinusoidal components with fixed amplitudes but random phase angles.

That the two different representations (2.8-1) and (2.8-6) of $I(t)$ lead to the same statistical properties is a consequence of the fact that they are always used in such a way that the "central limit theorem*" may be used in both cases.

This theorem states that under certain general conditions, the distribution of the sum of N random vectors approaches a normal law (it may be normal in several dimensions**) as $N \rightarrow \infty$. In fact from this theorem it appears that a representation such as

$$I(t) = \sum_{n=1}^N (a_n \cos \omega_n t + b_n \sin \omega_n t) \quad (2.8-6)$$

where a_n and b_n are independent random variables which take only the values $\pm [w(f_n)\Delta f]^{1/2}$, the probability of each value being $\frac{1}{2}$, will lead in the limit to the same statistical properties of $I(t)$ as do (2.8-1) and (2.8-6).

2.9 THE NORMAL DISTRIBUTION IN SEVERAL VARIABLES²⁰

Consider a random vector r in K dimensions. The distribution of this vector may be specified by stating the distribution of the K components, x_1, x_2, \dots, x_K , of r . r is said to be normally distributed when the probability density function of the x 's is of the form

$$(2\pi)^{-K/2} |M|^{-1/2} \exp \left[-\frac{1}{2} x' M^{-1} x \right] \quad (2.9-1)$$

where the exponent is a quadratic form in the x 's. The square matrix M is composed of the second moments of the x 's.

$$M = \begin{bmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{1K} \\ \cdot & \cdot & \cdot & \cdot \\ \mu_{1K} & \cdots & \mu_{KK} \end{bmatrix} \quad (2.9-2)$$

where the second moments are defined by

$$\mu_{11} = \overline{x_1^2}, \quad \mu_{12} = \overline{x_1 x_2}, \quad \text{etc.} \quad (2.9-3)$$

$|M|$ represents the determinant of M and x' is the row matrix

$$x' = [x_1, x_2, \dots, x_K] \quad (2.9-4)$$

x is the column matrix obtained by transposing x' .

* See section 2.10.

** See section 2.9.

²⁰ H. Cramér, "Random Variables and Probability Distributions," Chap. X., Cambridge Tract No. 36 (1937).

The exponent in the expression (2.9-1) for the probability density may be written out by using

$$x' M^{-1} x = \sum_{r=1}^K \sum_{s=1}^K \frac{M_{rs}}{|M|} x_r x_s \tag{2.9-5}$$

where M_{rs} is the cofactor of μ_{rs} in M .

Sometimes there are linear relations between the x 's so that the random vector r is restricted to a space of less than K dimensions. In this case the appropriate form for the density function may be obtained by considering a sequence of K -dimensional distributions which approach the one being investigated.

If r_1 and r_2 are two normally distributed random vectors their sum $r_1 + r_2$ is also normally distributed. It follows that the sum of any number of normally distributed random vectors is normally distributed.

The characteristic function of the normal distribution is

$$\text{ave. } e^{iz_1x_1+iz_2x_2+\dots+iz_Kx_K} = \exp \left[-\frac{1}{2} \sum_{r=1}^K \sum_{s=1}^K \mu_{rs} z_r z_s \right] \tag{2.9-6}$$

2.10 CENTRAL LIMIT THEOREM

The central limit theorem in probability states that the distribution of the sum of N independent random vectors $r_1 + r_2 + \dots + r_N$ approaches a normal law as $N \rightarrow \infty$ when the distributions of r_1, r_2, \dots, r_N satisfy certain general conditions.⁷

As an example we take the case in which r_1, r_2, \dots are two-dimensional vectors²¹, the components of r_n being x_n and y_n . Without loss of generality we assume that

$$\bar{x}_n = 0, \quad \bar{y}_n = 0.$$

The components of the resultant vector are

$$\begin{aligned} X &= x_1 + x_2 + \dots + x_N \\ Y &= y_1 + y_2 + \dots + y_N \end{aligned} \tag{2.10-1}$$

and, since r_1, r_2, \dots are independent vectors, the second moments of the resultant are

$$\begin{aligned} \mu_{11} &= \overline{X^2} = \overline{x_1^2} + \overline{x_2^2} + \dots + \overline{x_N^2} \\ \mu_{22} &= \overline{Y^2} = \overline{y_1^2} + \overline{y_2^2} + \dots + \overline{y_N^2} \\ \mu_{12} &= \overline{XY} = \overline{x_1y_1} + \overline{x_2y_2} + \dots + \overline{x_Ny_N} \end{aligned} \tag{2.10-2}$$

⁷ Incidentally, von Laue (see references in section 1.7) used this theorem in discussing the normal distribution of the coefficients in a Fourier series used to represent black-body radiation. He ascribed it to Markoff.

²¹ This case is discussed by J. V. Uspensky, "Introduction to Mathematical Probability", McGraw-Hill (1937) Chap. XV.

Apparently there are several types of conditions which are sufficient to ensure that the distribution of the resultant approaches a normal law. One sufficient condition is that²¹

$$\begin{aligned} \mu_{11}^{-3/2} \sum_{n=1}^N |x_n|^3 &\rightarrow 0 \\ \mu_{22}^{-3/2} \sum_{n=1}^N |y_n|^3 &\rightarrow 0 \end{aligned} \quad (2.10-3)$$

The central limit theorem tells us that the distribution of the random vector (X, Y) approaches a normal law as $N \rightarrow \infty$. The second moments of this distribution are given by (2.10-2). When we know the second moments of a normal distribution we may write down the probability density function at once. Thus from section 2.9

$$M = \begin{bmatrix} \mu_{11} & \mu_{12} \\ \mu_{12} & \mu_{22} \end{bmatrix}, \quad M^{-1} = |M|^{-1} \begin{bmatrix} \mu_{22} & -\mu_{12} \\ -\mu_{12} & \mu_{11} \end{bmatrix}$$

$$|M| = \mu_{11}\mu_{22} - \mu_{12}^2$$

$$x' = [X, Y]$$

$$x'M^{-1}x = |M|^{-1}(\mu_{22}X^2 - 2\mu_{12}XY + \mu_{11}Y^2)$$

The probability density is therefore

$$\frac{(\mu_{11}\mu_{22} - \mu_{12}^2)^{-1/2}}{2\pi} \exp \left[\frac{-\mu_{22}X^2 - \mu_{11}Y^2 + 2\mu_{12}XY}{2(\mu_{11}\mu_{22} - \mu_{12}^2)} \right] \quad (2.10-3)$$

Incidentally, the second moments are related to the standard deviations σ_1, σ_2 of X, Y and to the correlation coefficient τ of X and Y by

$$\mu_{11} = \sigma_1^2, \quad \mu_{22} = \sigma_2^2, \quad \mu_{12} = \tau\sigma_1\sigma_2 \quad (2.10-4)$$

and the probability density takes the standard form

$$\frac{(1 - \tau^2)^{-1/2}}{2\pi\sigma_1\sigma_2} \exp \left[-\frac{1}{2(1 - \tau^2)} \left(\frac{X^2}{\sigma_1^2} - 2\tau \frac{XY}{\sigma_1\sigma_2} + \frac{Y^2}{\sigma_2^2} \right) \right] \quad (2.10-5)$$

²¹ This is used by Uspensky, loc. cit. Another condition analogous to the Lindeberg condition is given by Cramer,²⁰ loc. cit.

(To be concluded)

Abstracts of Technical Articles by Bell System Authors

*Crossbar Toll Switching System.*¹ L. G. ABRAHAM, A. J. BUSCH, AND F. F. SHIPLEY. A new crossbar toll switching system was placed in service in Philadelphia in August 1943. Important improvements offered by this system include:

1. Transmission objectives are met more readily and substantial economies are obtained in outside plant and in repeater equipment.
2. Extended use of toll dialing results in operating economies and improved service to subscribers. Calls over circuits still on a ringdown basis are also handled more expeditiously and with operating economies.
3. Flexibility due to the use of sender and markers to control establishing connections will be useful in meeting future toll switching requirements.

*Low-Frequency Quartz-Crystal Cuts having Low Temperature Coefficients.*² W. P. MASON AND R. A. SYKES. This paper discusses low-frequency, low-temperature-coefficient crystals which are suitable for use in filters and oscillators in the frequency range from 4 to 100 kilocycles. Two new cuts, the MT and NT, are described. These are related to the +5-degree *X*-cut crystal, which is the quartz crystal having the lowest temperature coefficient for any orientation of a bar cut from the natural crystal. When the width of the +5-degree *X*-cut crystal approaches in value the length, the motion has a shear component, and this introduces a negative temperature coefficient which causes the temperature coefficient of the crystal to become increasingly negative as the ratio of width to length increases.

The MT crystal has its length along nearly the same axis as the +5-degree *X*-cut crystal, but its major surface is rotated by 35 to 50 degrees around the length axis. This results in giving the shear component a zero or positive temperature coefficient and results in a crystal with a uniformly low temperature coefficient independent of the width-length ratio. A slightly higher rotation about the length axis results in a crystal which has a low-temperature coefficient when vibrating in flexure and this crystal has been called the NT crystal. The NT crystal can be used in a frequency range from 4 to 50 kilocycles, while the MT is useful from 50 kilocycles to 500 kilocycles.

A special oscillator circuit is described which can drive a high-impedance NT flexure crystal. This circuit, together with the NT crystal, has been used to control the mean frequency of the Western Electric frequency-modulated radio transmitter.

¹*Elec. Engg., Transactions Section*, June, 1944.

²*Proc. I. R. E.*, April 1944.

*Electronics; Today and Tomorrow.*³ JOHN MILLS. John Mills, scientist, teacher, author, telephone and radio engineer, was first introduced to the electron while still a fledgling under the tutelage of that eminent scientist R. A. Millikan at Chicago University. That was before Millikan had, to quote the author, "first put salt on its tail." Electronics, as a special branch of physics and engineering, has come to adulthood under the eyes of this observant author. From intimate association with this and allied fields John Mills writes knowingly. It is a book intended for the intelligent layman rather than for the expert.

"Electronics; Today and Tomorrow" likewise contains much of interest concerning the electronics of yesterday but, as one would expect, very little about the electronics of tomorrow because, as the author points out, much of this "should await the victorious conclusion of the present conflict." The book generally presents an interesting introduction to many things electronic and throughout is interspersed with examples of present day techniques which employ electronic devices. All line drawings, diagrams and photographs have been omitted.

The author begins his latest book with a brief capitulation of familiar engineering applications such as long-distance telephony, broadcasting, sound motion pictures and television which have been achieved through the use of electronic devices. He goes on to an historical account of some underlying discoveries and a discussion of atomic structure. An introduction to static electricity with the classical concepts of positive and negative charge is followed by establishment of the ideas of charges in motion, electrical current, discharges in gases, X-rays and their generation.

The remainder of the book is divided into Part I—Electron Tubes, and Part II—Electronic Devices. In Part I the author follows the development of the art in nearly chronological order from diodes through the modern multi-electrode tubes giving uses and applications of each and explaining the purpose of the grids introduced. Part II discusses more complicated structures such as cathode-ray tubes, kinescopes, iconoscopes, electron microscopes, kodatrons, magnetrons, rumbatrons, klystrons, etc. and elaborates upon their practical applications. This is, in fact, a wondrous field of developments. Finally there is a chapter on cyclotrons which the author says "comes into this book because it requires for its operation a powerful oscillator—or oscillator plus powerful electron amplifier—which will supply a high voltage at a frequency of megacycles." While many industrial applications, such perhaps as tin plating, might be included on the same basis, nevertheless the discussion of the cyclotron does give an opportunity to explain the concepts of modern physics more completely than was undertaken earlier in the book and is very interesting reading.

³Published by D. Van Nostrand Company, Inc., New York, N. Y., 1944.

*Impedance Concept in Wave Guides.*⁴ S. A. SCHELKUNOFF. The impedance concept is the foundation of engineering transmission theory. If wave guides are to be fully utilized as transmission systems or parts thereof, their properties must be expressed in terms of appropriately chosen impedances or else a new transmission theory must be developed. The gradual extension of the concept has necessitated a broader point of view without which an exploitation of its full potentialities would be impossible.

In the course of various private discussions it has been noted that there exists some uneasiness with regard to the applicability of the concept at very high frequencies. In part this may be attributed to relative unfamiliarity with the wave guide phenomena and in part to the evolution of the concept itself. Some particular aspects of the concept have to be sacrificed in the process of generalization and although these aspects may be logically unimportant, they frequently become psychological obstacles to understanding in the early stages of the development. For this reason several sections of this paper are devoted to a general discussion of the impedance concept before more specific applications are given; then by way of illustration it is proved that an infinitely thin perfectly conducting iris between two *different* wave guides behaves as if between the admittances of its faces there existed an ideal transformer. This theorem is a generalization of another theorem to the effect that when the two wave guides are alike, the iris behaves as a shunt reactor. Actual calculation of the admittances and the transformer ratio depends on the solution of an appropriate boundary value problem.

More generally, wave guide discontinuities are representable by *T*-networks. In some special cases these networks lack series branches, and in other cases the shunt branch.

*Theory of Cathode Sputtering in Low Voltage Gaseous Discharges.*⁵ CHARLES HARD TOWNES. To determine the amount of sputtering in a glow discharge three functions must be known: the number of ions of a given energy striking the cathode, the amount of cathode material released from the cathode by each ion, and the fraction of material released from the cathode which diffuses away. Expressions derived for these allow determination of the dependence of total rate of sputtering on the geometry of the discharge, pressure, cathode fall, current, and constants of the gas and cathode surface. The result is most accurate for very low voltage, high pressure discharges. Comparison with experimental data shows quantitative agreement under these conditions.

⁴*Quarterly of Applied Mathematics*, April 1944.

⁵*Phys. Rev.*, June 1 and 15, 1944.

Contributors to this Issue

A. R. D'HEEDENE, B.S. in Mechanical Engineering, New York University, 1924; Bell Telephone Laboratories, 1924-. Mr. D'heedene has been engaged in the development of wave filters, particularly filters using quartz crystal units or operating at very high frequencies.

R. M. C. GREENIDGE, B.S. in Mechanical Engineering, Harvard University, 1924. Engineering Department, Western Electric Company, Inc., 1924-25; Bell Telephone Laboratories, Inc., 1925-. Engaged in the development of crystal units for filters and oscillators.

WILTON T. REA, B.S. in Physics, Princeton University, 1926. American Telephone and Telegraph Company, Department of Development and Research, 1926-1934; Bell Telephone Laboratories, Inc. 1934-. Prior to 1941 Mr. Rea was concerned with telegraph development problems. Since then the group of which he is in charge has been engaged full time on war projects.

S. O. RICE, B.S. in Electrical Engineering, Oregon State College, 1929; California Institute of Technology, 1929-30, 1934-35. Bell Telephone Laboratories, 1930-. Mr. Rice has been concerned with various theoretical investigations relating to telephone transmission theory.

THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS
OF ELECTRICAL COMMUNICATION

The Conquest of Distance by Wire Telephony
Thomas Shaw 337

Some Aspects of Powder Metallurgy
Earle E. Schumacher and Alexander G. Souden 422

Abstracts of Technical Articles by Bell System Authors 458

Contributors to this Issue 460

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

THE BELL SYSTEM TECHNICAL JOURNAL

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



EDITORS

R. W. King

J. O. Perrine

EDITORIAL BOARD

M. R. Sullivan

O. E. Buckley

O. B. Blackwell

M. J. Kelly

H. S. Osborne

A. B. Clark

J. J. Pilliod

S. Bracken



SUBSCRIPTIONS

Subscriptions are accepted at \$1.50 per year. Single copies are 50 cents each.
The foreign postage is 35 cents per year or 9 cents per copy.



Copyright, 1944
American Telephone and Telegraph Company



FRANK B. JEWETT

The Bell System Technical Journal

Vol. XXIII

October, 1944

No. 4

THE CONQUEST OF DISTANCE BY WIRE TELEPHONY

A Story of Transmission Development From the Early Days of Loading To the Wide Use of Thermionic Repeaters

By THOMAS SHAW

EDITORIAL FOREWORD

SOME few months ago, in anticipation of the retirement of Dr. F. B. Jewett, an informal committee undertook to discover such action as the Journal might appropriately take to commemorate the event. The various possibilities finally narrowed down to one, a historical review appearing for various reasons to be the most suitable.

The period to be covered by the review was not difficult to fix. For sentimental reasons its beginning should naturally tally with Dr. Jewett's appearance upon the scene of telephone engineering, but as this followed close upon the invention of the loading coil, such a beginning had more than sentiment to recommend it.

The review is carried through the creation of the high vacuum tube to the demonstration by large scale practical application that this was the keystone of an art which would open up a new era in transmission of the voice. An examination of the record shows that the last twenty-five years of the art of telephone engineering have been adequately chronicled from year to year, almost from month to month, in the technical press. The immediately preceding period of approximately fifteen years covered by the review was badly in need of a historian in spite of the fact that in some respects the events of those years were as significant as any that have occurred subsequently.

Such considerations led to the decision to record these events while the story as it stood in the minds of certain of the chief participants was readily available. But while a committee may reach a decision, it is likely to prove a poor instrument of accomplishment. In consequence, the task of compiling the history has fallen upon the shoulders of a single individual, and we believe a very competent one. Mr. Shaw is to be congratulated in capturing to an unusual degree the spirit of the period which intervenes between the introduction of the loading coil and the completion of the first transcontinental line. He has compiled his history only after a painstaking review of the written record and many interviews with its surviving prin-

principal actors. Needless to say, he has been aided by the fact that he was, himself, a participant in much that he relates.

As in every effort of this sort, it has been necessary to set up rather definite boundaries in advance. The decision was reached deliberately to confine the present discussion to the broad phases of telephone transmission, with very little reference to the concomitant, and indeed related, developments and improvements such as occurred in the domains of substation apparatus, central office equipment, and operating methods.

Without striving for effect, and without forsaking a simple and easily comprehended engineering vernacular, Mr. Shaw makes the reader sense the momentous nature of the work in progress and the basic importance of the decisions under discussion. One sees in a new light, as he reads, the difficulties which were patiently but determinedly overcome in creating the first successful loaded phantom open-wire circuit; in reducing the crosstalk unbalance in early cables; in evaluating the relative merits of the multiple-twin and the spiral-four; and in obtaining the balancing networks needed to operate repeaters in tandem on a very long line. And, of course, there are other matters too numerous to mention here which are similarly dealt with. All in all, it is a recital whose simplicity is in sharp contrast to the intricate nature of much of the work it narrates.

And as one lays it down he feels a strongly renewed admiration for the executives who visioned, guided, counselled, and in the days of rough going had the courage to back their judgments, and more especially that of their engineers, to a magnificent extent.

We are now on the point of losing by retirement one of the best loved of these executives. Starting as a young recruit in 1904, through outstanding merit he was destined to rise so rapidly in responsibility as to become a very influential counsellor within a few years, and ever since has occupied a commanding position with respect to the Bell System's entire research and development program. No individual is more intimately associated with the scientific achievements of the System throughout the last forty years, in the minds both of the public at large and those within our organization. Under the circumstances, it is not easy to find an entirely adequate method of signaling the respect and good-will we entertain toward him. But for a host of reasons—and for many more than the inherent modesty of the individual himself would allow to be pointed out—the following narrative is implicitly biographical. It must, therefore, bring back many cherished memories. Moreover, it stands as testimony to his excellent scientific judgment and courageous and sympathetic administration. It is an opportunity, therefore, which all welcome, especially the author and his intimate advisers, to dedicate this review to Dr. Frank Baldwin Jewett.*

* The text of this review is different in some degree from that published in monograph form on the date of Dr. Jewett's retirement.

INTRODUCTION

The universal telephone service now provided by the Bell System has become such a "taken-for-granted" factor in our every-day national business and social life that one may easily forget the existence of the many regional frontiers which greatly restricted the usefulness of the telephone as recently as three decades ago.

The technical developments which made economically practicable the complete elimination of these regional frontiers were worked out in this country during the first two decades of the century. In spite of their technical and social importance, there is still lacking a connected recital that sets forth the various coordinated efforts by which the difficulties inherent in the long distance transmission of the voice were gradually overcome. Substantial amplification would be required to do justice to the concurrent accomplishments of the engineers who worked on the related problems involving outside plant, equipment, traffic, apparatus, and manufacturing questions. Without the important contributions made by these engineers in the associated departments of the American Telephone and Telegraph Company and Western Electric Company, there could not have been complete success.

Mention should also here be made of the fact that during the period covered by the story, steady improvement of transmission was effected in subscribers' exchange services.

The story as it unfolds divides naturally into four parts. The first is concerned with the 1904-1907 period when the A. T. & T. Co. headquarters staff was located at Boston and includes a discussion of the then current state of the art as a general background for the subsequent developments. The second part has to do primarily with the important sequence of achievements of the 1907-1911 period in New York which step-by-step prepared the way for the development of transcontinental telephony. The third part is primarily concerned with the transcontinental project itself, including the planning of the project. The fourth and concluding part reviews the subsequent establishment of a Bell System backbone network of repeatered, non-loaded, 165 mil lines interconnecting the large cities, and includes the removal of loading from the transcontinental line.

CHAPTER I

1904 to 1907 at Boston

A CAREER SEEKS THE MAN

IN the spring of 1903 Dr. George A. Campbell, then in charge of the work on "Loaded Lines and Theory of Telephone Transmission" in the Engineering Department of the American Telephone and Telegraph Company at Boston, visited Professor Harry E. Clifford of M.I.T. to inspect a 10,000 cycle generator that had just been acquired for some experimental work. While they were discussing the generator, a young instructor walked by. He was called back by Clifford and introduced to Campbell as Dr. Frank B. Jewett.

In the few minutes conversation that resulted, Campbell was much impressed with the charm of Jewett's personality and his alertness, high intelligence, and maturity. The thought flashed through his mind, "Here is the type of man we want in the Telephone Company." After Jewett had passed along, Campbell told Clifford his thoughts and impressions. The latter countered by remarking with evident satisfaction that Jewett was under contract to M.I.T. for the 1903-04 academic year.

The next year, however, Campbell renewed his efforts well in advance of the time for academic contract commitments and, in due time, Jewett visited the American Telephone and Telegraph Company engineering headquarters for an interview with Messrs. Hammond V. Hayes, Howard S. Warren, and G. A. Campbell. Warren, who was Campbell's immediate superior, was in charge of the so-called "Equipment Division," reporting to Hayes. At the time Warren had an authorization for a new man to work on protection problems, and he had become interested in Jewett as a prospect, in consequence of Campbell's recommendation. Hayes was one of the "Triumvirate" or Engineering Committee, that managed the Engineering Department in behalf of the Chief Engineer, Joseph P. Davis, then living in semi-retirement on account of illness.

The reactions of Hayes and Warren to Jewett's personality were similar to those earlier experienced by Campbell, and the interview resulted in a definite offer to Jewett and a tentative acceptance. This became final a few days later, after Jewett had convinced Hayes that he would be worth more than the amount previously suggested. The starting salary was \$30.76 per week, or \$1600 per year, which was big starting money. The standard

starting rate then current for new college men without post-graduate work was \$600 per year.

In his first few months with the Telephone Company Jewett worked on a wide variety of transmission problems, under Campbell's supervision, and handled considerable department correspondence. This was in accordance with the department's policy for an educational period prior to concentration on a specific line of work.

His first important 1904 assignment was a study of the Jacques' patent 767818 which had been offered for purchase. This patent had to do with a variety of schemes for improving transmission on long telephone lines. Jewett made an adverse report on the basis of theoretical studies and experiments which convinced him that the appearance of improvement was greater than the substance. His analysis was so fair and clear that it brought forth a note of commendation from some of the principal executives¹ of the organization, without known precedent for engineers just beginning their telephone careers, and must have been very heartening to its recipient.

1905 HAPPENINGS

A reorganization of the Engineering Department effective on January 1, 1905, resulted in Mr. Hammond V. Hayes becoming the chief engineer. The January 1905 organization chart on page 399 is the first official chart on which Dr. Jewett's name is listed. It is of incidental interest to note that he is the only individual mentioned by name who has remained in Bell System service up to 1944. There were a total of 195 employees in the Engineering Department, including a small, substantially autonomous, "operating" division under G. M. Yorke, whose headquarters were in New York City. This division was in fact the Engineering Department of the Long Distance Lines Department.

Early in 1905, Jewett made a good start on the protection job which had been in prospect when he was engaged. In his report² to Warren on the work done in 1905 Campbell listed the protection work as being one of three major activities, the other two being problems resulting from disturbances by alternating current railways, and the inspection of commercial transmission conditions. An intriguing feature of one of the protection development projects was the use of low inductance choke coils in series with the line at points adjacent to the protector blocks, in order to reduce protector

¹ Specifically Mr. Frederick P. Fish who was President of the American Telephone and Telegraph Co., and was also widely known as a successful patent attorney. Also from Mr. Thomas D. Lockwood who was in charge of the Telephone Company's Patent Department, and Mr. Hayes, himself.

² Some abstracts from Campbell's report to Warren, including the text of Jewett's 1905 report on the protection work, are given in Appendix I.

maintenance by curbing the severity of the lightning surges. Although the initial experiments were quite encouraging, the more extensive service trials proved insufficient advantage to warrant standardization.

Early in 1905 Jewett also started to build up a splendid record as a personnel recruiting agent for the headquarters staff.³

Jewett's 1905 engineering work, however, was not wholly taken up by protection problems or personnel recruiting. He also handled a large correspondence with the field engineers on current transmission engineering problems, including loading and phantom working, and made a number of special transmission and patent studies. This experience substantially broadened his training in the engineering work, and provided a helpful background for the assumption of new responsibilities on January 1, 1906, when he succeeded Dr. Campbell as head of the Electrical Department, reporting directly to Warren. For some time Campbell had been anxious to concentrate on theoretical research problems and he welcomed the availability of Jewett as a replacement. That Jewett was ready for department supervision responsibilities after less than 16 months' service with the Telephone Company proved the capabilities of the man and verified the initial appraisals of his potentialities made by Messrs. Hayes, Warren, and Campbell.

Now that the story has Jewett well started on his telephone career, it is appropriate to review the state of the art, and briefly consider organization responsibilities. Also the laboratory facilities and methods of transmission testing are briefly described.

STATE OF THE ART

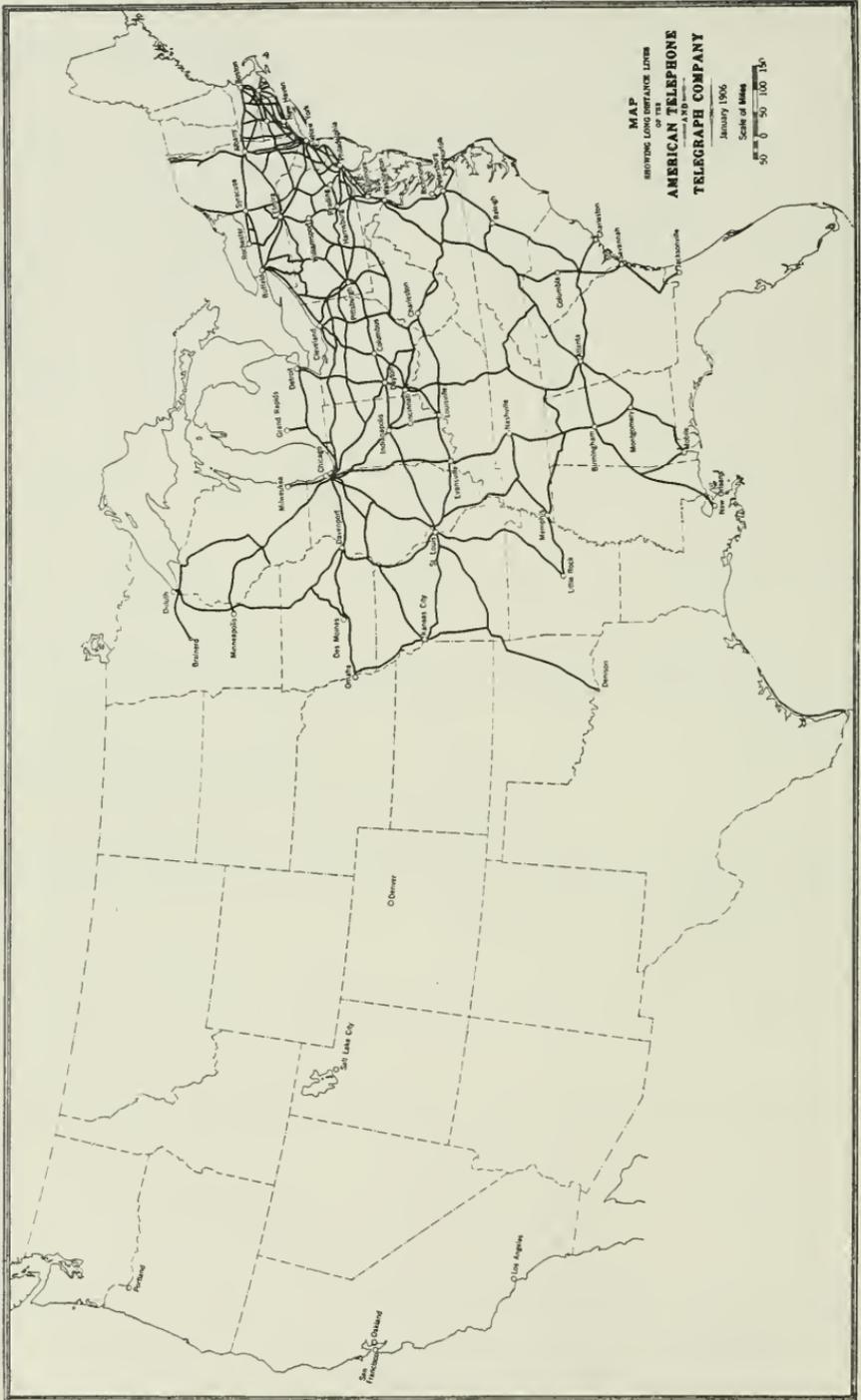
Regarding the general status of telephony at the beginning of 1906 we can take it from Frederick L. Rhodes' account in the "BEGINNINGS OF TELEPHONY" (published 1929) that the art had been very well begun and that the Bell System plant had been placed on a sound engineering basis. Much remained to be done, however, in all branches of the art, and in some of the fields which assumed great importance in later years the surface had hardly been scratched.

A few high spots of the 1906 status are briefly mentioned below to give some indication of what had been done and in some instances what remained to be done.

1. The telephone wire plant was substantially on a metallic circuit basis (excepting some rural subscriber lines).

³ His early selectees included F. J. Chesterman now Vice-President of the Bell Telephone Company of Penna.; O. B. Blackwell now Vice-President of the Bell Telephone Laboratories; H. S. Osborne, now Chief Engineer of the American Telephone and Telegraph Co.; John Mills, Director of Publication, and W. H. Martin, now Director of Station of Apparatus Development, Bell Telephone Laboratories.

2. Paper-insulated, twisted-pair cable construction dominated the exchange plant, with about 80% of the cable underground. 22-gauge cable was coming into extensive use.
3. The preparation of comprehensive conduit plans for the larger cities had started soon after the development of paper-insulated cables in the early nineties. By 1906, this work had been broadened to include definite forecasts of the future requirements of those cities, and the resulting development plans showed not only the most economical size and distribution of the conduits and cables, for a period extending about 15 years into the future, but also the proper number, locations, and sizes of the central offices. This work, in later years termed "Commercial Surveys and Fundamental Plans," enabled the operating companies to keep abreast of the continually advancing business needs with a minimum of reconstruction.
4. Common battery switchboards had been installed in all of the large cities replacing the magneto boards, and were being installed in the smaller cities and towns. The use of the new switchboards made local batteries and magneto generators no longer necessary at the individual subscriber stations, and resulted in great improvements in the speed and quality of the telephone service. Switchboard lamps used as line and supervisory signals were also factors in the improved service.
5. High-grade telephone instruments were in universal use, including the "solid back" transmitter which was much superior to its predecessors with respect to power and freedom from carbon packing.
6. Message rate service had been introduced in the larger cities, supplementing flat rate service, and was still an important factor in the rapid rate of station growth.
7. An accompanying map shows the A. T. & T. Co. long distance lines at the beginning of 1906. The long distance circuits were almost entirely in open-wire construction, little cable being used except to provide entrance facilities to city toll offices, and at river crossings. The loading standardized in 1904 for 104 mil open-wire pairs made such circuits approximately equivalent in transmission to non-loaded 165 mil pairs, their range being of the order of 1000 miles. Study of the problems involved in loading 165 mil circuits had begun.
8. The leased-wire telegraph business had expanded to become an important source of revenue and was becoming a basic factor in the expansion of the long distance telephone plant. The telegraph circuits that were leased as private lines were usually obtained by compositing the telephone circuits, and thus the same wires were used simultaneously for telephone and telegraph.



9. Phantom working for non-loaded open-wire lines had gotten a good start in the 1904-05 period, in consequence of the standardization of the first satisfactory phantom repeating coil (37-A). The complete commercial exploitation of the phantom circuit, however, awaited the development of usable quadded cable and of phantom group loading.
10. Several different standard loading systems had been made available for different fields of service on telephone cables, and improved types of cable suitable for use with loading had been developed. In the initial standard loading, the theoretical cut-off frequency was about 2300 cycles. The principal early uses of loaded cable were for circuits between metropolitan city and suburban offices, long entrance cables, and long switching trunks. These installations had yielded large economies by permitting the abandonment of plans for installing sizable networks of these types of circuits in coarse-gauge low capacitance cable, which had been formulated just prior to the invention of loading. A good beginning had been made in the use and in the planning of loading for intercity toll cables, the longest in use being the Boston-Worcester (44 mi.) 1904 cable. Plans for the New York-Philadelphia (90 mi.) and the New York-New Haven (80 mi.) loaded cables had been started in 1905.
11. Substantially continuous efforts to develop a telephone repeater had started inside and outside the Bell System soon after the invention of the telephone. These efforts usually involved receiver-transmitter combinations, and the designation "telephone relay" was quite common. It was not until 1904, however, that there became available a commercially usable result, namely the Shreeve receiver-transmitter type mechanical repeater. This device was being used only on non-loaded open-wire lines, with not more than one repeater in the circuit. Improvements which increased its field of use were worked out later in connection with the transcontinental line developments.
12. In the reference period, competition by the independent telephone companies was very strong and on the increase, as indicated by the following statistics regarding approximate numbers of stations:

<i>Item</i>	<i>At the end of 1905</i>	<i>At the end of 1907</i>
Bell Connecting Companies.....	246,000	826,000
Non-Connecting Companies.....	1,596,000	2,280,000
Independent Co. Totals.....	1,842,000	3,106,000
Bell System Stations.....	2,240,000	3,035,000

During the latter part of 1909, the Bell stations began again to outnumber the independent stations. As another manifestation of the competitive situation, it is of interest to note that in the middle of 1905 a majority of the cities having more than 10,000 stations (total) had both independent and Bell exchanges.

ORGANIZATION

In the period under consideration, 1906 and thereabouts, the Engineering Department in Boston was broadly responsible for Bell System engineering, development, and research work, and also had important responsibilities in inspection work. The organization units which handled the different types of work are indicated in appended organization charts.

The Department established engineering standards for plant design, prepared central office specifications, and advised the associated companies (then termed licensee companies) on current plant and traffic problems. By means of circulars, bulletins, and specifications, and in routine correspondence, it advised the field how to use new developments.

In conference committees and correspondence, it outlined the service requirements for telephone cable and the bulk of the telephone apparatus manufactured by the Western Electric Company—items on which the development work was done by Western Electric Company engineers, at New York or Chicago. In its own laboratories at Boston, it carried on considerable research work and also development work on many special items such as telephone instruments, loading coils, phantom repeating coils, and telephone repeaters.

TRANSMISSION

Within the Boston engineering department, most but not all of the transmission engineering work and all of the transmission development work was done by the group of nine engineers which became Jewett's responsibility on January 1, 1906. However, "cost studies" for plant extension projects, and exchange area loop and trunk studies, were then made by the Construction Division headed by F. L. Rhodes. Jewett's group also worked on electrical protection and inductive interference problems.

The laboratories were in a ground floor annex on Oliver Street. The principal "accessories" for transmission testing and impedance measurements included several single-frequency inductor-type generators ranging in frequency from about 200 to 5000 cycles, two "sound-proof" test rooms, two shielded-bridges for a-c measurements, fixed and variable inductance standards, capacitance standards, and several boxes of sectionalized artificial lines, one of them designed to "simulate" a long 104 mil open-wire line, while the others provided several adjustable lengths of standard 19-gauge reference cable. Facilities also were available for connection to out-

side lines in the plant of the American Company and the New England Company for experimental purposes.

Basically, the transmission tests were talking tests, usually with both ends of the circuit accessible at the test point, i.e., loop tests. The transmission equivalents of lines, and apparatus losses, were determined in terms of miles of reference cable. In tests to determine apparatus losses in lines, the losses were sometimes estimated on a percentage basis, when it was inconvenient to use the reference cable. Transmission quality judgments involving frequency distortion effects were usually expressed with a variety of non-standard adjectives, ranging from "sharp" to "boomy" or "drummy" and including others with a more salty personal flavor.

The non-availability of portable "high frequency" tone generators prevented the making of single frequency measurements in the field. However, in the period under discussion, considerable progress had been made in laboratory measurements of low amplitude alternating currents by means of thermocouples.

1906 AND 1907 AT BOSTON

During 1906, much effort of Jewett's group was devoted to field investigation and analytical studies of conditions affecting long distance telephone service, in consequence of transmission complaints by important users of the service. Much of the poor transmission was found to be due to defective apparatus and to departures from standard maintenance and operating practices. This transmission inspection work later became so extensive as to require the organization of a separate group of engineers. By progressive evolution, such investigations and corrective measures eventually led to the thorough organization of transmission maintenance work as we know it, now such an important factor in making the actual transmission performance of commercial telephone circuits closely approximate their theoretical design performance.

In 1906 a substantially increased fraction of the department effort was concentrated upon various phases of electrical interference work. This was made necessary by the increasing use of alternating current traction on interurban trolley lines, and the projected single-phase electrification of some important railroad systems, notably the N. Y., N. H., and H. R. R. This particular project continued to require a lot of attention for over a decade, initially in the electrification to Stamford and later in the extension to New Haven.

Engineering and apparatus problems arising from the rapid growth of loading, and the need to realize its maximum benefits took a great deal of time. Thus, loaded underground toll cables between New York and Philadelphia, and between New York and New Haven, were completed during 1906.

Further development work on phantom transposition systems and improvements in the balance of phantom repeating coils added to the advantages of phantom working on non-loaded open-wire lines.

To indicate the range and variety of the work done by Jewett's group, a copy of his report to Mr. Warren for the year 1906 is given in Appendix II.

1907 REORGANIZATION

Early 1907 saw a serious reaction following the business boom which reached its climax in 1906. Difficulty had been experienced by the American Company in disposing of a large issue of bonds and "the financial sky was filled with the scudding clouds that foretold the impending storm. A period of retrenchment and doubt had begun." This situation resulted in the retirement of Mr. Fish as president and the election of Theodore N. Vail on May 1, and was followed by a quick drastic reorganization of the telephone organization under Vail's careful planning.

To carry out his broad plans on engineering, development, and research, Vail selected as the new Chief Engineer for the American Company John J. Carty, who at the time was Chief Engineer of the New York Telephone Company. Carty's reorganization activities during the summer of 1907 resulted in a consolidation of the development laboratories of the Bell System in the Engineering Department of the Western Electric Company at New York. This new organization included a substantial portion of the Chicago group of development engineers, and several members of the American Company's Boston group. The amalgamated organization expanded almost from its inception, and nearly two decades later became the Bell Telephone Laboratories, Inc.

The 1907 reorganization also resulted in the Western Electric Company's taking over all of the inspection activities that previously had been carried on by units of the American Company's Engineering Department. The Engineering Department itself was drastically reduced in size and in September 1907 moved to New York along with executive departments. Charts dated June 1907 and December 26, 1907, pages 400 and 401 respectively, show the organization set-up prior to and after the reorganization.

The late spring threats of drastic reorganization had been quite disturbing to the Boston engineers, especially to those who had only recently started their telephone careers. Several of them, including Jewett, began to wonder whether they might not have made a mistake in joining the Telephone Company. A number of attractive college teaching offers which reached Jewett at about that time inclined him towards a resumption of his academic career broken in 1904, but the temptation was thrust aside after he had made a special visit to New York to interview the new Chief Engineer relative to the prospects for future advancement in the Telephone Company.

CHAPTER II

The 1907-1911 Period in New York

IMPORTANT early developments of this period led to the successful loading of open-wire phantom lines and their side circuits; the commercial application of loading to 165 mil open-wire circuits; and the development of duplex (quadded) cable and of phantom group loading for such cables. Jewett's prestige rose high in consequence of his personal efforts and his supervision of these developments, and was further enhanced by the basic roles these developments played in the New York-Denver line, and the Boston-Washington cable projects, which are also described in this chapter. The Denver line proved to be, as was intended, a major preparatory step in the westward march to achieve transcontinental telephony, and then universal telephony within the United States.

The important but more or less routine engineering work that was necessary to maintain continuous progress in the telephone transmission art went forward along with the specific developments in long distance telephony that are described in detail herein. Mention should also be made on the continuing fundamental work on the reduction of noise and crosstalk. Especially in the long distance services, this was a vital necessity as the lines became longer and longer. The rapid extension of the use of loading and of phantom working over lines and cables, followed by the introduction and the wide use of telephone repeaters, substantially increased the complexity of the noise and crosstalk problems, and greatly magnified the importance of the work. The steady improvement of transposition systems was an important part of the effort on the open-wire lines.

1907 HAPPENINGS

Following the move of the reorganized Headquarters Staff Engineering Department to New York, and partly in consequence of conditions that had led to the recent reorganization, but mainly because of the critical general business panic which exploded in Wall St. in October 1907, engineering and development activities of the Telephone Company operated at a relatively low voltage for a considerable period. Fortunately, the steps that Vail had taken to improve the financial status of the Company had been effective, and the storm was weathered without important changes in the rate of station growth, and without substantial distress of any kind. While plant expansion was slowed down, the stringency did not prevent the start

of important new development projects, or the continuation of important work which had been started at Boston.

The first new major project was that of developing phantom group loading for open-wire lines. The theoretical work on this problem started early in October 1907, and the design requirements for the new types of loading apparatus were put up to the Western Electric Company in December 1907. The general objective was to make possible the exploitation of the economies inherent in a full application of phantom working in the expensive open-wire plant. Other developments, mentioned later, were also essential to the full achievement of this objective. The important fact to remember in this general connection is that for a period of several years prior to the development of phantom group loading it was feasible to load 104 mil circuits, and to phantom non-loaded 104 mil circuits, but it was not possible to combine the advantages of phantom working and loading. Two new types of loading coils were necessary, one which became known as the side circuit loading coil for use on the phantom pairs, and the other for use on the superposed phantom itself. The original standard open-wire loading coils were not suitable for use on side circuits because of the transmission impairments and the unbalances that they would have introduced into the associated phantom circuits. The critical problem in the new loading apparatus was to obtain satisfactorily low crosstalk among the associated side and phantom circuits. The results of this development work are described in a subsequent section.

Among the important old projects that were continued and pushed in the months that followed the move from Boston were (1) studies and experiments to enable loading to be used with satisfactory results on 165 mil open-wire circuits, and (2) problems involved in the use of telephone repeaters on loaded lines.

The problem of loading the 165 mil circuits was primarily one of improving and stabilizing the insulation of the circuits, so that during wet weather and the subsequent drying-out periods the transmission impairments caused by leakage losses would not materially offset the transmission loss reduction obtainable with the added inductance. In the early commercial attempts to load 165 mil circuits (beginning with the New York-Chicago line, 1901) the loading eventually proved to have much too high an impedance in relation to the wet weather line insulation, and it was removed late in 1905 because under unfavorable weather conditions the transmission equivalent became (temporarily) worse than that of a non-loaded 165 mil line. The solution of the loading and insulation problems for the 165 mil lines required a great deal more experimental work than had been involved in the successful application of loading to the 104 mil pairs, primarily because of the much greater sensitivity of the heavier conductors to leakage effects. The open-

ing of the New York-Denver Line in 1911 proved, however, that the essential problems had been solved. A subsequent discussion includes a brief statement of the high spots of this and other developments that were essential to its success.

The early work on the problems involved in the use of repeaters on loaded lines was not so successful as that on the other concurrent major projects, a topic that we shall return to in connection with the planning of the trans-continental telephony project.

1908 HAPPENINGS

During 1908, Jewett's department initiated several additional important developments, including duplex (quadded) cable and low loss repeating coils for toll lines.

Over a long period, prior to 1908, many sporadic and unsuccessful experiments had been made, here and abroad, to obtain quadded cable suitable for phantom working. By the middle of 1908, however, very encouraging results had been obtained in the Bell System development work on open-wire phantom loading. Forecasts of the substantially universal use of loaded, phantomed, lines brought into sharp focus the need for loaded quadded entrance cables in the future open-wire toll plant. Also, if a satisfactory type of quadded cable could be developed and loaded, very large economies could be anticipated in long distance telephone cable systems that as yet were in the dream stage. These incentives were tremendous relative to those that governed the previous unsuccessful experiments referred to, and in fact compelled the success that was achieved in due course by the concentrated engineering efforts of the American Company and Western Electric Company, beginning in 1908. The first question to be decided by experiment and study was concerned with the type of construction that would offer the best chance of ultimate success. The leading competitors were the spiral-four type quad, and the multiple-twin quad, consisting of twisted pairs, twisted about one another. The twisted-pair quad eventually won out partly because of crosstalk considerations, but a not-negligible factor in the decision was the fact that if the new cable should not turn out to be completely satisfactory for phantom working, the side circuits of the twisted-pair would have characteristics more closely similar to those of non-quadded twisted-pair cable than would the side circuits of spiral-four quads. This was a powerful plant flexibility and homogeneity argument.

The cable development became commercially fruitful in 1910, and is described in a subsequent section of this story.

The 1908 repeating coil project mentioned in an earlier paragraph included high-efficiency phantom-deriving repeating coils especially for use on

165 mil open-wire pairs, and a high-efficiency non-phantom type repeating coil for Type B composite ringers. The existing standard 37A phantom-deriving repeating coil had been developed for use on lines operated on a 16-cycle ring-down basis, and had very good "ring-through" characteristics. In consequence of this feature, the speech transmission loss was quite substantial, being of the order of 1.5 db per coil at each end of a phantomed circuit. Two such coils, at opposite ends of a phantomed non-loaded 165 mil circuit, were equivalent to an extension of the length of the line by about 100 miles. This was too much of a transmission and economic penalty to be acceptable on expensive 165 mil circuits. By sacrificing the 16-cycle ring-through properties, it turned out to be a relatively simple job to reduce the transmission loss in the new coils to values below 20% of the loss in the 37A coil. In circuits equipped with the new coils, the signaling was accomplished by "composite" ringing (135-cycles).

HEADQUARTERS STAFF ENGINEERS VISIT THE PACIFIC COAST

There occurred late in 1908 and early in 1909 a Pacific Coast visit of several headquarters staff engineers which had an important place in the sequence of events that preceded the American Company's decision to provide transcontinental telephone service. Jewett participated in this expedition, and was joined later by Messrs. Carty, Gherardi, and others. The initial purpose of Jewett's trip was to advise the Pacific Tel. & Tel. engineers how to improve transmission conditions in certain parts of their territory, notably in the Oakland area, and on trunks to San Francisco, and also in the Los Angeles-Pasadena area. Aggressive competition by local independent companies was a factor in these problems. Improvements were also needed in the Pacific Company's long distance toll plant. An extensive use of loading was indicated. There were also a number of pressing inductive interference problems.

Before Jewett had finished his work on these various problems, Messrs. Carty and Gherardi reached San Francisco to consider with the Pacific Company executives some revisions in their 1909 budget, covering extensive plant additions that had become desirable. (Here again the competitive situation was a factor.) It was inevitable that Carty should become more deeply concerned with the telephonic isolation of the Pacific Coast when he was there and unable to talk to his staff in New York, than when he was in his own office in the east. This isolation was very real and oppressive; not only was there a large geographic gap in the wire plants of the Associated Companies, between the Pacific Coast area and the middle west, but also the limit then current on telephone transmission over the best available type of circuit was considerably less than one-half of the minimum transcontinental distance. Under the circumstances, it was natural that during his Pacific Coast trip Carty should spend considerable time with

his New York assistants, and with the Pacific Company engineers, in surveying the principal problems involved, and the prospects for transcontinental telephony in terms of the development work then under way and of potential future researches, particularly on telephone repeaters. Apparently, the prospects were encouraging.

Carty was stimulated in this study by pressure from President Vail, who happened to visit San Francisco while Carty, Gherardi, and Jewett were there. It appears that Vail was under some pressure from Pacific Coast business men who were then very busy planning the Panama-Pacific Exposition (originally scheduled for 1914 but later postponed to 1915), and who wanted him to promise that San Francisco would be put in regular telephonic communication with the eastern cities when the Fair opened. Being a good business man himself, Vail was sympathetic to these appeals. Realizing that engineering difficulties would be involved, he consulted Carty, who as usual, was unwilling to commit himself without a careful survey of the prospects and possibilities. Presently, Carty made a favorable report, and Vail told the Fair management that the telephone company would attempt to provide the desired transcontinental telephony. It thus happened that before he returned to New York Carty added the transcontinental line to his list of "must" objectives.

THE ENGINEERING SITUATION IN APRIL 1909

A major reorganization of the engineering department became effective on March 13, 1909, soon after Carty's return from the Pacific Coast. As shown in the chart on page 402 the reorganized department had two major divisions respectively reporting to B. Gherardi as Engineer of Plant and K. W. Waterson as Engineer of Traffic. Jewett reported to Gherardi. A third division of the department handled engineering work on legal cases.

In a memorandum of April 8, 1909 addressed to Vice-President Thayer, his immediate superior, Carty discussed the planning of the new organization and asked for additional personnel to enable him to carry on the new duties and responsibilities in associate company relations which had been assigned to his department, and to undertake certain important new engineering and development work, without neglecting the important work then under way. This memorandum includes such a beautifully clear and significant exposition of the engineering situation that substantial extracts are included in Appendix III. Carty's discussion of the principal projects in which Jewett's department was or would be involved are included in full under the headings: Phantom Circuits and Duplex Cables; Further Development of Pupin Invention; The Problem of the Telephone Repeater. From this discussion, it is clear that Carty expected that it would be possible to accomplish speech between New York and Denver over loaded 165 mil

open-wire circuits, but that this would be the geographical limit for loaded 165 mil circuits without also using repeaters which were not yet practicable on the loaded lines.

In building up the justification for the development of a "more powerful" telephone repeater, Carty wrote in part: "There is nothing in the nature of the case to discourage us in this line of work, and the art seems to have so many possibilities and the results to be obtained . . . are so far-reaching that the work . . . should be pushed vigorously. If we successfully load the Denver line and thereby accomplish speech between New York and Denver, the development of a successful repeater would enable us to accomplish speech between San Francisco and New York.⁴ The achievement of this result would mean universal telephony throughout the United States and its importance is so apparent that no argument is needed to demonstrate it." At this point it is appropriate and permissible to note that New York-Denver transmission was commercially accomplished in 1911, and that just prior to this achievement a vigorous and successful attack was launched on the new repeater problems. Meanwhile, the experimental work continued on the general problem of applying the Shreeve mechanical repeater to loaded lines.

The "more powerful" telephone repeater which Carty had in mind was a hypothetical inertialess repeater of an entirely new type. In the previously mentioned Pacific Coast analyses of the problems that must be solved to achieve transcontinental telephony, Jewett had convinced Carty that there was a good chance of obtaining a new and satisfactory type of repeater if research workers trained in the modern electronic physics could be hired and put to work on the problem. In Chapter III of this story, Jewett's personal contribution to the planning of this research program is considered at greater length.

OTHER 1909 HAPPENINGS

In general, the year 1909 was marked by accelerating, favorable, progress in the major transmission development and engineering projects previously mentioned. The principal discordant notes in the otherwise harmonious and tuneful concerto were caused by expanding difficulties in learning how to use telephone repeaters effectively on loaded lines.

A 1909 event that was responsible for starting one of the most important engineering projects in the 1910-1915 period occurred on March 4, at the time of the inauguration of President Taft. By wrecking all of the open-wire lines out of Washington, an unusually severe blizzard telephonically and telegraphically isolated the capital from the rest of the country for a

⁴ By reference to the quotations in Appendix III the reader will see that Carty's engineering group also clearly visualized that the repeater would be the open sesame to successful *radio* telephony.

period of several hours. This event dramatized the need for storm-proof communications to the capitol, and led to a decision by President Vail that a complete underground telephone cable system should be established along the Atlantic seaboard, between Boston and Washington. Vigorous development activity on the new types of cable and loading coils that would be required got underway early in 1910. A full discussion of this development is given later on under the heading "Boston-Washington Loaded Duplex Cable Project."

By the spring of 1909, the development work on quadded cable had reached a stage which made it desirable to start the development of new types of cable loading coils suitable for use on phantom and side circuits. This work benefited from the earlier work on the open-wire phantom loading. Since it then appeared that there would be little use for duplex cable on a non-loaded basis, the work on the cable loading coils was coordinated with the further work on the new type of cable, leading to a joint trial in 1910 on the Boston-Neponset project described later.

1910 ACHIEVEMENTS

Progress during 1910 was especially important and interesting. It included the first (Bell System) loaded submarine cable installation, which is of special historical interest even though it was not directly related to the major transmission projects previously discussed. During 1910, the initial objectives of these major projects were realized in full measure, and before the end of the year the gains in the new engineering knowledge were being consolidated for very important new engineering projects, notably the New York-Denver line and the Boston-Washington underground cable. All of these various projects are separately discussed below.

CHESAPEAKE BAY LOADED SUBMARINE CABLE

This was the first Bell System submarine cable to be provided with submarine loading. It was an intermediate cable, crossing upper Chesapeake Bay, in an open-wire line providing service from Baltimore to the Eastern Shore points, greatly shortening the route. It was a 17-pair, 13-gauge, paper-insulated cable and had two underwater loads. The engineering and installation of many loaded submarine cables that were subsequently installed in shallow water crossings of river or bay include practices that originated in the 1910 Chesapeake Bay project.

OPEN-WIRE PHANTOM LOADING

Returning to the story of the major transmission developments in which Jewett's department took a leading part, attention will first be given to open-wire phantom loading. This was proved feasible in a commercial trial on an

open-wire phantom group of 104 mil conductors, between Newtown Square and Brushton (test stations near Philadelphia and Pittsburgh, respectively) installed during August 1910.

By design, the side circuit transmission characteristics were substantially identical with those of the loaded non-phantomed circuits then in extensive use. A slight impairment resulted from the increase in circuit resistance caused by the inserted phantom loading coils. These coils were installed at the same points as the side circuit loading coils, at systematic intervals of about eight miles, a distance set by the need to coordinate the coil spacing with the line transposition systems. The phantom loading coil inductance (0.163 henry) was chosen to provide a theoretical cut-off frequency close to that of the side circuit (approx. 2400 cycles). This resulted in a phantom circuit impedance approximately 60 per cent of that of the side circuits, and an attenuation nearly 20 per cent lower than that of the side circuits. This advantage resulted in the phantom being preferred for long-haul service. The increase in transmission efficiency obtained by loading the phantom (about 2.5 to 1) was practically as large as that obtained in the side circuits.

Since this was a pioneering project, it is understandable that the crosstalk results were not all that could be desired. There was some real satisfaction, however, in the fact that the crosstalk was not too close to the borderline of being intolerable. The crosstalk was due to unbalances in the line and in the loading coils. In commercial service, unbalances in the phantom-deriving repeating coils and in the composite telegraph sets were also factors in the crosstalk performance. In the course of time, in consequence of improvements in the phantom transposition systems and experience in the manufacture of the line and terminal apparatus, substantial improvements in the service crosstalk characteristics were secured.

The loaded phantom circuit was much more susceptible to noise induction than the side circuits, and increased the need for good line maintenance.

At this point, a few remarks regarding the conservative policy followed in this phantom loading development are appropriate. So far as the loading apparatus development work itself was concerned, a trial installation could have been made much earlier than the summer of 1910. The early laboratory work on the proposed initial loading coil designs showed several minor changes to be desirable from the crosstalk standpoint. After these were made, the designs appeared to be free from inherent dissymmetry that might cause crosstalk. The question as to whether the coils would have satisfactory balance when manufactured on a quantity production basis, however, could only be determined by undertaking manufacture of a sizable lot. When the question of making a trial installation was first considered in-

formally with the Long Lines plant people, no suitable types of additional new facilities were in prospect, in consequence of the somewhat slow recovery of general business activity following the 1907 panic. Prior to starting production of the new types of coils, models were turned over to the Long Lines telegraph experts for tests to determine whether objectionable impairment in the superposed telegraph service would result in consequence of the increased magnetic coupling between the telegraph circuits, contributed by the loading coils. The favorable report on this feature was tinged with an informal suggestion of regret that the wide application of phantoms in the long distance plant would reduce the aggregate number of wires that would be available for the leased wire telegraph services.

In arranging for potting the loading coils used in the trial installation, the decision was made to encase the coils individually so that in the event of unsatisfactory results the phantom loading could be removed without disturbing the side circuit loading. The phantom coil was considerably larger than the side circuit coil, and a new case had to be developed for it. The practice of separate potting of the individual coils continued for several years, mainly for flexibility and maintenance reasons. Not long after the trial installation, the manufacture of the non-phantom type open-wire loading coils was discontinued in favor of the new side circuit type. Gradually, the bulk of the existing non-phantomed loaded circuits in the open-wire plant was made suitable for phantom working. The displaced non-phantom type loading coils were returned to the factory for "conversion" into side circuit type coils, by partial rewinding of the original cores.

LOADING OF 165 MIL OPEN-WIRE CIRCUITS

The development efforts to improve the wet weather insulation of 165 mil wires sufficiently to make loading commercially practicable culminated in an experimental installation of loading on a New York-Chicago circuit during 1909 and early 1910.

The initial steps in this trial were (a) to change the transposition arrangements from the single-pin type to the drop-bracket type in order to avoid tying to the same insulator the two wires being transposed, and (b) to install bridge wire insulators at all bridling points, including the loading coil and lightning arrester leads. Comparative wet weather tests of the single-pin and the drop-bracket transposition arrangements made previously had indicated the new method to be about 20 per cent better, and tests with the bridle wire insulators had indicated their use would substantially eliminate low insulation at bridling points.

The bridle wire insulator was the final result of a long period of development. It provided sheltered dry spots on the rubber-insulated braided leads of loading coils and lightning arresters, and on bridle wires to test stations

and cable terminals. The need for these dry spots had become apparent from analyses of line tests which showed that after the braid on the wire had weathered and had begun to disintegrate, a considerable period of time elapsed after rain ceased before the line insulation returned to its usual dry weather excellence. The insulated wires passed through the insulator and at the point of exit the conductors were soldered to a metal insert moulded into the insulator. The bridle wires themselves supported the insulators at a point close to the connections to the line wires. It is of interest to note, in passing, that a patent was granted to Jewett on some design features of this insulator.

Returning to the discussion of the New York-Chicago loaded line experiment, it was found that the insulation improvements described above were insufficient to provide satisfactory transmission performance during periods of continuous bad weather. During fair weather periods, however, the transmission was as good as had been expected. The experiment thus proved beyond question the need for a new type of line insulator having substantially better insulating properties than those of the standard toll line insulators, which were made of glass and had a single petticoat.

Renewed studies of this particular question led to the rush development of a moulded double-petticoat porcelain insulator. The possibilities of porcelain insulators had been under consideration for several years, notwithstanding adverse cost factors. The accumulated test data on porcelain insulators generally similar in design to the standard glass insulators indicated that after a long period of exposure on roof racks the wet weather insulation was about twice as good as that with the glass insulators. Moreover, theoretical studies indicated that a properly designed double-petticoat porcelain insulator should be about twice as good as the single-petticoat porcelain insulator. The possibility that the opacity of the porcelain might unduly encourage insects to build their nests under the petticoats and thereby impair the wet weather insulating properties, however, could not be allowed for quantitatively in the preliminary estimates of the potential over-all improvement, due to the limited and conflicting evidence on this question.

Consideration of all of the factors involved, including favorable price estimates, led to a decision in October 1909 to substitute the new double-petticoat porcelain insulator on the experimental loaded line. An accumulation of manufacturing difficulties delayed the completion of the installation, so that the transmission observations and over-all line insulation tests did not get under way until the spring of 1910.

After a suitable test period it was found that although the wet weather line insulation was not so high as had been expected it was sufficiently good to warrant the general commercial use of loading on 165 mil circuits. Ac-

cordingly, the improved insulation features of the experimental loaded line, including the new porcelain insulator, were recommended⁵ for this use.

The loading arrangements standardized for the 165 mil circuits were similar to those which for several years had been standard for 104 mil circuits, viz., 0.265 henry coils installed at intervals of about 8 miles. With line insulation of 5 megohm-miles or better, the transmission range of the loaded 165 mil circuits was about 2.3 times as great as that of non-loaded 165 mil circuits.

LOADED DUPLEX CABLE:

The pioneering development work on duplex cable and on new types of loading coils for the cable phantom and side circuits culminated in a commercial installation between Boston and Neponset, which became ready for service early in September 1910.

By the fall of 1909, the experimental work and analysis of test data on experimental lengths of cable having multiple-twin quads and spiral-four quads respectively, and on the required new types of loading coils, had progressed sufficiently to make it desirable to undertake trial manufacture, preferably for a project that would meet a need for new facilities. In November 1909, it was decided that the Boston-Neponset cable project would be a suitable objective. About 5.8 miles of 37-quad 13-gauge cable was required, extending south from Boston, partly for use as an entrance cable for the American Company's Central and Shore (loaded open-wire) lines, and partly as a suburban trunk cable for the New England Company. The entrance facilities were to be provided with phantom group loading, in anticipation of extensive phantom working on the open-wire lines. Since phantom working was not needed or desired on the suburban trunks, the loading for the New England Company was limited to the quaddled pairs, using side circuit type coils. This particular plan was more valuable from the standpoint of development experience and gave greater plant flexibility

⁵ Before the end of 1911, an appreciable degradation was noticed in the wet weather line insulation of the initial installations. This appeared to be due to several causes which could not be separately appraised, including (1) an unexpected retention of deposits on the glazed surface, i.e. the insulator was not self-cleaning; (2) the nesting of insects underneath the petticoats; and (3) trouble with a large number of defective insulators. In general, the impaired insulation was not large enough to warrant the removal of the porcelain insulators, except those that were physically defective, nor was the "bug trouble" sufficiently serious to warrant the establishment of routine cleaning operations. Meanwhile, the prices of the porcelain insulator increased drastically in consequence of the continued manufacturing difficulties, which also greatly limited the supply of acceptable insulators. About the middle of 1912 it was decided to start using double-petticoat glass insulators on new installations of loaded 165 mil lines, in place of porcelain insulators. These glass insulators had become available in consequence of development work undertaken late in 1910 for the Western Union Telegraph Co. (at that time closely affiliated with the Bell System). The new glass insulators were much less expensive than the porcelain insulators and their insulating properties were nearly as good. Subsequent experience showed them to be fairly satisfactory for use on loaded 165 mil circuits.

than if non-quadded pairs and non-phantom type loading coils should have been used.

Cable Details: The multiple-twin quad was chosen in preference to the spiral-four quad on the basis of practical features involved in manufacture and installation and the resulting arrangement, in case phantom working should not prove successful. That is to say the twisted-pair side circuits would better serve the customer's plant flexibility needs than would the side circuits of spiral-four quads. Also the early experiments had indicated a high probability of superior balance among the very important within-quad couplings.

Additional factors that resulted in the subsequent standardization of multiple-twin quads were:

1. Their phantom capacitance is about 60 per cent higher than that of the side circuits, whereas in spiral-four quads the phantom capacitance is upward of three times as high as that of the side circuits. Consequently, when the phantom and sides are loaded for equal cut-off frequency and at the same spacing, the resulting impedances are such that the attenuation of the multiple-twin phantoms is considerably lower than that of their associated side circuits, whereas the attenuation of the spiral-four phantoms is inherently much higher than that of their side circuits. Since the phantom loading for the open-wire lines had provided phantom circuits which were preferable to their own (loaded) side circuits, due to their lower attenuation; it also seemed desirable that the loaded cable phantom should be better than the loaded side circuits.
2. The ratio of phantom circuit to side circuit capacitance in multiple-twin quads is close to that in the open-wire lines. In consequence, the loaded entrance cable impedances are better related to the open-wire impedances than would be practicable with simple loading on spiral-four quads. These comparisons again assume the phantom and side circuits to be loaded at the same spacing, with coil inductances giving equal cut-off frequencies.

In the design of the multiple-twin quad cable, the "staggered-twist" principle which had been found necessary from the crosstalk standpoint for use in loaded non-quadded pair cables was applied to the individual circuits of a quad. The lengths of twist were different for the two side circuits and a still different length of twist was used in the phantom. Adjacent quads had different combinations of pair and quad twists, and adjacent layers were stranded in opposite directions.

Notwithstanding these basic design precautions a great many manufacturing difficulties were encountered in preventing the phantom-to-side capacitance unbalances from being very much too large. In the first

lengths of commercial cable made up, the unbalances were worse than those in the 1908 experimental lengths. From then on, engineers from the Engineering Departments of the Western Electric Company and the American Company cooperated closely with the factory engineers throughout the entire manufacturing period in working out important fundamental improvements. One of the greatest difficulties was in obtaining sufficiently symmetrical twisting of the individual pairs, prior to forming the quads. Machine limitations and the need for conductors of identical size and ductility, insulated with exactly the same thickness of paper, were factors in this problem. Other difficulties too many and too involved for present discussion were also encountered.

At the beginning of manufacture it was appreciated that it would probably be impossible with known techniques to obtain quadded cable that would be completely satisfactory from the crosstalk standpoint, especially phantom-to-side crosstalk, without resorting to capacitance unbalance adjustments during installation of the cable. Some preliminary consideration was given to the use of balancing condensers. Further studies of possible methods of field balancing led to the development of a technique for measuring the capacitance unbalances in adjacent lengths of cable and selectively splicing the conductors of the quads of one length to those of another length in such a manner that the like-type unbalances would tend to annul one another. Suitable field test sets were developed for determining the magnitudes of the cable capacitance unbalances and their relative phase relations. The planned splices made in accordance with the proposed technique became known as capacitance unbalance test splices.

In splicing the Boston-Neponset cable, a total of 7 capacitance unbalance test splices were made in each full loading section at intermediate points approximately $\frac{1}{8}$ of a full loading section apart. The first set of test splices was made at the $\frac{1}{8}$, $\frac{3}{8}$, $\frac{5}{8}$, and $\frac{7}{8}$ section points. Then "semi-final" tests were made at the $\frac{1}{4}$ and $\frac{3}{4}$ points. The test splice made at mid-section was most important because it was the final test splice. Splices required at points between the test splices were made on a random basis. The test splices were made primarily for the reduction of phantom-to-side unbalance. When individual quads had objectionably high side-to-side unbalances, reductions in the residual unbalances could usually be obtained by planned splicing to other quads having high side-to-side unbalances.

The test splicing procedure was very effective in reducing the pile-up of objectionable unbalances. In general, the maximum residual unbalances per loading section were kept below a predetermined tolerable value. The average residual phantom-to-side capacitance unbalance per full loading section turned out to be of the same order as the average unbalance in the individual (approx.) 500-foot lengths when they left the factory. Statisti-

cal considerations indicate that if the cable should have been spliced on a random basis, without regard to capacitance unbalance in individual lengths, the r.m.s. average residual unbalance per loading section would have been about four times as great as that actually obtained.

On the whole, crosstalk results, including the effects of the loading coil unbalances, were considered fairly satisfactory for an initial pioneering effort, but not good enough as a standard of excellence to work to in subsequent projects which were to be of an entirely different order of magnitude and importance in the scheme of nation-wide telephony.

Since the cost of the capacitance unbalance test splicing was small relative to the total cost, the general technique used on the Boston-Neponset cable was subsequently standardized for general use in installing quadded cable. Eventually, substantial reductions in the amount of the test splicing resulted from improvements in cable manufacture.

Loading Details: Except for its phantom working feature, the side circuit loading for the Boston-Neponset cable was similar to the old standard medium-weight loading, originally developed for non-quadded cable. 175 mh coils were installed at intervals averaging about 8520 feet. A total of 72 side circuits were loaded, using two cases, each containing 36 coils at each load point. Eighteen phantom circuits were loaded with 106 mh coils at an average spacing of about 8450 feet. The differences in average spacing resulted from manhole space limitations which led to the phantom coils being installed systematically at manholes next in line to the associated side circuit loading points, at distances ranging from 214 to 490 feet. This layout would have simplified the removal of the phantom loading, if the phantom-to-side crosstalk should have been large enough to make phantom working impracticable. The phantom loading just described conformed to the established cut-off frequency standard for cable loading (approx. 2300 cycles) and gave an attenuation loss reduction of the same order as that provided by the side circuit loading. The absolute attenuation in the loaded phantom was approximately 20% below that in the associated side circuits. The nominal impedances of the cable phantom and side circuit loading were about 800 and 1300 ohms, respectively.

The new types of loading coils used on the Boston-Neponset cable were generally similar in basic design features to the side circuit coils and phantom coils used in the open-wide trial installation of phantom loading, but were much smaller in dimensions. Mainly because of size differences, the unbalances in the cable coils were much smaller than those in the open-wire coils.

EXPLOITATION OF THE NEW DEVELOPMENTS

The pioneering developments just described were so basically important in extending the limits of long distance telephony, and reducing the cost of

long distance facilities, that a substantial amount of engineering information regarding them was made available to the field in conferences, and in routine correspondence on current engineering projects, in advance of the completion of the development work. A coordinated quantitative statement was given in General Engineering Circular No. 107, "Aerial Loading and Duplex Cable," which was issued on August 19, 1910, in advance of the completion of the trial installations of open-wire phantom loading and loaded duplex cable, so as to assure that the new developments would be fully taken into account in preparing the 1911 Provisional Estimates and in planning the 1911 construction program.

In the open-wire plant, the new loading and phantoming developments had extensive application even before the end of 1910. The following pertinent quotation is from a paper, "Long Distance Telephony in America," read by John J. Carty at the second international conference of European Telephone & Telegraph Administrations, Paris, Sept. 1910.⁶

"Aerial Loading: At the present time there are about 52,000 miles of loaded No. 12 NBSG Circuit in the United States and about 1000 miles of No. 8 BWG loaded circuit. There are at present under construction, or intended for completion by January 1, 1911, about 17,000 miles of No. 12 NBSG loaded circuit, and about 13,000 miles of No. 8 BWG loaded circuit. Of this latter, about 3800 miles, namely four circuits from New York to Chicago will be arranged for phantom working, . . ."

The Carty paper also was prepared in advance of the completion of the development work. In general it could be considered a European edition of G.E.C. 107, but it went beyond the latter in mentioning some high spots of certain spectacular new engineering projects, namely the New York-Denver line and Boston-Washington underground cable.

In the cable plant, also, there was accelerating activity in the installation of loaded duplex entrance cables, beginning in the latter part of 1910. This was especially desirable in connection with open-wire phantom lines that were used also for composite telegraph service, due to complications otherwise involved in carrying the telegraph circuits through the entrance cables. The use of loaded quadded cable for toll cable facilities also started quickly and expanded rapidly, the Boston-Washington project being the outstanding initial commitment.

From what has been said in the preceding paragraphs, however, it must not be inferred that the development ended with the completion of the initial commercial installations. As time went on, the service requirements became increasingly severe, especially as regards crosstalk, and there has been substantial continuous activity ever since in the laboratory, factory, and field in the reduction of crosstalk unbalances in quadded cable, lines,

⁶ The Carty paper also was published in the March 1911 issue of "Telephone Engineer" (Chicago).

and phantom loading apparatus, and in other apparatus associated with phantom circuits and their side circuits.

THE NEW YORK-DENVER LINE

This project was especially significant in utilizing the recent radical advances in the telephone transmission art to achieve a specific long distance objective which constituted a recognized necessary preparatory step in a broad fundamental plan for transcontinental telephony.

This line made use of the phantom circuit of loaded 165 mil open-wire phantom groups installed between New York and Chicago (via Buffalo) and between Omaha and Denver, during 1910. The intermediate Omaha-Denver portion of the circuit initially consisted of a 165 mil non-phantom pair loaded with the new side circuit type coils. Some time later, a second 165 mil pair on the line was moved to pins adjacent to the first mentioned pair so that these two pairs could be phantomed and the phantom loaded. From then on, the New York-Denver circuit was a continuous phantom circuit. The new high-efficiency type of phantom-deriving repeating coil was used throughout this installation.

The construction work on the initial layout of the line was completed in December 1910, and the first through talk was made on December 29, 1910. (These are the reasons for describing this project as a 1910 development.) Commercial service, however, did not start until May 8, 1911. The intervening time was utilized in clearing up noise and crosstalk trouble which the time schedule and the lack of complete engineering information had made it impossible for the engineers to predetermine and take care of in advance. The initial transmission tests showed the side circuits to be satisfactory with respect to transmission and noise. The phantom, however, was very noisy and the phantom-to-side crosstalk was quite heavy. The line crosstalk difficulties were found to be mainly due to transposition irregularities, including omitted transpositions. For noise reduction, considerable retransposition work was necessary in regions where inductive interference prevailed and some rerouting in entrance cable portions involving the use of selected pairs in existing non-quadded cables.

When the line was opened for public use, the transmission performance was as good as had been expected when the project was planned. Crosstalk and noise were well within tolerable limits. The theoretical equivalent was about 28.5 db (appreciably better than the equivalent of the older non-loaded 165 mil circuits between New York and Chicago). According to the standards for long distance transmission that were worked to in the period under discussion, the transmission was considered to be satisfactory between terminal stations in New York and Denver, but the margin of transmission was not great enough to provide really satisfactory service

when long switching trunks were involved at the two ends. Consequently, as soon as the subsequent developments in telephone repeaters would permit, experimental repeaters were put into use on the New York-Denver connections.

The Denver line was not a through or terminal circuit, ready for use on call. It was built up when needed, with switches at Morrell Park (Chicago) and Omaha. When it became necessary to use side circuit portions as substitutes for the phantom circuit portions, the over-all transmission was several db worse than when the phantoms were used. This was partly a result of the higher attenuation of the side circuits, and partly because one of the side circuits was arranged for telephone connections and leased wire telegraph at several intermediate points between Omaha and Denver.

There was a heavy use of grounded d-c telegraph service on the composited line wires. At times, serious impairment to the telephone transmission was caused by non-linear distortion resulting from transient magnetization of the loading coil cores by the superposed telegraph currents. This effect became known as "Morse flutter" and later on as "telegraph flutter." It had previously been noticed on shorter loaded lines, but as the composited loaded lines became longer and longer, the flutter interference became more and more serious. Subsequent developments, first used on the transcontinental line, resulted in a considerable reduction of telegraph flutter per unit length of line, but there always was an appreciable amount on long loaded lines when they were used for composite telegraph circuits.

BOSTON-WASHINGTON LOADED DUPLEX CABLE PROJECT

As previously indicated, a main purpose of this project as planned in 1909-1910 was to provide storm-proof communications between the principal eastern cities. This project is historically significant as being the first long-distance cable system to use quadded cable and phantom group loading. It was also the longest telephone cable system ever designed for use without repeaters. Fortunately, a satisfactory type of telephone repeater became available for commercial use before the project was completed, and its value was thereby greatly enhanced.

General: The over-all length, Boston to Washington, was about 455 miles, with New York close to the half-way point. Underground conduit already existed for about half of the total distance—Boston to Providence, and New Haven to New York to Philadelphia to Wilmington. Heavy-loaded, non-quadded, 16, or 14, or 13-gauge cables were already in commercial use in different parts of the route but, in general, since they had been designed for shorter distances, they were not sufficiently efficient to be used in tandem connections with new cable as portions of Boston-New York or New York-Washington all-cable facilities without also using telephone

repeaters. The magnitude of the over-all project was so great that the construction work and the manufacturing effort had to be undertaken in several steps, spread over a period of years.

The first step was to lay a new underground conduit between Wilmington and Washington via Baltimore, and install a composite coarse-gauge loaded cable between Philadelphia and Washington, thus closing the gap between New York and Washington. It was an economically fortunate coincidence that when it became necessary to engineer the cable and the loading, a good start had been made on the development of duplex cable and of cable phantom group loading. The status of these developments, however, was such that it was very far from a certainty that long-distance loaded duplex cable facilities could be made satisfactory from the crosstalk standpoint. On the other hand, the economic stakes were very great; if the crosstalk could be kept within tolerable limits, the complement of coarse-gauge circuits would be nearly 50% larger than if a coarse-gauge non-quaddled cable should be used, and the transmission equivalents obtainable in the loaded phantoms would be appreciably lower than those on loaded non-quaddled pairs using the same size conductors. The increment cost in getting the extra complement of higher grade circuits consisted principally of the cost of the phantom loading coils and the cable capacitance unbalance adjustments. This was judged to be small in proportion to the potential value of the circuits. The decision to proceed with the development of the duplex cable and phantom group loading was made in the spring of 1910, while the quaddled cable for the Boston-Neponset project was under production, and prior to the start of production of the new phantom loading apparatus.

A fundamental transmission objective was to obtain facilities which upon completion of the project would meet talking standards desirable for Boston-New York and New York-Washington connections, with a few circuits suitable for emergency use between Boston and Washington in case the open-wire lines should go down. These requirements called for 10-gauge conductors. Also, there were requirements for all-cable circuits connecting intermediate points, which called for 13-gauge or 16-gauge conductors. Since the cable route intersected the existing open-wire routes at many points, patches could be made for the protection of the long distance service normally handled in open-wire.

Philadelphia-Washington Section: The requirements above summarized resulted in the design of a cable having 7 quads of 10 B&S gauge conductors, 18 quads of 13-gauge conductors, 6 non-quaddled 13-gauge pairs in the interstices of the layer of 10-gauge quads, and 18 non-quaddled 16-gauge pairs in the interstices of the 13-gauge quads. All quads were of the multiple-twin type. The loading cost study resulted in a decision to use a new weight of loading, medium-heavy, with coils installed at intervals of

about 1.4 miles, intermediate between that of the spacings for standard heavy and medium loading. The coil inductances were chosen to provide a cut-off frequency of about 2300 cycles, which was standard for cable loading for nearly two decades. Heavy loading was also considered, but its efficiency would not have been enough better to justify the extra cost and the extra manufacturing and installation efforts in working to difficult rush schedules.

Since the loading coils were designed to be in approximate cost-equilibrium with the cable conductors, those used on the 10-gauge quads were nearly as large as the standard open-wire loading coils, and much larger than those used on 13-gauge quads, which in turn were appreciably larger than those used on the 16-gauge pairs. Also, for cost-equilibrium reasons, the phantom loading coils were larger than their associated side circuit coils and were potted with them, with the cross-connections between them made inside the cases at the point where the connections to the stub cables were made.

The manufacture of the Philadelphia-Washington section of the cable and the loading coils was completed during 1911, and commercial service started on March 22, 1912. A great many manufacturing difficulties were encountered in preventing the cable and coil unbalances from reaching unacceptable magnitudes, including some difficulties not previously experienced in the Boston-Neponset project which were in part due to the larger sizes of conductors.

As the associated phantom and side coils were potted in the same case it was possible to obtain some crosstalk advantages by poling the unbalances in one type of coil against those in the other type. In order to maintain schedules, it was sometimes necessary to accept loading apparatus and cable lengths having objectionably high unbalances. In such instances, special arrangements were made for installing these items at points remote from the principal toll centers so that the resulting crosstalk would be substantially attenuated before reaching the telephone subscribers.

In installing the cable, capacitance unbalance test splices were made at 7 points in each loading section, following the procedure used on the Boston-Neponset cable, but working to more severe limits on residual unbalances. In the over-all tests made prior to commercial service, the crosstalk was found to be within permissible limits, but was in excess of values considered satisfactory. In consequence, renewed and successful efforts were made to obtain better crosstalk performance in the next section of the project, that between Hartford and New Haven.

The transmission tests on the Philadelphia-Washington cable showed the attenuation to be about 13% better than had been estimated on the coarse-gauge circuits. This was in part due to conservatism in estimating, but mainly due to special manufacturing efforts to dry out the cable so as to

reduce the dielectric losses. Also, the loading coils were somewhat better than expected. In a report to Gherardi, Jewett wrote:

"These tests indicate a most successful and gratifying outcome for a piece of work which has taxed our energies almost to the limit and which it must be admitted was an extremely large experiment at the time the original decision to lay a cable was made. The results obtained seem to me conclusive proof, if such proof is still needed, of our ability to forecast transmission results without recourse to laboratory demonstration . . ."

It is of interest to note that the attenuation in the 10-gauge cable circuits was approximately $\frac{1}{3}$ lower than that of non-loaded 104 mil open-wire circuits, and only 50% higher than that of loaded 104 mil and non-loaded 165 mil open-wire. The attenuation in the 13-gauge circuits was approximately 70% higher than that in the 10-gauge circuits.

Considerable telegraph flutter was noticed in the long cable circuits when they were used simultaneously for telephony and telegraphy. This led to the restriction of the magnitudes of the superposed telegraph currents and to the development of more sensitive telegraph relays. Several years later, (1916), the development of more stable types of loading coils proved to be beneficial in the reduction of telegraph flutter in the new facilities on which they were used.

Remainder of Project: Work started on the manufacture of cable and coils for the New Haven-Providence section of the Boston-Washington project before the Philadelphia-Washington section was completely installed. The section between Hartford and New Haven was placed in service February 13, 1913. Some manufacturing difficulties temporarily stopped cable production, with the result that the Hartford-Providence section did not get into service until October 1913. This installation closed the last gap in the underground cable system between Boston and Washington, and led to the following statement by President Vail in the report to stockholders for the year 1913:

"During the year 1913 we have made such further advances in the art of loading and balancing underground circuits, and have so greatly improved the intermediate apparatus, that it is now possible to talk satisfactorily by underground wires from Boston to Washington, in part through types of cable formerly suitable for short haul distances only. These short haul cables make up 47 per cent of the total cable in the line."

The "intermediate apparatus" referred to were telephone repeaters that had been made available for a few long haul circuits, as mentioned later in the transcontinental repeater development story. In another paragraph of the report, Vail remarked:

"By using the underground in connection with the overhead, the Seaboard cities from Washington to Boston could be no longer isolated by storms destroying the overhead wires."

The storm-proof communications objective having been achieved, the remaining sections of the coarse-gauge quadded cable project were installed in accordance with the need for additional circuits in particular sections of the Boston-Washington route. These additional coarse-gauge cables became available for service in the following sequence: Boston-Providence, November 1914; New York-New Haven, April 1917; and New York-Philadelphia, 1918. The listed dates apply to the initial loading complements, which usually included all of the 10-gauge quads and half of the 13-gauge quads.

Several of the shorter sections made use of a new cable layup having 3 quads of 10-gauge conductors, 30 quads of 13-gauge conductors, and 18 pairs of 16-gauge conductors. This change in layup was a significant result of the new repeater developments then current, which greatly reduced the need for 10-gauge circuits. By dropping out 4 quads of 10-gauge conductors and the 6 interstice 13-gauge pairs (a total of 18 circuits), it was possible to provide 12 additional quads of 13-gauge conductors, giving a net gain of 18 circuits.

CHAPTER III

The Transcontinental Telephony Project

A. PLANNING THE PROJECT

FROM what was said regarding the 1908-09 Pacific Coast trip and Carty's analysis of the engineering situation early in 1909, and the accounts of the subsequent development work, there should be no occasion for surprise in the statement that the extension of the westward limit of commercial telephony to Denver in 1911 set in motion a planned series of new development and research projects for the specific purpose of achieving transcontinental telephony. It should not be inferred, however, that early in 1909 a complete, detailed, plan had been worked out and approved, and was ready to be carried out on a rush basis upon receipt of the go-ahead signal from the chief executive. Otherwise, it would not have been necessary to engage in the new studies and the planning work during the period November 1910-April 1911 which preceded the beginning of the necessary new fundamental researches on the telephone repeater and line problems.

Mr. Carty's memorandum of April 9, 1909 to Mr. Thayer, previously mentioned, outlined a broad plan, one part of which involved learning how to load 165 mil open-wire circuits, but knowing in advance that that step would not get beyond Denver. The other part, of greater basic importance but much more complicated and less predictable, involved the repeater research problems. These parts were carried out in orderly sequence.

The short story of the transcontinental line given in the Alfred Bigelow Paine biography of Mr. Vail, "In One Man's Life" (1921), indicates that business policy and financial questions were factors in timing the go-ahead signal to Carty. This account, however, blithely ignores the engineer's technical problems in developing the necessary new instrumentalities, and the inherent uncertainties. Emphasis was placed on costs. Some of the directors conceded it would be a good thing to do, but believed the venture would not pay. Vail finally decided, "Oh, well, if it is a good thing let's do it, anyhow". He was looking forward to a fulfillment of the company's objective of universal service which had been clearly pictured in the original incorporation papers of the American Company in 1885.

It should be emphasized here that the transcontinental project planning by the engineers did not wait until the Denver line had been placed in commercial service, since long before then they were confident of its success.

By November 1910 the development situation in relation to transcontinental telephony had begun to crystallize, and during the following months definite plans were made.

PRELIMINARY TRANSMISSION STUDY

In this planning work, it was logical that the possibilities of reaching the Pacific Coast on loaded lines without repeaters should again be considered. A transmission-cost study made by Jewett and his associates (F.B.J. Memo of December 6, 1910 to Gherardi) demonstrated that it would be economically unsound to attempt New York-San Francisco transmission without developing an entirely new type of telephone repeater, then considered to be a development possibility.

On a non-repeater basis, to provide a satisfactory grade of transmission according to standards of that day, it would have been necessary to use a phantom circuit on #5 BWG wire (220 mil diameter, 774 lbs. per wire mile) having entirely new types of insulators and loaded with new types of ultra-high-efficiency coils, all the way from coast to coast. On the other hand, with a suitable type of repeater used in conjunction with loading, it should be possible to use the existing 165 mil wires between New York and Denver in conjunction with new 165 mil wires from Denver to the Coast. Jewett believed that the cost of the repeater solution would be very small relative to the cost differences between the two types of lines. The concluding paragraphs of the December 6, 1910 memorandum are quite revealing:

“As a result of this preliminary study, I am more than ever impressed with the very great need for producing a satisfactory repeater for operation on loaded lines if we are to establish a truly universal service on the North American continent on a paying basis as well as one of true economy.

“From a preliminary study of the situation, I feel very confident that if this repeater matter is tackled in the proper manner by suitably equipped men working with full coordination and under proper direction the desired results can be obtained at a relatively small cost. I feel, however, that to achieve this result it will be necessary to employ skilled physicists who are familiar with the recent advances in molecular physics and who are capable of appreciating such further advances as are continually being made, also that the work must be carefully supervised by some one having a full understanding of the requirements.”

REVIEW OF REPEATER SITUATION

This project transmission study occurred at a time when the repeater problems were being carefully reviewed and analyzed, looking backwards, and forward.

Looking backwards, it will be recalled that mention has been made on several occasions in this story of the efforts which were being made to learn how to use the mechanical telephone repeater on loaded circuits. Little

has been said, however, regarding the results of this work, which was carried out entirely on loaded cables, because there was no progress in commercial utility to report. Nevertheless progress had been made, but of a somewhat negative character.

Much had been learned regarding the inherent limitations of the repeater element itself.

The loaded circuits of that era which were sufficiently regular for good transmission without repeaters were being found not to be sufficiently regular when used with repeaters. Moreover, similar types of loaded circuits in the same cable were found to have large differences in their impedance characteristics.

In all of the experiments the 21-type repeater circuit was used. In this circuit, the line sections between which the repeater works must be closely equal in impedance, in order to avoid transmission distortion by interaction of the input and output currents, which may be very disastrous to intelligibility, especially when attempts are made to operate two or more repeaters in tandem in the same circuit.

It thus happened that the limitations of the mechanical repeater element, the repeater circuit, and the lines, combined to accentuate each other's effect in piling up the practical difficulties.

A PLAN EVOLVES

The November-December 1910 analyses of what had been done and what remained to be accomplished resulted in an engineering decision to renew the attack on the repeater problems on an "all-out" basis, according to a plan which would be designated as a "four-prong" offensive in the military language of today. The following statement of this fundamental plan is quoted from the Work Order No. 7655, "General Repeater Study," prepared by Jewett, and which was officially approved by Carty on April 1, 1911.

"Nature of Work

"A general study to determine the proper characteristics for the best telephone repeater, its circuit, and the general terminal and line conditions that must be fulfilled to make this repeater available for both loaded and non-loaded lines. This study will include—

"(1) A complete study of the characteristics of the existing receiver-transmitter (Shreeve) type of repeater with a view to determining whether the action of this repeater cannot be improved upon and whether modifications in the repeater element, its circuit or in the line conditions will make it suitable for general use on loaded lines.

"(2) A study of other possible repeater ideas, particularly in the domain of molecular physics. Certain characteristics of discharge of electricity through gases and vapors seem to offer considerable possibility of obtaining a telephone amplifier

that will be suitable for use on loaded or non-loaded lines and which will give the desired amplification without a great deal of distortion.

“(3) A mathematical and laboratory study of two-way repeater circuits with a view to determining the best form of repeater circuit to be used in combination with any desired repeater element and any kind of loaded line.

“(4) A mathematical and experimental investigation of loaded line characteristics in the existing plant, and a determination of what changes, if any, must be made in the construction and installation of loading coils and cables in order to make loaded lines suitable for the application of telephone repeaters.”

The executive decision to make the “all-out” attack on the repeater problems as an economically necessary step to transcontinental telephony was largely influenced by Jewett’s advice and his confidence in the possibilities offered by the recent advances in electron physics. The decision proved to be very timely. By starting in time, and by continuous, vigorous, activity in the allied research, development, engineering, manufacturing, and construction problems, it was possible to have the Transcontinental Line ready for commercial service when the Pacific-Panama Exposition opened at San Francisco early in 1915. The advantages to all concerned in having the Transcontinental Line ready prior to the Fair are obvious.

RESPONSIBILITIES FOR THE WORK UNDER THE PLAN

Returning to the discussion of the plan, the specific program for group and individual responsibilities in doing the work conformed to proposals made by Jewett in his memorandum to Gherardi, “Repeater on Loaded Lines,” dated December 22, 1910, from which the following paragraphs are taken:

“As a result of my study of the matter it seems to me that as the results from all four investigations must ultimately mesh with an existing, or to-be-built, telephone plant, the general supervision of the problem, the formulation of the specific problems and the general coordination of the work can probably best be done by someone in this Engineering Department rather than by someone at the Western Electric Company.

“With regard to the probable best assignment of the specific divisions, it seems to me that the investigation of the present carbon-button type of repeater is clearly a problem for the Western Electric Company under the general supervision noted above. Also, it seems clear that as the investigation of new repeater principles will undoubtedly involve a large amount of laboratory work, this also is a matter for Western Electric investigation. As regards the other two, namely, the study of repeater circuits and the investigation of loaded line characteristics, I believe that best results will be obtained by having as much of the detailed work as possible done here. Primarily, the investigation of repeater circuits is a matter for theoretical and mathematical consideration, and secondarily, a matter for experimentation. In the making of this investigation it will be necessary to utilize all of the results obtained from the work on repeater element and line characteristics.

"My reason for thinking that the phase of the investigation which concerns the characteristics of loaded lines should be handled directly by us is that much work of this kind will have to be done by us in connection with our phantoming, phantom loading, duplex cable design, coil design, and superimposed telegraph work and unnecessary duplication can undoubtedly be accomplished by combining the two problems.

"The various phases of the problem are so interlocked that the utmost cooperation, between ourselves and the Western Electric Company, and between the various men engaged on the problem is absolutely essential to the accomplishment of successful results, and as under the present organization this Engineering Department is responsible for the specification of what shall or shall not go into the operating plant, I believe, as noted above, that the immediate supervision of the problem as a whole had best be located here. What I have in mind is that this problem should be handled in much the same way as the problem of phantom loading the duplex cables was handled. In this case part of the detailed work has been done by us and part by Western Electric Company, and while the general supervision has been with us there had been the utmost cooperation between all concerned, whether here or at the Western Electric Company.

"With proper handling and with proper men engaged on the various phases of the problem, I feel very confident that fruitful results should be obtained within a reasonable time."

It was most logical that the responsibility for the general direction of all this work should be assigned to Jewett in the beginning, and that he should continue to hold this broad responsibility when he was transferred to the Western Electric Company on April 1, 1912 to become an Assistant Chief Engineer, reporting to C. E. Scribner, the Chief Engineer.

ORGANIZING OF PERSONNEL TO DO THE WORK

The new work order was officially sent to the Western Electric Company on May 27, 1911, to cover that part of the work which had been delegated to the recently organized Research Branch of the Western's Engineering Department, under the direction of E. H. Colpitts, its first Research Engineer. Other Western engineering units joined in the development work later on.

Early in 1911, Jewett's own department acquired new personnel primarily to work on the "loaded lines characteristics" phase of the fundamental repeater study and on repeater circuit questions. These were R. S. Hoyt, transferred from the Special Development Laboratory of the Western Electric Company, and John Mills, fresh from a teaching job at Colorado College. At one time or other, practically all other members of Jewett's rapidly expanding department worked on the transcontinental line problems. As a research consultant, Dr. G. A. Campbell made important theoretical contributions. The American Company organization set-up as of December 1, 1911, after the 1911 crop of engineering graduates had been assimilated, is given on page 403.

The great organizational achievement was, of course, the creation of the Research Branch of the Western's Engineering Department, early in 1911. Jewett had important responsibilities in this planning, working in conjunction with Messrs. Carty and Gherardi of the American Company and Messrs. Scribner and Colpitts of the Western Electric Company. Jewett was also active in recruiting personnel. His most fruitful contribution to this phase of effort was the engagement of Dr. Harold D. Arnold, who eventually succeeded Colpitts as the Research Engineer, and who might have risen to positions of even greater responsibility but for his untimely death in 1933. The hiring of Dr. Arnold was the result of personal negotiations with Professor Robert A. Millikan of Chicago University, an old friend and former associate when Jewett was studying for his doctor's degree, and who had become widely recognized as a leading American expert in the whole realm of electronic physics. In a 1931 radio address by Millikan, Jewett's statement to him regarding the requirements for a satisfactory telephone repeater was quoted as follows:

" . . . Such a device, in order to follow all of the minute modulations of the human voice must obviously be practically inertialess, and I don't see that we are likely to get such an inertialess moving part except by utilizing somehow these electron streams which you have been playing with here in your research work in physics for the past ten years. . . ."

Millikan was requested to recommend a man whose familiarity with the electronic technique and whose character would qualify him as being competent to attack the Telephone Company's research problems on repeaters. In due course Millikan recommended Arnold, who was then working in the Ryerson laboratory for his doctor's degree, and Jewett sponsored him. He reported for work with the new Research Branch of the Western Electric Company early in January 1911, knowing what was expected of him. Arnold's outstanding personal contributions to the project, as discussed later, fully justified Millikan's confidence and Jewett's expectations.

The first Western Electric organization chart to show the new Research Branch is that dated January 1, 1912, page 404. The chart on page 405 shows the complete engineering personnel of the departments for which Jewett became responsible on April 1, 1912, as Assistant Chief Engineer. Another chart dated July 1, 1912, page 406, shows Jewett's departments in relation to the complete Engineering Department of the Western Electric Company.

B. ACHIEVING TRANSCONTINENTAL TELEPHONY

This section of the story summarizes the significant research and development efforts that solved the basic problems of transcontinental tele-

phony, culminating in the first talk over the New York-San Francisco line by Mr. Vail on July 29, 1914 and in the official opening of the line for commercial service on January 25, 1915. Many amusing stories are told of the efforts of the engineers to do their final testing on the line without transmitting any of their voices from coast to coast, the injunction having gone forth that under no circumstances was anything to happen that would detract from Vail's first talk.

In general, the present story does not discuss the personal contributions of individual engineers and physicists which were essential to the complete success of the project. Some information on these matters is available in an article, "The Line and The Laboratory," written by John Mills, and published in the January 1940 issue of the Bell Telephone Quarterly, along with other articles commemorating a quarter century of transcontinental service.

In the beginning, the Western Electric Research department and the Transmission Engineering department of the American Company were most actively engaged in the project. In the course of time, these departments were expanded to handle the increasing amount of work and other associated departments in these organizations became involved in the cooperative efforts. The engineers of the Long Lines department, and of the Pacific and Mountain States companies also, did their own very important jobs, and last but not least, so did the manufacturing organization of the Western Electric Company.

THE IMPROVED MECHANICAL REPEATER

The 1911 analyses of the then available form of Shreeve repeater, receiver-transmitter mechanical type, indicated the principal defect to be a very marked natural period about midway in the telephone talking range, in which range the amplification was very much greater than at low and high voice frequencies. This caused distortion and tendency to sing well within the audible range. Other serious defects were a variable amplification with different magnitudes of input energy, the amplification with low levels being markedly less than with high level input, non-linear distortion, and a tendency for periodically variable amplification from instant to instant. Inertia of the moving parts was a congenital handicap that could not be completely overcome. The diaphragm of its receiver portion had to vibrate at any and all speech frequencies, and simultaneously drive at the same vibration rates the movable electrode of the carbon-button transmitter.

The analysis just summarized resulted in design modifications which materially improved the performance characteristics. Specifically, these modifications improved the magnetic circuit, reduced the movable mass, and raised the natural period of the vibratory system to the upper part of

the voice range, approximately 2200 cycles. The sensitiveness of the repeater, which was greatest at its natural frequency, was reduced by means of a resonant shunt-type electrical filter of about the same critical frequency. This arrangement provided more uniform frequency-amplification characteristics and better quality. The modified repeater, however, was still not entirely satisfactory with respect to the variation of amplification with respect to input levels.

The 1912 improvements produced an amplifier element that was good enough for experimental use at Philadelphia in September 1912 on a loaded New York-Baltimore circuit in the New York-Washington cable, using the 22-repeater circuit subsequently described. Somewhat later there was an experimental installation on the New York-Denver loaded open-wire circuit, using the 22-type repeater arrangements with three repeaters in tandem. Further development work on refinements and auxiliary devices made the mechanical repeater fairly satisfactory from the quality, volume, and life standpoints, and several commercial installations were made during 1913 and 1914, including a number of points along the Boston-Washington cable, initially at Philadelphia, to improve and protect the service along this route.

The improved repeaters, code 3A, were used for a few days in the initial 3-repeater service (at Pittsburgh, Omaha, and Salt Lake City) on the transcontinental line in January 1915, as alternatives for vacuum tube repeaters installed at the same points. The New York-San Francisco service with the mechanical repeaters was fairly satisfactory, but not as good as that with the vacuum-tube repeaters, which were retained in service.

The inferior characteristics of the improved mechanical repeater, relative to the vacuum tube repeater, led to restrictions in its general use. The principal disadvantages of the mechanical repeater were those previously commented upon. These were such as to become more and more serious with increasing lengths of circuits, involving increasing numbers of repeaters in tandem. Moreover, even under most favorable operating conditions, the maximum repeater gain was well below that obtainable with the vacuum tube device. The inertia of its moving parts restricted its frequency range application to voice-frequency telephony. Within a few years after the opening of the transcontinental line no more installations were made, and vacuum tube repeaters were substituted in old installations. That is to say, the vacuum tube repeater soon became the standard.

NEW TYPES OF REPEATER ELEMENTS

(a) *The Mercury Arc Repeater*

The early theoretical survey of the possibilities of developing essentially inertialess telephone repeaters focussed attention on gaseous discharge

devices as having great promise, and Arnold's initial research efforts were concentrated in this field, using a mercury arc. The suggestion to use the mercury arc as an amplifier was old, having been advanced in this country by Peter Cooper Hewitt, the inventor of the mercury vapor lamp, but it had never proved to be feasible in a practical way, because of its variable amplification, inefficiency, noise, and distortion. Arnold contributed new features based in part on novel phenomena discovered in his experimental work. These substantially reduced or eliminated the defects above listed. Patents were issued to him in due course, and rights were also obtained under the Hewitt patents.

The basic element of the Arnold amplifier was a stream of ionized molecules of mercury vapor flowing vertically from the positive electrode at the upper end of an evacuated tube to the negative electrode, a pool of mercury at the lower end, the energy for maintaining the arc being furnished by a direct current power source which included in its circuit a stabilizing choke coil and a regulating rheostat. Within the tube there were two auxiliary side electrodes (cathodes) symmetrically disposed with respect to the axis of the tube and closely spaced thereto. There also was a starting electrode located within an associated condensing chamber. The ionized mercury vapor stream was vibrated transversely between the two auxiliary side cathodes, by virtue of the electromagnetic action of the input telephone current flowing through coils which were mounted on the pole pieces of an external electromagnet and so disposed that the axis of the magnetic field was perpendicular to the axis of the ionic stream. The output circuit included a transformer which had a split primary winding, with its two main terminals connected to the two side cathodes respectively, and its mid-point connected to the negative terminal of the d-c power source previously mentioned. When there was no input telephone current in the receiver coils the arc stream flowed steadily, and no current was induced in the secondary windings of the output transformer because the equal currents from the two side cathodes flowed in opposite directions in the two halves of the primary winding and inductively annulled each other's effect on the core. On the other hand, when a telephone current flowed through the magnet coils, the arc stream was magnetically deflected first to one side cathode and then to the other, depending upon the magnetic polarity. This caused changes in the magnitudes of the currents flowing in the individual halves of the primary winding of the output transformer; as the current in one half-winding increased that in the other half-winding necessarily decreased, and vice versa. The resultant induced current in the secondary winding of the output coil had similar frequency components to the incoming telephone current and much greater energy, which was supplied by the current from the external battery, or other power source.

The theoretical principles of the device were studied to provide a background for straightening out initial design kinks, and to provide suitable auxiliary apparatus for associating the amplifier with the telephone circuit. Considerable difficulty was encountered in securing a sufficiently long life of the mercury tube to permit experimental use. The first field experiments occurred late in December 1912 at Philadelphia, on loaded circuits in the New York-Washington cable. During the next two years a number of other experimental installations were made. Sentimentally significant experiments were made on the transcontinental line in the late spring of 1915, using three and sometimes four repeaters in tandem. In some respects the over-all transmission performance was good, but it was not so generally satisfactory as with the vacuum tube repeaters. Considerable difficulty was encountered in starting and maintaining the mercury-arc amplifiers.

By the end of 1913, it was becoming apparent that the high vacuum tube amplifier was destined to become the leading type. As a matter of fact, the development work on the mercury arc device began to slow down late in 1912, soon after the work started on the improved audion, as subsequently discussed. The mercury arc repeater was never used in commercial service. The experimental installations were dismantled during 1915, and the development case was officially closed in October 1916.

The success that was quickly achieved in the development of a satisfactory high vacuum tube repeater, as described in the following pages, leaves unanswered the question as to whether the mercury arc repeater could have been developed to become a thoroughly satisfactory voice-frequency amplifier. Its more complicated structure, its need for more complicated and more expensive auxiliary devices for associating it with working telephone circuits, and the greater difficulties in operation and maintenance were serious handicaps. Also, it was more limited with respect to repeater gains, and with respect to working-frequency band. The device as developed was used only as a voice-frequency amplifier.

(b) The High Vacuum Tube Amplifier

Arnold's research and development work on thermionic amplifiers started in November 1912 soon after a laboratory demonstration on October 30 and 31 by Lee deForest of the amplifying properties of his Audion tube, which for several years had been extensively used as a detector for wireless telegraph signals. The Audion as submitted was a much simpler device than the mechanical and mercury arc amplifiers. It consisted of an evacuated glass tube containing three elements: (1) a filament which emits electrons when heated by an external "A" battery, (2) a metal-plate electrode which attracts and collects these electrons when maintained at a

suitable positive potential by an external "B" battery, and (3) a grid electrode placed close to the filament in the path of electron flow to the plate electrode and used so as to exercise a very sensitive control of the electron stream. This third element was deForest's pioneering contribution to the art. In using the Audion as a telephone amplifier, the grid and filament terminals serve as the input terminals and the plate and filament terminals as the output terminals. Under proper operating conditions, variations in input voltage applied to the grid circuit so affect the flow of electrons as to produce amplified voltages in the plate or output circuit. Using energy drawn from the plate battery, an increase in energy is delivered to the output line. In deForest's recent use of the Audion as a radio receiving amplifier the grid circuit included a series condenser; this was also included in the arrangement offered as a telephone amplifier.

The laboratory demonstration had been arranged with Mr. Carty by John Stone Stone, a mutual friend and an independent research worker who had acquired a theoretical knowledge of telephone transmission problems when he was a member of the Boston headquarters staff of the telephone company during the nineties. Colpitts and Richards participated in the complete demonstration and Jewett in the final stages. This demonstration of the Audion was entirely qualitative in scope, under simple but adequate circuit conditions. With very low input levels, the speech currents were greatly amplified without perceptible impairment in intelligibility. However, when the speech input approached the levels that would be encountered in any commercial use of repeaters, the amplification was greatly reduced and very noticeable distortion and noise resulted. Under these conditions, a blue haze was prone to appear in the tube and it disappeared when the input level was reduced. As the plate potential was progressively raised, a permanent condition of blue haze developed and the device ceased to amplify.

On the following day, November 1, the Audion was called to Arnold's attention. He promptly repeated and extended the experiments, using deForest's tubes and auxiliary apparatus which had been loaned for that purpose. Arnold's broad training in electron physics enabled him without study or delay to explain the blue haze phenomena and to prescribe a remedy. The keynote of the explanation was that the blue haze was due to ionization of gas present in the device and the remedy was to secure a much higher vacuum. The medium vacuum in the Audion test samples was a normal result of the best evacuation processes then used by incandescent lamp manufacturers. Better results could be secured by laboratory processes known to research physicists and still better results could be expected from a new type of molecular pump then recently described in technical literature. Arnold's preliminary analysis also indicated that a

better type of filament, providing a more profuse emission of electrons and a much longer life, could be used in place of deForest's tantalum filament. All in all, Arnold painted an exceedingly intriguing picture regarding the practically certain prospects of developing a really good telephone amplifier from the deForest Audion. There was, of course, considerable chagrin that these prospects had not been recognized much earlier in the Telephone Company's research work on repeaters, but no time was wasted in attempts to develop alibis. Arrangements were made for Arnold to spend most of his time on the vacuum tube job and several assistants were provided. From then on, the work on the mercury arc amplifier element slowed down. To get full freedom in the development and use of the improved audion, patent rights were purchased from deForest and later on, as commercial use approached, it became desirable to obtain patent rights from other outside inventors, American and foreign, who had been working on electronic repeaters.

Arnold was the first worker in the electronic field to determine the physical laws of operation of the 3-element high vacuum tube, this being his initial personal contribution to the development. In the concurrent and subsequent experimental work it was found that the grid condenser of the deForest circuit was a basic factor, along with the previously mentioned blue haze, in the paralysis of the Audion as an amplifier, when large input currents of the magnitudes involved in wire telephony were employed. Under these conditions it was found that the grid condenser acted as an electron trap which, by piling up a negative charge on the grid, would cut off the plate current and block the tube, even if the vacuum should be high enough to prevent blue haze phenomena. Consequently, the condenser was eliminated from the grid circuit. For some time, a grid leak was substituted (a very high resistance, grid to plate). This was subsequently replaced by a battery inserted in the grid circuit to maintain the grid at a positive potential relative to the filament; this held the grid impedance to a definite value and improved stability. Early in 1914, Arnold began using a negative "C" battery in the grid circuit and thereby increased the sensitivity and the stability. This led to the potentiometer input method of controlling gain. Still later, it was learned that Lowenstein had anticipated Arnold in the use of the negative C battery, and patent rights were obtained.

In carrying out the development of the vacuum tube repeater by "methods of pure science which were brought to its study in the spirit of research," many workers were involved and a host of problems had to be solved. A new manufacturing art had to be created in the research laboratory. Optimum conditions and means for connecting the vacuum tube elements into the repeater circuit had to be worked out. From the beginning, these involved the 22-type repeater circuit, discussed later on.

The work on the vacuum tube repeater during 1912 and 1913 was summarized in the 1913 report on Work Order 7655 as follows:

“ . . . The result has been the ability to construct an Audion amplifier to give distortionless amplification between desired limits of current input; to give outputs of energy far above any value that is normally met in telephony; to act as a potential or as a current transformer with the ability when connected two or more in series of giving current amplifications of as large amounts as 50 times or more; to present to the circuits between which it works practically constant impedance. The present form of the Audion gives practically perfect repetition and amplification of currents delivered to it.”

It is of interest that on February 18, 1913 a laboratory demonstration of the promising possibilities of the audion-type repeater was made for President Vail and other executives on a 900-mile non-loaded artificial 104 mil open-wire line. This was a one-way test having several repeaters in tandem, and the tubes did not have a high vacuum, due to limitations of the then available apparatus in the laboratory. The first use of the high vacuum tube amplifier on commercial circuits was on October 18, 1913 when a 22-repeater installed at Philadelphia was placed in service on a New York-Baltimore loaded cable circuit.

Further developments resulted in improved amplifiers becoming available in the transcontinental line when Mr. Vail first talked over it in July 1914, and when commercial service started the following January. The important over-all service characteristics of the line, including the part played by the repeaters, are considered in a subsequent section of this story under the heading, “The First Transcontinental Circuits.”

REPEATER CIRCUITS

The development work on repeater circuits for the transcontinental project had for its basic theoretical background a classical mathematical analysis of the relations between line impedance irregularities and repeater gains in two-way repeater circuits, reported by Dr. G. A. Campbell in May 1912. The study included the currently used two-way, one-repeater circuit (21-circuit), and the two-way, two-repeater circuit (22-circuit), which had not been commercially used, although it had been invented by W. L. Richards in 1895, long before the availability of a commercially usable repeater element. In a preliminary report dated March 7, 1912,⁷ Campbell had recommended the 22-circuit on the basis of its greater stability and unrestricted flexibility, as subsequently discussed. The March 7 memorandum also discussed the four-wire repeater circuit which later became very

⁷ A complete copy of this memorandum is given in the last item listed in the attached Bibliography.

important commercially in the application of repeaters to long loaded, small-gauge, toll cables.

The maximum repeater gain which could be obtained in two-way circuits without sustained singing that would block transmission, or near-singing that would degrade intelligibility, was expressed by Campbell as a function of the differences of the impedances of the circuits involved.

In the 21-circuit, the single repeater element amplifies transmission in both directions, and its usable gain is a function of the differences of the impedances of the circuits between which it works.

In the 22-circuit where the two repeater elements each function as one-way repeaters, the usable gain is a function of the differences of the impedances of the line East and the artificial line required to balance it, and of the differences of the impedances of the line West and its own balancing artificial line. In the general case, involving lines with irregular impedance-frequency characteristics, twice the power amplification feasible with the 21-circuit would be allowable if the lines should be connected through a 22-circuit, assuming the use of balancing lines having "average" impedances as described later. Statistically, the average balance obtainable between any one line of a given type and the "average" balancing line here assumed in the 22-circuit is 3 db better than the average balance obtainable between any single line and others of its type as involved in the use of the 21-circuit. In the limiting theoretical case, which may be approached but not attained in practice, if one of the lines using a 22-repeater should be perfectly balanced by its associated artificial line, singing could not occur at that repeater irrespective of the degree of the unbalance between the impedance of the other line and its associated artificial line.

The foregoing discussion is adequately summarized in the statement that when high repeater gains are required the lines using the 22-repeater do not need to be so uniform in their impedance frequency characteristics as would be necessary with 21-repeaters. This was very important in the trans-continental telephone project because of the serious practical difficulties involved in the reduction of line impedance irregularities to very small values, as discussed later.

Furthermore, the 22-circuit has an overwhelming superiority in stability over the 21-circuit, under conditions that require the use of repeaters in tandem. This follows from its characteristic property of transmitting the amplified energy in one direction only—the desired direction away from the source—whereas each 21-repeater transmits in both directions and one half of its amplified energy starts backwards towards the source, thereby setting up among the successive repeaters objectionable circulating currents which impose severe restrictions on the repeater gains.

The 22-circuit also has important practical service flexibility advantages in that the lines between which it works may be of radically different types, provided each line has associated with it an artificial balancing line having closely similar impedance-frequency characteristics. This flexibility feature also permits the 22-repeater to be used as a terminal repeater. Since in such service the terminal impedances (switching trunks and loops) vary over a wide range, operating flexibility requires the use of a compromise impedance balancing network instead of one that simulates the line. In consequence, the terminal repeater gains are restricted to values much smaller than those obtainable with intermediate repeaters.

In considering Campbell's proposal to use the 22-circuit instead of the 21-circuit, it initially appeared that the artificial balancing lines for use with loaded circuits would have to be complicated multi-section loaded lines which themselves would tend to possess appreciable irregularities in their own impedance-frequency characteristics. Since the technical difficulties involved were critical handicaps, and because of adverse cost factors, the question arose as to whether a simpler form of balancing network could be devised. The study of this problem resulted in the development of a simple 3-element 2-terminal network, to balance a regularly loaded line terminated at about 0.2 fractional section. To provide flexibility in use, a simple procedure was devised for building out this "basic network" to match the actual line termination when different from approximately 0.2 section termination. For example, if the loaded line should be terminated at mid-coil, i.e., with a half-weight loading coil, the basic network would be built out to full-section using a shunt condenser of proper capacitance and then a series inductance equivalent to the half-coil would be inserted in tandem. An alternative procedure would be to build out the line at the repeater station.

The simple basic network above mentioned consists of a fixed resistance equal to the nominal impedance of the loaded line, in series with an inductance shunted by a capacitance, these elements being proportioned to shape properly the reactance component of the impedance-frequency characteristic. The special virtue of the line termination chosen for the basic network design is that at 0.2-section termination the resistance component of the characteristic impedance of a regularly loaded line is approximately constant over the most important part of the working-frequency band.

It was this possibility of constructing a simple balancing network instead of a complicated loaded artificial line which made practicable the use of the 22-repeater circuit.

The simple type of basic network above mentioned was first used in 22-repeater circuit trials on loaded circuits in the Boston-Washington cable. Different proportioning of the elements was of course required for the later

uses with loaded open-wire lines. Later on, when the 22-repeater circuit was used on non-loaded lines, new types of simple basic networks were devised to simulate the impedance of the non-loaded lines. It is also appropriate at this point to mention the fact that when commercial service started with 22-repeater circuits, the balancing networks of the repeaters included not only the basic networks, and building-out devices when required, but also apparatus for balancing line terminal apparatus which otherwise would have contributed objectionable impedance irregularities to the line. Such auxiliary apparatus usually included a repeating coil to balance the line repeating coil, and a simple 4-terminal network to balance the composite telegraph sets, when involved.

An additional very important new feature provided in 1912 for the 22-repeater circuits was the use of a low-pass electrical wave filter in the branches of the circuit where each repeater element functions as a one-way amplifier. As their cut-off frequency was about 300 cycles below that of the loading cut-off, these filters suppressed the unwanted and unneeded frequencies near and above the loading cut-off, thereby making these frequencies negligible factors in repeater singing phenomena. This was of special importance because the simple basic balancing networks above described did not simulate loaded line impedance at these frequencies, and had added importance where vacuum tube repeaters were involved in consequence of their tendency to amplify the same amount at all frequencies.

Subsequently, low-frequency filters were included in the 22-circuit to suppress unwanted frequencies below approximately 250 cycles, in which band are occasionally present currents resulting from the operation of superposed telegraph circuits, and noise current produced by induction from power circuits. Line irregularities are not important factors in the repeater balance problem at these frequencies when proper types of basic networks are used.

The line experiments made with the improved mechanical repeater and the mercury arc and vacuum tube repeaters and the commercial installations previously mentioned used the 22-type repeater circuit. Different types of auxiliary apparatus (input and output transformers, etc.) were of course required to obtain optimum results with the different types of repeater elements. When the experiments involving tandem 22-repeaters showed objectionable impedance irregularities to be caused by the impedances presented by the repeaters to the line, improvements in the repeaters were made to reduce these effects.

In due course, the 22-repeater became the standard two-way repeater. The use of the 21-repeater was restricted to special situations not requiring more than one repeater of relatively low gain and located near the middle of the circuit.

IMPROVING THE LINE

When the development work for the transcontinental project started it was realized on the basis of the earlier work that impedance irregularities in the existing types of loaded circuits were large enough to set objectionable limitations upon the gains obtainable with two-way telephone repeaters. The recent availability of the Vreeland mercury arc, variable frequency, oscillator had made it possible to get a considerable number of impedance-frequency curves at close frequency intervals, but as yet the specific irregularities in the impedance curves had not been correlated with their individual causes. This prior work had been concentrated on loaded cables.

For a considerable time the new line studies were concentrated on loaded open-wire lines. Since the spacing irregularities were known to be as great or greater than in the cables, and since other sources of irregularity such as intermediate and entrance cables were also present, it was expected that the open-wire impedance-frequency curves would be even more irregular than the cable curves. And such was found to be the case, but in a much greater degree than anticipated.

Inductance Irregularities: The discussion will initially be directed to the inductance irregularities, since they proved to be the principal problem. In the course of the line measurements it happened that one set of impedance curves had a very unusual systematic sequence of ups and downs with rising frequency. This was especially intriguing since the usual curves had non-systematic bumpy characteristics. The cause was found to be an omitted load at a particular load point. This incident resulted in the development of a formula for estimating the position of an impedance irregularity in terms of the frequency spacing of resulting bumps in the impedance-frequency curve and the velocity of transmission.

A comprehensive series of impedance measurements were then started on a long loaded artificial cable at the laboratory, since it was much more simple to measure than to compute. Also, the magnitudes and circuit position of individual and multiple irregularities could easily be controlled. Very valuable data were collected in this manner.

When the loaded open-wire measurements were resumed, it was noticed that the changes in the impedance-frequency irregularity patterns of particular lines changed substantially from time to time. These changes were found to be due to large inductance changes, up or down, in individual loading coils, and the cause was eventually found to be the magnetizing or demagnetizing action of strong transient line currents induced by lightning discharges. In lines exposed to lightning, sooner or later the inductance of all exposed coils would drop well below the factory value. A coil thus partially magnetized by one shock would sooner or later be partly demagnetized by a subsequent shock and these experiences would be repeated again

and again in different degrees. Moreover, there were no systematic relations among the effects on coils at different points in the same circuit, or on coils in different circuits at the same loading points. These effects usually occurred without mechanical injury to the coil windings, or to the associated lightning arresters with which each coil was protected against breakdown, and it was this fact that had delayed recognition of lightning as being the probable cause. Confirmation of this deduction was obtained when tests made on lines unexposed to lightning showed that the coil inductances were close to their factory adjustment values.

Laboratory tests showed the magnetizing effects of lightning surges to be of the same order as the residual magnetizing effects of superposed direct currents, ranging up to several amperes in amplitude. The high magnetic retentivity of the continuous wire-type toroidal cores of the loading coils was a basic factor in these phenomena.

The necessity for accepting exposure to lightning surges as a normal service experience for open-wire loading coils forced consideration of the practicability of providing new designs having much greater magnetic stability. Experimental work was started on (non-magnetizable) solenoidal type air-core coils having finely sectionalized windings, and on toroidal wire-core coils having series air-gaps in their magnetic circuit to decrease its retentivity.

Fortunately the statistical study which was made to determine the limits that should be placed upon individual line irregularities in order to avoid undesirable restrictions on the repeater gains showed that it would not be necessary to use perfectly stable coils, i.e., the air-core coils. The concurrent work on the wire-core coils with air-gaps had indicated that by properly proportioning the air-gaps the inductance changes that should be expected from magnetization by lightning surges could probably be kept to tolerably low values. A single air-gap would have been sufficient to provide the required stability, but crosstalk considerations and other factors made it desirable to have two air-gaps symmetrically located at diametral points in the toroidal cores. The resultant designs were better in all important respects than the air-core coils and were inferior only with respect to magnetic stability, which difference as above noted was tolerable. To assist in the control of the inductance deviations in the lines, the new loading coils were manufactured to $\pm 1\%$ precision inductance limits. On the older standard designs, $\pm 6\%$ manufacturing deviations had been allowed. The new coils had somewhat lower nominal inductance values than the old coils, so that their average service inductance values after partial magnetization by lightning surges would be about the same.

The size advantage of the wire-core coils with air-gaps made it possible to pot the three loading coils for an open-wire phantom group, connected as a

phantom loading unit, in a 3-compartment case. This was the beginning of the use of "phantom loading units" in open-wire loading. To provide installation flexibility, cases were also developed for individual side and phantom loading coils. The over-all dimensions for all cases were small enough to avoid limitations on the number of circuits that could be loaded at the same loading points along any line. Double-pole H-fixtures, however, were required on routes having a large number of wires.

These air-gap type loading coils and their cases remained standard for open-wire loading until subsequent developments in the art, notably improvements in the repeater and the use of open-wire carrier systems, resulted in the gradual abandonment of open-wire loading.

Following the transcontinental project, wire-core coils with air-gaps were also developed for use on coarse-gauge duplex cables of the Boston-Washington type. It is also of interest to note that while the work on the transcontinental project was still under way a very good start was made on the development of the compressed magnetic powder core-type loading coil for small-gauge cables. The high stability characteristics of this general type of coil became an important factor in the wide use of telephone repeaters in the long distance cable plant.

Spacing Irregularities: Taking up the consideration of the impedance irregularities caused by loading spacing irregularities, the need for a substantial improvement was duly proved, and precision limits of ± 2 per cent in the spacing were established for lines to be used with high-gain repeaters, starting with the transcontinental project. In general, the required precision in new lines could be secured by proper engineering care, including more uniform transposition layouts, since coordination with the coil spacing was necessary. In applying the new high stability coils along old routes, as was necessary on the transcontinental line sections east of Denver, relocation of many of the loading points was found to be desirable. In these loading rearrangements and sometimes also on new lines, it was occasionally found to be desirable to tolerate the use of geographically underlength loading sections and build out their total capacitance to the theoretically desirable values by using shunt condensers. Mica condensers having suitable dielectric strength were used for this purpose, protected by loading coil type lightning arresters. In lines used for phantom working, a network of six condensers was used to provide the optimum building-out capacitance for the side circuits and their associated phantom. Subsequently, building-out condensers and stub cables also found use in the repeatered loaded cable plant for correction of objectionable spacing deficiencies.

Incidental Cables: In improving the open-wire lines for repeater operation, substantial development and engineering effort was also devoted to the reduction of impedance irregularities caused by incidental cables. Es-

pecially on the transcontinental line, such cables were avoided when practicable, and those that could not be avoided were made as short as practicable. In some instances long incidental cables were avoided by locating the repeaters in a test station at the city outskirts—Brushton (Pittsburgh) and Morrell Park (Chicago), for example.

Several types of treatment were applied to incidental cables that could not be avoided. Short cables having capacitances materially less than that of an open-wire loading section were taken into account in the layout of the open-wire loading. This method was also applied to bridle wire at test stations.

Long cables were provided with a new type of impedance-matching loading. The coil inductances and spacings were such that the loaded cable would have about the same nominal impedance and the same cut-off frequency as the loaded open-wire circuits, these being the requirements for minimizing the junction reflection effects. This was the first use of impedance-matching loading on incidental cables in loaded lines. Previously, it had been the general practice on entrance cables to use some standard weight of toll cable or exchange cable loading, for example on the Boston-Neponset cable described earlier in this story. These former loading practices reduced the cable attenuation and the junction reflection loss, both being desirable objectives, but resulted in junction impedance irregularities large enough to be objectionable on repeated circuits. In due time, it was found desirable also to use an extra light-weight impedance-matching loading for incidental cables in non-loaded open-wire lines, when used in conjunction with telephone repeaters. Later on, this need became especially important in lines used for carrier telephone systems, and suitable types of high cut-off, impedance-matching, carrier loading, were developed.

Line Insulation: Because of the high impedance of the loaded line and its great length, it was particularly important to keep the leakage losses as low as possible. An interesting problem in this connection was the effect of the salt on the line in the vicinity of Great Salt Lake in Utah. The Mountain States Company equipped itself to take care of the situation by using steam from the boiler of an old Stanley steam automobile to clean the insulators, when necessary.

THE FIRST TRANSCONTINENTAL CIRCUITS

The construction of an entirely new phantom group between Denver and San Francisco began during the summer of 1913 via Rawlins, Salt Lake City, Winnemucca, Sacramento, and Oakland. An interesting account of the construction problems is given in an article, "The Circuits Go Up," by H. H. Nance and R. M. Oram, which is one of a series of articles commemorating a quarter century of transcontinental service, published in the

January 1940 issue of the Bell Telephone Quarterly. The ceremonies that occurred at the time the line was opened for public service and the subsequent series of demonstrations of transcontinental service are also described in that article.

The new line had all of the latest improvements to provide regularity in the loading. These ideas were also applied to the lines east of Denver, making use of the new high-stability loading coils, and including respacing of the loading points where desirable. Also, a great many changes in transpositions were necessary. In addition to the open-wire phantom group between New York and Chicago, there was one between Boston and Chicago via Buffalo, and a New York-Buffalo phantom group for use as part of an alternate route to Chicago. Philadelphia, Baltimore, and Washington were also connected to this network by additional phantom groups or by non-phantom pairs to Pittsburgh.

As previously mentioned, the first coast-to-coast conversation occurred on July 29, 1914 when President Vail spoke the first words to cross the continent. The engineering tests that preceded this ceremony had been made on long isolated sections of the circuit. In the period that preceded the opening of the New York-San Francisco circuits for public service on January 25, 1915, a great deal of work was done to make the lines suitable for commercial service and to train personnel in the operation and maintenance of the lines, including the repeaters. In some sections, noise troubles⁸ had to be cleared by special transpositions. Crosstalk conditions required considerable attention. Some of the last minute improvements in the repeaters and other apparatus were utilized.

The New York-San Francisco circuits as first used commercially had some temporary or experimental features, especially as regards the repeaters. For several weeks, the transcontinental circuits used three intermediate repeaters located at Brushton (Pittsburgh), Omaha, and Salt Lake City. Then additional repeaters were used at Morrell Park (Chicago), Denver, and Winnemucca. Later on, permanent repeaters were substituted for the experimental repeaters. This change from three to six intermediate repeaters was made primarily to obtain greater flexibility in operation and to provide long-haul service, including leased telegraph service to some inter-

⁸ This reference to noise reduction work on the transcontinental line makes it appropriate to mention at this point the very important fundamental investigation of inductive interference between electric power and communication circuits which was made in the period 1913-1917 by the "Joint Committee on Inductive Interference" appointed by the Railroad Commission of the State of California in 1912. The field engineering staff which planned and conducted the technical studies and prepared reports thereon included engineers from the transmission division of the American Telephone and Telegraph Company Engineering Department. Mr. H. S. Warren made important contributions in the planning and conduct of the investigation, and was elected to an honorary membership of the committee. The conclusion and principal reports of the work have been published and widely used (refer bibliography).

mediate points. In this connection, it should be remembered that the transcontinental circuits were not through circuits, ready upon call, but were built up by switches at two or three intermediate points as required.

It is of interest to note that the change from three repeaters to six repeaters did not significantly affect the overall transmission performance. The gains in the individual repeaters had to be reduced in order to avoid objectionable interaction effects.

The transmission performance and circuit data given below apply to the New York-San Francisco circuits, having six vacuum tube repeaters in tandem:

Over-all Length.....	3359 miles
Transmission Losses	
Bare Line.....	53 db
Apparatus.....	7 db
Over-all, Line and Apparatus.....	60 db
Total Repeater Gain.....	40 db
Net Equivalent.....	20 db
Over-all Transmission Time.....	0.067 second

The net equivalent above given is a dry weather value. Under bad weather conditions the line loss approximately doubled. Adjustment of the repeaters, made manually, was required to keep the overall equivalent within reasonable bounds.

The transmission band that was effectively transmitted ranged from about 350 to 1250 cycles, defining the transmission band as that between the lowest and highest frequencies whose transmission was not more than 10 db higher than that of the transmission of 1000 cycles. At frequencies between 400 and 1000 cycles, the over-all loss was appreciably less than at 1000 cycles. At frequencies above 1250 cycles, approximately 50% of the theoretical loading cut-off frequency, the line losses including the loading coil losses piled up so as to effectively suppress transmission. The excess transmission losses at the low voice frequencies were due to losses in the line terminal apparatus (repeating coils, composite sets) and in the repeater auxiliary apparatus. Although the 900-cycle frequency band effectively transmitted by the transcontinental circuits was only about $\frac{1}{3}$ as wide as that required by the present standards for long distance transmission, it was acceptable to the early users of the service, and did not noticeably handicap the public interest in the large number of country-wide demonstrations that were made in the "Hello, Frisco!" era.

In the concluding section of his report to the American Company stockholders for the year 1914, President Vail appraised the significance of the transcontinental line and related developments as follows:

"It is a long step from a hardly intelligible telephonic conversation between two rooms, to a perfectly easy, low-voiced conversation between the extremes

of our land, East, West, North, South. Remarkable as this is, the progress made during the epoch of which this was the culminating point has been still more remarkable, but so quietly has it all been accomplished that it has been hardly appreciable. During the past ten years more has been done to increase the utility and availability of the telephone service, more has been done to increase its reliability, and greater obstacles have been overcome, than during its whole preceding existence.

“What has been accomplished perhaps never will be surpassed, the present contains the germs of the future development. Commercial practicability will be more controlling in the future than technical practicability.”

CHAPTER IV

The Establishment of a Transcontinental Network of Repeated 165 Mil Lines

This concluding part of the transmission development story is mainly concerned with the establishment of a country-wide network of 165 mil lines interconnecting all important cities, following the completion of the New York-San Francisco line. The American Company engineering department took the initiative in this work, and the Western Electric research and development groups handled their parts of the work under Jewett in accordance with their usual organization responsibilities. During the early part of this period, December 1916, Jewett became the chief engineer of the Western Electric Company. Early in 1921, he became Vice-President of the Western Electric Company in charge of the telephone department, having over-all responsibility for engineering and manufacture.

PLANNING A BACKBONE NETWORK

Immediately after the opening of the New York-San Francisco line, the making of plans to exploit the new developments and the new engineering knowledge got well under way. On March 1, 1915, Carty approved the American Company Work Order 8230, "Network of No. 8 Gauge Circuits Equipped With Telephone Repeaters Connecting All Important Cities of the United States." The work was to consist of:

- (A) Determination of the best routes, and the sequence of installation.
- (B) Determination of changes in loading and transpositions to fit these lines for repeater use.
- (C) Choice of repeater equipment and circuit arrangements.
- (D) Determination of the best operating methods.

When this project was authorized it was expected that loaded 165 mil open-wire circuits would be universally used in the backbone network. This part of the plan, however, was soon modified in consequence of transmission studies which showed even more attractive possibilities in the use of non-loaded 165 mil lines having additional repeaters to make up for the increased line losses. The expected advantages of this proposed change were:

- (a) Elimination of the telegraph flutter impairments that were quite troublesome on the long loaded circuits.
- (b) More uniform attenuation and impedance characteristics under varying weather conditions.

- (c) Better quality of speech transmission obtainable by the effective transmission of a much wider frequency band.
- (d) A reduction in costs.

Additional advantages of non-loaded lines for voice-frequency telephony became apparent from subsequent studies and experiments:⁹

- (1) The practicability of securing materially lower net losses, in consequence of the effect of the higher velocity of transmission in reducing disturbances caused by echo currents.
- (2) A reduction of delay distortion, resulting from the fact that the velocity of transmission of the upper speech frequencies in the non-loaded line is approximately constant with frequency, whereas in the loaded line the velocity decreases substantially with rising frequency, especially near the cut-off.

In September 1915, quantitative line tests were made to verify the theoretical expectations per items (b) and (c), above. The tests involved a comparison of the transmission over a non-loaded New York-Denver circuit, via Pittsburgh and St. Louis using six intermediate repeaters, with that over the New York-Denver section of the loaded transcontinental line using three intermediate repeaters. The non-loaded circuit had a somewhat lower net loss and noticeably better quality, and, of course, a complete freedom from transmission distortion by telegraph flutter. As a result of these tests, plans for using loaded lines in certain parts of the proposed backbone network were quickly modified to call for non-loaded lines and additional repeaters. In some instances where the lines were already loaded with old types of coils that were susceptible to magnetization by lightning surges, the change in plans avoided the expense of installing new high-stability type loading.

CHANGE IN ENGINEERING ATTITUDE TOWARDS LOADING

This decision to use non-loaded 165 mil lines instead of loaded 165 mil lines for parts of the continental backbone network was of great significance in that it marked the beginning of a new engineering attitude with respect to the fields of use for loading and for repeaters. Up to that time, loading had been accepted as the dependable and indispensable method of improving transmission in long distance circuits and extending their range, and repeaters had been regarded primarily as auxiliary devices for stretching the transmission benefits obtainable with loading. It will be recalled that, prior to the transcontinental development project, the type of repeater

⁹ The non-loaded lines also had important possibilities in the application of carrier telephone systems, the commercial development of which got well started during the 1915-1920 period under consideration in Chapter IV.

which was available could not be used as an adjunct to loading, and even on non-loaded lines its use was greatly restricted.

From 1916 on, the vacuum tube repeater was recognized in its own right and potentialities as an independent instrumentality for improving transmission. For nearly a decade, repeaters and loading were competitors in the open-wire plant, sometimes used together as a team. Beginning in 1916, loading was removed from many 165 mil lines on which it was planned to use repeaters, and this practice continued at an accelerating rate during the early twenties to facilitate the exploitation of open-wire carrier telephone systems. GEC-812, issued June 1918, definitely discouraged the provision of loading on new 165 mil circuits. On 104 mil circuits, however, the competition between loading and repeaters was much closer. Partly due to production limitations on repeaters, the mileage of loaded 104 mil circuits increased rapidly during the war period, and reached a peak about 1923. Not long afterwards, the practice of loading 104 mil circuits stopped and the removal of existing loading accelerated, so as to provide maximum plant flexibility for the use of repeaters and open-wire carrier telephone and carrier telegraph systems.¹⁰ On all types of non-loaded lines used in conjunction with repeaters, however, loading continued to have an important function in the transmission treatment of the unavoidable incidental cables.

In the long distance cable field, repeaters and loading were continuously developed over a period of more than two decades to work together as equal partners in a team, each making its own optimum contribution on a basis that provided the desired over-all transmission performance at about the minimum total cost.

DEVELOPMENT WORK

Returning to the evolution of the continental backbone line network in terms of repeatered non-loaded 165 mil lines, it was fully appreciated at the beginning that because of the increase in the number of repeaters and the changes in the line, improved types of repeaters and auxiliary apparatus would be required. The principal need was an improvement in the gain-frequency characteristic. This involved among other matters a reduction of the frequency distortion characteristics of the auxiliary apparatus. It was also found desirable to stabilize the repeater gain and to improve the impedance of the repeater presented to the line so that it would more closely match the line impedance.

During 1917 and 1918, analyses of extended tests on repeatered lines and cables laid the foundation for computation techniques that enabled the over-all transmission performance of repeatered circuits to be predicted with

¹⁰ It was not until 1934, however, that the use of open-wire loading ceased completely.

close accuracy. The principal factors were found to be the velocity of transmission, the number and spacing of repeaters, the line attenuation, and the reflections at significant points of irregularity; i.e., at the line terminals, and at the repeaters. These studies and line tests clearly demonstrated the importance of the transmission velocity in echo current phenomena, and the limitations on transmission performance imposed by the echo currents.

To achieve transcontinental transmission on non-loaded 165 mil lines became an objective in the repeater development work. Although this development was started in good time, the pressure of war work interfered so that not much progress occurred until late in 1918, when arrangements were made for a trial of the improved repeaters on a non-loaded circuit between New York and Chicago. The success of this tryout in 1919 led to arrangements being made for unloading the transcontinental circuits west of Chicago. At the new repeater points west of Chicago entirely new repeaters were installed; those used at other points and in the new non-loaded Chicago-New York circuits were modified to have equivalent transmission features, including 3000-cycle filters, which became a characteristic feature of the 22-repeaters for non-loaded lines. Different sections of the transcontinental line became available on a non-loaded basis at intervals during the spring of 1920.

THE UN-LOADED TRANSCONTINENTAL LINE

The through circuits had a total of twelve intermediate repeaters. The net loss was about 11 db, or 9 db below that of the loaded line, and the effective transmission band was twice as wide. The expected improvements in stability under varying weather conditions were realized in full. The better repeater balances that were obtainable with the inherently more uniform non-loaded lines were factors in the greatly reduced net loss. The factor of fundamental importance, however, was the approximately 3.5 to 1 increase in the velocity of transmission, which shortened the time interval between the direct transmission and echoes from points of impedance irregularity and thereby reduced the disturbing effects of the echoes. When the unloaded transcontinental circuits were demonstrated to a conference of Bell System presidents held at Yama Farm, N. Y., on May 25, 1920, the "sense of nearness" made possible by the high transmission speed was especially commented upon as a component of the improved transmission quality.

Following the unloading of the transcontinental line, other 165 mil circuits on important routes were unloaded, and soon there was a complete backbone of non-loaded 165 mil circuits operated with the improved repeaters.

CONCLUSION

The end of the present story is at hand.

It has shown the principal high spots of a most fruitful period in the development of long distance telephony, either when the specific accomplishments are taken into account in their own right or when they are considered as foundations for the subsequent developments which first made universal telephone service on the North American continent economically practicable, and then by radio links the inclusion of this network within a world-wide international network.

The prediction in the Carty memorandum of 1909 that a successful telephone repeater would also unravel the problems of radio telephony was amply fulfilled six years later.

The advent of the high vacuum tube repeater closed the era under review and has of course proved to be an outstanding legacy for the succeeding years. The initial aim during its commercial development was to overcome distance, then to improve the transmission standards. In time, these efforts made excellent transmission performance substantially independent of distance.¹¹

As these objectives were approached and realized, the general development emphasis turned in various ways towards the reduction of the costs of the long distance facilities and the provision of much larger circuit groups; by the application of carrier telephone systems to open-wire lines, incidentally requiring the removal of the remaining voice-frequency loading; by the extensive use of repeated, loaded, cables along routes where large groups of circuits were required, and where stability of service had become very important; and in recent years, by the application of wide-band carrier telephone systems to open-wire and to non-loaded cable, and to coaxial conductor systems. In these developments, the repeater played its own basic part in offsetting attenuation; and other new developments, in particular the electric wave filters, distortion corrective networks, and regulating devices played their parts in shaping and controlling the transmission medium. These various developments were basic to the improved speed of service and the reduced rates, which, along with the high quality transmission standards, have stimulated an ever increasing demand for long distance telephone service.

Some important aspects of these developments which followed the first transcontinental line, and other associated developments, are described in Jewett's article, "Transcontinental Panorama," published in the January

¹¹ The following pertinent quotation is from the 1922 company report to the stockholders: "In faithful reproduction of speech at a distance so that the person listening will understand with ease, so that the speech transmitted will be of proper volume and quality without distortion, our engineers and scientists have achieved what seemed to be the impossible. On the through lines, distance has been eliminated."

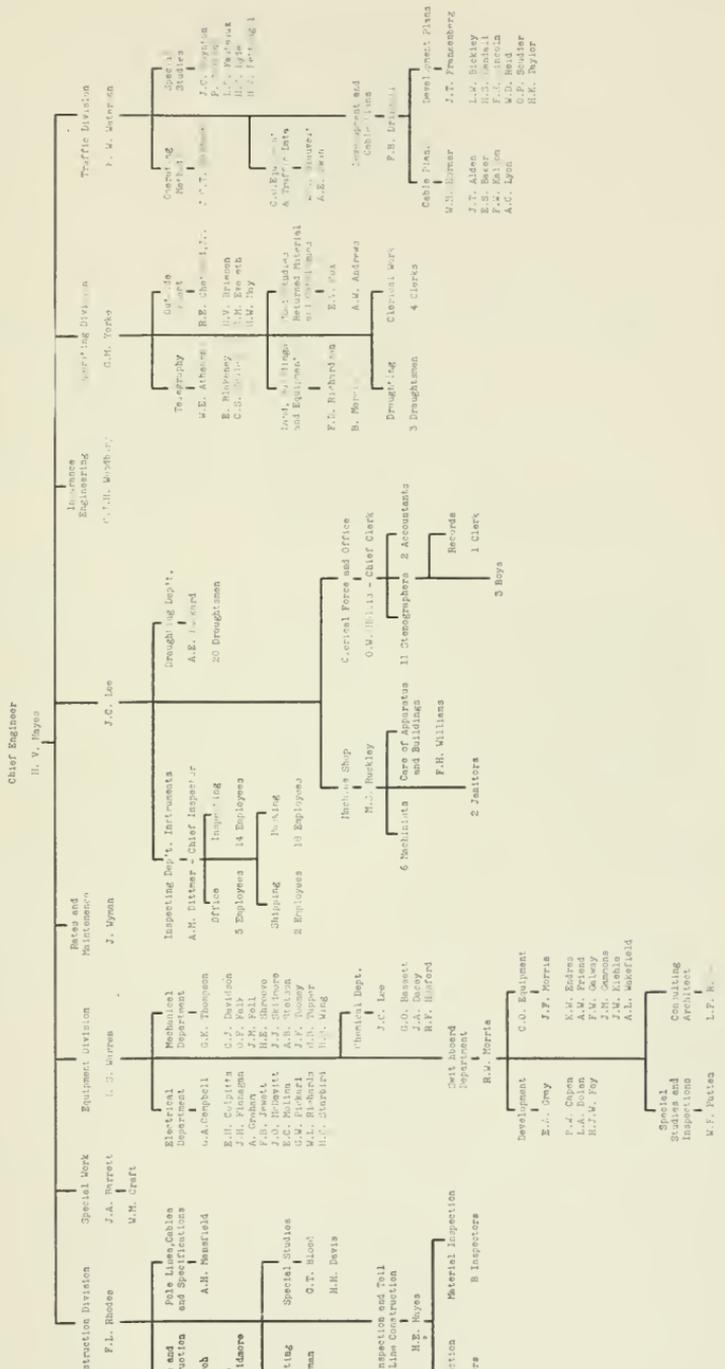
1940 issue of the Bell Telephone Quarterly to commemorate a quarter century of transcontinental telephone service. Its subtitle is revealing: "From the Solution of the Specific Problems of Transcontinental Telephony Have come Changes and Advances in the Art Which Have Affected Every Aspect of the Service." The author of that story is the individual who had the responsibility of leadership for the success of the transcontinental telephony project.

It is a project which has continued to grow mightily until the original phantom group has become some five hundred circuits distributed over five cross-country routes four of which are open-wire lines, and the fifth and latest an underground cable, with what might be called reverberative effects that have worked as leaven in every phase of telephony.

And finally, as the closing paragraph, I quote the conclusion of Jewett's own article, because it so well typifies the spirit of the man himself and also expresses in a minimum of words what has unquestionably been his most important contribution, viz., an organization of research scientists and development engineers and designers working harmoniously as a composite mind upon a single problem of vast technical ramifications as well as infinite details.

"It has been my good fortune to have had a part in a great adventure, some of whose principal features I have attempted to sketch out for you. I would be less than honest, and far less than generous, however, if I allowed any of you to depart with a false impression of my personal contribution. The achievements embody the contributions of many men, my associates (some of whom I do not even know), working through the years as a team of which I have been a member."

ORGANIZATION
 ENGINEERING DEPARTMENT
 American Telephone and Telegraph Company
 195 Employees - January, 1935.



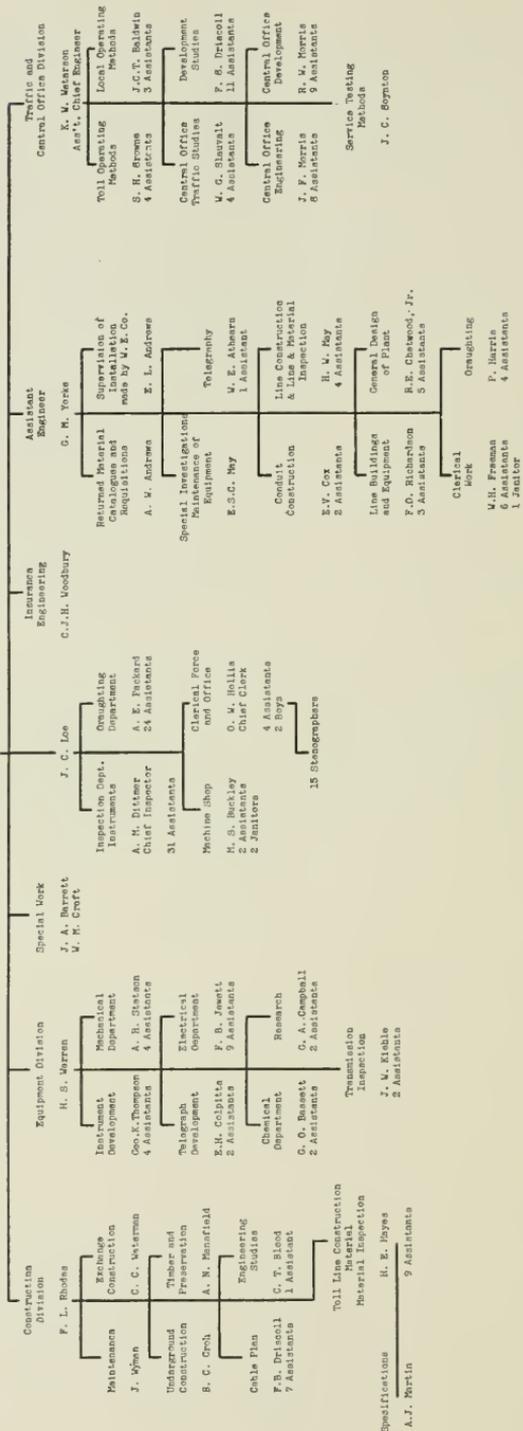
ORGANIZATION

ENGINEERING DEPARTMENT

American Telephone and Telegraph Company

236 Employees - June, 1907

Chief Engineer
Harwood V. Hayes



ORGANIZATION

ENGINEERING DEPARTMENT
AMERICAN TELEPHONE AND TELEGRAPH CO.

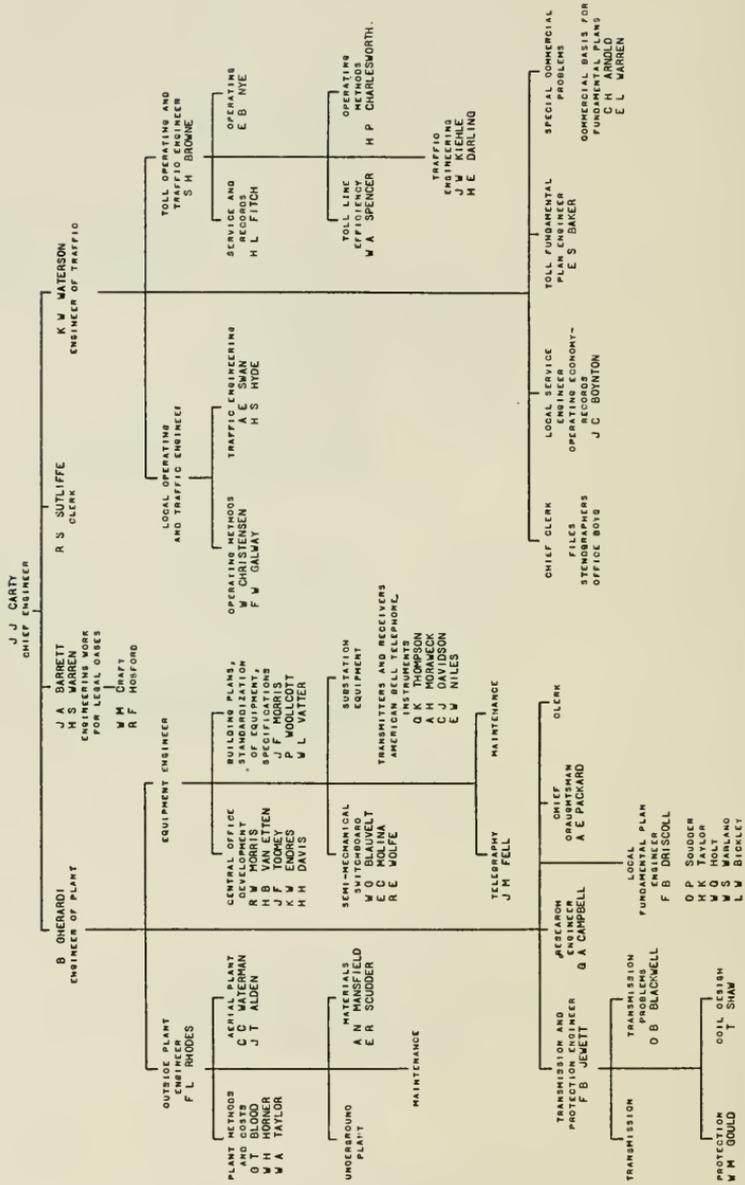
DECEMBER 26, 1907.

Chief Engineer

J. J. Gerty.

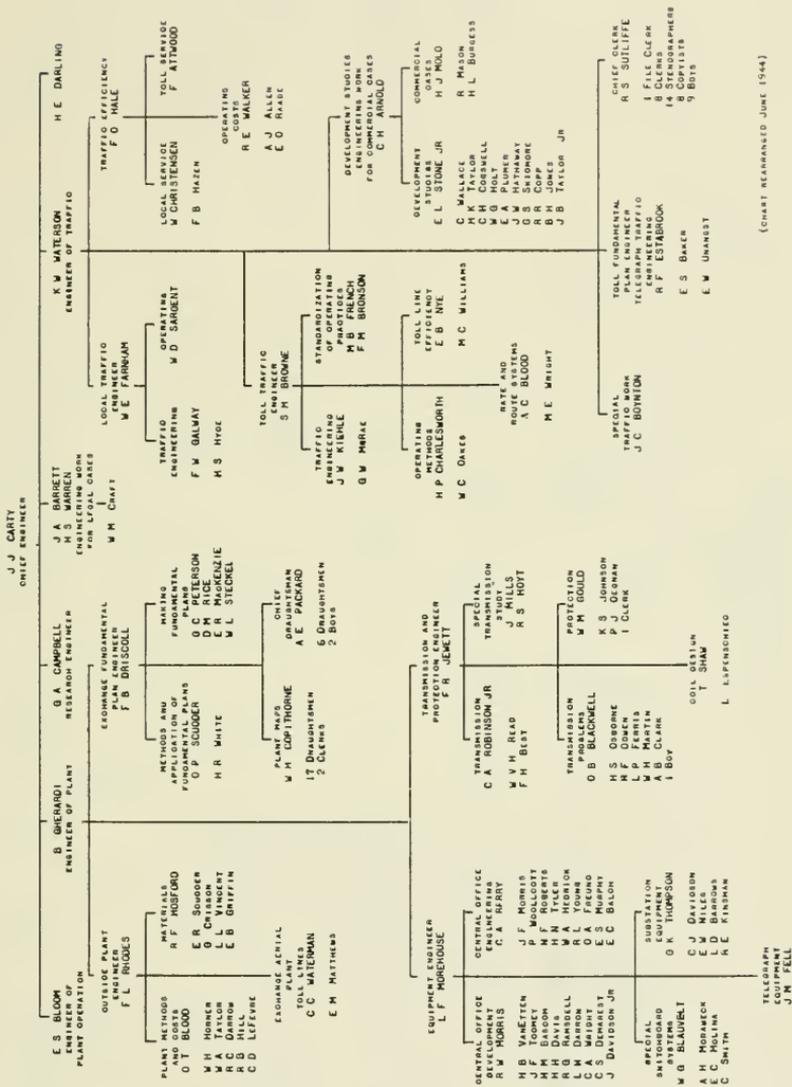
	<u>Outside Plant</u>	<u>Special Work</u>	<u>Power Interference</u>	<u>Electrical Buildings</u>	<u>Equipment and Buildings</u>	<u>Toll Traffic</u>	<u>Fundamen-tel Plans</u>	<u>Clerical</u>	<u>Draughting</u>
<u>General Problems</u>	F. L. Rhoads	C. H. Arnold	J. A. Barrett W. M. Craft	H. S. Merran	B. Cherdard				
<u>Construction</u>	C. C. Mesterman J. T. Alden								
<u>Line Material</u>	A. N. Mansfield H. A. Taylor								
<u>Apparatus</u>	C. K. Thompson E. H. Hites								
<u>Research</u>									
<u>Telegraphy</u>	J. M. Fell								
<u>Standardiz'g</u>									
<u>Equip. & Buildings</u>									
<u>Local Traffic</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									
<u>Chief Clerk</u>									
<u>Files Clerks</u>									
<u>Steno-graphers</u>									
<u>Boys</u>									

ORGANIZATION
ENGINEERING DEPARTMENT
AMERICAN TELEPHONE AND TELEGRAPH COMPANY
MARCH 15, 1909



(CHART REARRANGED JUNE 1944)

AMERICAN TELEPHONE AND TELEGRAPH CO
ENGINEERING DEPARTMENT
191 EMPLOYEES DECEMBER 1, 1911



BIBLIOGRAPHY

In addition to the references in the text, the following will be of interest. In some instances, the articles are partly or largely concerned with developments during the period covered by the story, although the publishing dates are years later. In other instances involving late publication, the articles are of special interest in correlating the early developments with those subsequent to the period covered by the present story.

- B. Gherardi, "The Commercial Loading of Telephone Circuits in the Bell System," *Trans. A.I.E.E.*, June 28, 1911.
- B. Gherardi, "Some Recent Advances in the Transmission Efficiency of Long Distance Circuits," (Published Notes of a Talk before the Telephone Society of New York), April 18, 1911.
- F. B. Jewett, "Long Distance Telephony in America," Pamphlet, Paper presented to International Congress of Applied Electricity, Turin, Sept., 1911.
- "Inductive Interference Between Electric Power and Communication Circuits," Selected Technical Reports With Preliminary and Final Reports of the Joint Committee on Inductive Interference. Published by the Railroad Commission of the State of California, San Francisco, Calif., April 1, 1919.
- B. Gherardi and F. B. Jewett, "Telephone Repeaters," *Transactions, A.I.E.E.*, October 1, 1919.
- G. A. Campbell, "Physical Theory of the Electric Wave Filter," *Bell System Technical Journal*, Nov., 1922.
- R. S. Hoyt, "Impedance of Smooth Lines and Design of Simulating Networks," *B.S.T.J.*, April, 1923.
- H. S. Osborne, "Telephone Transmission Over Long Distances," *Transactions A.I.E.E.*, October, 1923.
- H. H. Nance, "Some Very Long Telephone Circuits of the Bell System," *B.S.T.J.*, July, 1924.
- R. S. Hoyt, "Impedance of Loaded Lines and Design of Simulating and Compensating Networks," *B.S.T.J.*, July, 1924.
- George Crisson, "Irregularities in Loaded Telephone Circuits," *B.S.T.J.*, Oct., 1925.
- T. Shaw and W. Fondiller, "Development and Application of Loading for Telephone Circuits," *B.S.T.J.*, April, 1926.
- A. B. Clark, "Some Recent Developments in Long Distance Cables in the United States of America," *B.S.T.J.*, July, 1930.
- F. B. Jewett, "Carty—The Engineer and the Man," *Bell Lab. Record*, Sept., 1930.
- Wm. R. Ballard, "The High Vacuum Tube Comes Before the Supreme Court," *Bell Lab. Record*, July, 1931.
- M. A. Weaver, "The Long Struggle Against Cable Crosstalk," *Bell Telephone Quarterly*, Jan., 1935.
- F. B. Jewett, "Dr. George A. Campbell," *B.S.T.J.*, Oct., 1935.
- "The Collected Papers of George A. Campbell," A.T.&T. Co., 1937: Introduction, by E. H. Colpitts; "Three Early Memoranda on Loading," Page 10; "Repeater Circuits," page 533.

APPENDIX I

EXTRACTS FROM ANNUAL REPORT* OF THE ELECTRICAL DEPARTMENT FOR THE YEAR 1905

The problems upon which the main work of the year has been spent have been protection, disturbance from alternating current railways and the inspection of commercial transmission conditions. These together with the correspondence have consumed about one-half of the time of the department if the routine work such as the inspection of loading coils is excluded from consideration.

The attempt has been made this year to make the ledger numbers a more correct and more detailed record of the work of the department than has been the case in the past. There may be some question as to whether this is worth while, but I have thought it desirable to carry the plan through the entire year and in accordance with this plan I quote below the list of orders with the amount of time which has been charged to each. This will take the place of any special list of the problems which have come up during the year, and it will assist in showing how nearly eleven years work has been divided among these problems. The list gives the date when the ledger number was opened, the ledger number, the title of the number and the number of days charged to the ledger number during the year. In order to indicate somewhat the character of the work on each subject, the time charged has been divided into two classes. One including the work which has been on the whole independent and original and may be referred to as development, (D), and the other the work which has been more of the nature of assistance, (A). All figures are brought up to December 20th, 1905.

* * * * *

PROTECTION

Jan. 27, '05	3787	Protection of Substation Sets.	79D.	41A.
Mar. 2, '05	3860	High Frequency Tests on #7 Fuse.	3D.	3A.
Mar. 16, '05	3896	Rating of Fuses.	2D.	3A.
Apr. 17, '05	3928	High Tension Protection.	132D.	3A.
Apr. 17, '05	3930	Protection Against Lightning.	9D.	0A.
Mar. 31, '05	3952	Examination of and Specifications for Protector #73-A.	2D.	0A.

The work on protection has been in charge of Mr. Jewett, who has had a large amount of correspondence to attend to in this connection, in addi-

* Letter, G. A. Campbell to H. S. Warren, 12/30/1905.

tion to the work indicated by the ledger numbers listed above. He reports as follows:

During the year a large number of experiments have been conducted with a view to ascertaining the amount of protection to terminal apparatus and the immunity from fire hazard afforded by the present standard protectors. These experiments have shown that, aside from one exception in the case of substation protectors, the present protective apparatus properly installed is capable of furnishing adequate protection to the central office and substation equipment for crosses of any potential.

With regard to the fire hazard incident to the operation of the protectors, the results obtained from the experiments showed that an appreciable danger from fire existed when the voltage of the circuit crossed with the telephone lead exceeded a certain value. In comparison with the high potentials now generally employed by light and power companies, this critical voltage was very low.

To decrease the hazard of a fire resulting from the operation of the protectors and to secure a substation protector on which the maintenance from lightning would be a minimum, an entirely new system of protection was developed and its elements subjected to these tests included numerous operating trials at the General Electric Company's Works at Lynn, and also two sets of tests at the power house of the New Milford (Connecticut) Power Company. In these latter tests the protective elements were subjected to a potential of 33,000 volts under such varying conditions as would be met with in practice.

In connection with the development of the system of complete protection and in conformity with the results obtained from the tests, the following pieces of apparatus have been designed:

- (1) Substation Protector—providing for open-space cutouts and impedance coils—to be used at substations the lines of which are subject to severe static disturbances.
- (2) Outside Fuse—to be used at substations on lines which are exposed to high tension circuits.
- (3) Protector Mounting—to be used in cable boxes where open-space cutouts are required.
- (4) Cable Terminal—provided for open-space cutouts and impedance coils—for use at cable terminals where the entering lines are subject to severe static disturbances.
- (5) Metal Block Arrestor—an open-space cutout for use in connection with 1, 3 and 4.

These pieces of apparatus have been tested under laboratory conditions and are now being manufactured by the Western Electric Company. A large number are to be installed by some of the licensee companies and their

operation and maintenance carefully watched during the coming summer. It is hoped that the data obtained in this way will afford sufficient information to enable us to so modify our present protective practice as to afford adequate protection in all cases, with a reasonable amount of maintenance on the protectors.

* * * * *

APPENDIX II

Memorandum from Dr. Jewett to Mr. Warren:

December 22, 1906.

I have enumerated below under a number of different more or less general headings the most important things upon which we have been engaged during the past year. Subjects marked with an asterisk (*) are those upon which work is still being done although in most cases one or more reports have been filed on certain phases of the investigation.

CABLES.

- *1. Work in connection with Conference Case #23-A on the development of cable for loading.
- *2. A design of a switchboard cable for #2 Private Branch Exchanges.
3. A study of wool insulated cable for switchboard use.

COILS.

- *1. A general study of iron for retardation and repeating coil cores.
2. The development of coils for use in a new form of Private Branch Exchange circuit.
3. Thorough investigation of the efficiency of repeating coil T-602 (25-K) for use as a terminal transformer on loaded lines.
4. The development of a method for manufacturing balanced phantom repeating coils.
- *5. The development of a shellac insulation for loading coil core wires.
- *6. The development of a submarine loading coil.
- *7. The re-designing of an aerial loading coil.
- *8. The development of extra light loading coils.
9. The development of a compensating coil for use on telephone lines exposed to alternating current railway induction.

CONDENSERS.

1. A study of the alternating current capacity and conductance of #21 type condensers.

DISTURBANCES (Noise and Electrolysis).

1. A test at Derry, Pennsylvania, on the Scott compensator for reducing disturbances on telephone lines exposed to alternating current railway induction.
2. Test at East Pittsburg, Pennsylvania, of the Scott compensator for reducing disturbances on telephone lines exposed to alternating railway induction.
3. An investigation of the induction caused by alternating current railways along the lines of the Indiana and Cincinnati Traction Company.
4. A study of the Providence, Warren and Bristol trunks and preparation of the proper treatment of these lines to reduce excessive noise troubles.
- *5. A study of the induction and electrolysis troubles likely to result from the alternating current electrification of the New York, New Haven and Hartford Railroad.
- *6. A study of alternating current electrolysis.
- *7. A study of battery crosstalk under varying conditions.

INSULATION.

1. A study of enamel wire insulation in connection with various forms of apparatus.
2. A study of enamel insulated wire submarine cables.
- *3. A study of the insulation afforded by different types of aerial wire insulators under different weather conditions.

LOADING.

- *1. The development of equipment for new loading coil testing room at the Western Electric Company's factory in New York.
2. A study of the loading of the Boston, Salem and Beverly low capacity cable.
3. A study of extra light loading for terminal and trunk cables.
4. A study of the loaded cables used for inter-office trunks in Kansas City.
- *5. A general study of load coil efficiencies.
6. A study for a loaded cable between Albany and Schenectady.
7. A study for a loaded cable between Jordan City and Salt Lake City.
8. A study of loading for inter-office trunk cables in Chicago.
9. A study for loaded toll cables in Kansas City.
10. A study for loaded toll and trunk cables between Brockton, Taunton and Middleboro.

11. A study for the loading of open wire circuits for the New England Telephone and Telegraph Company.
12. A study for the loading of open wire circuits for the Southern Bell Telephone Company.

MEASURING APPARATUS.

1. The design of a new artificial cable using Ward Leonard enamel resistances.
2. The development of a method for determining transmission equivalents by means of an alternating current dynamometer measurement.
3. The development of a design of a universal receiver shunt for use in determining line equivalents.
4. The design and construction of a capacity unbalance testing set for the Central District and Printing Telegraph Company.
5. The development and construction of a testing set for use in transmission investigations.

PROTECTION.

1. The completion of a design on the experimental system of high potential protection for substations and the preparation of a circular letter on the same.
2. The development and construction of a high potential protected cable terminal.
3. An investigation of the Birsfield central office protector.

REPEATERS.

1. A study of the distortion introduced by the presence of telephone repeaters in an open wire circuit.
2. A study of the efficiency of telephone repeaters when used in tandem on long haul circuits.
3. The design of an induction telephone repeater.

SWITCHBOARDS.

- *1. A study looking to the development of fire-proof switchboard resistances.
- *2. A study for the development of an efficient type of Private Branch Exchange circuit.

TELEGRAPH.

1. A study of phantoplex interference and a development of remedies for the same.
2. An investigation of the Field quadruplex telegraph.

3. Work and tests connected with the development of the Bleakeney-Chetwood balance for use on telegraph lines subjected to heavy alternating current induction.
4. Tests of Mr. Athearn's circuit for use on lines subjected to alternating current railway induction.

TRANSPOSITIONS.

1. The development of a transposition system for short spur and loop lines.
2. The development of a transposition system for short toll lines.
3. The development of a phantom system of transpositions for the aerial line between Philadelphia and Atlantic City.
4. The preparation of a complete set of phantom transposition drawings and the preparation of a circular letter on phantom transpositions.

TRANSMISSION.

- *1. Transmission study of the Boston and Maine circuits.
2. A study of the bridging effects of various types of ringers.
3. Transmission tests on the Boston and Wellesley loaded cable.
4. Transmission tests on the Boston-Hingham loaded cable.
5. Transmission investigation of the Boston-Chicago repeater circuits.
6. Transmission tests on the Grant-Brushton loaded cable.
7. A transmission investigation of the 48-volt exchange at Bryn Mawr, Pennsylvania.
8. A study of the conditions affecting long distance service from Buffalo, Rochester and Louisville.
9. A study of the conditions affecting long distance service between the Union Pacific Railroad Company's private branch exchange in New York and the Illinois Trust and Savings Bank's private branch exchange in Chicago.
- *10. The commencement of a general study of the conditions affecting long distance transmission.
11. The determination of the transmission losses in various types of cord circuits.
12. The determination of the transmission losses in various types of outside distributing wire and switchboard cable.

MISCELLANEOUS.

1. A study of the effect of multiplying call wires and terminal offices.
- *2. A study of multiplex telephone (Weintraub system).
3. A study of the Cooper-Hewitt rectifier for use in charging storage batteries.

4. A study of dielectric cables.
5. The construction of a new capacity balance room for the laboratory at 15 Oliver Street.
- *6. An investigation of the properties of low hysteresis iron.
7. A study of the costs involved in various types of construction for the New York, New Haven and Hartford Railroad electrification.

In addition to the above work which has been more or less in the nature of experimental and theoretical studies a large amount of time has had to be devoted to general correspondence since the number of outgoing letters during the past year is practically 50% larger than the number of outgoing letters in 1905. For the period from January 1st to December 20th, 1905, the number of outgoing letters was 768, while for the same period in 1906 the number of outgoing letters was 1120.

(Signed) F. B. JEWETT.

APPENDIX III

A. T. & T. Co. ENGINEERING SITUATION, APRIL 1909

In a memorandum dated April 8, 1909 to Vice-President Thayer, to whom he reported, Mr. J. J. Carty then Chief Engineer of the American Telephone & Telegraph Company gave an appraisal of the engineering situation as it appeared to him.

The following partial copy of Carty's memorandum includes in full his discussion of the transmission items on which Dr. Jewett and his department worked. The discussion of Outside Plant, Equipment, Traffic, and Operating items is not included. Other parts of the memorandum are also omitted from the attachment as not being important pertinent background material for a story primarily concerned with Dr. Jewett's career. Rows of asterisks indicate deleted material. (T.S.)

April 8, 1909.

MEMORANDUM for Mr. Thayer,
Vice President.

I have considered the best way to strengthen our forces so as to properly carry out the work which will necessarily devolve upon this department in consequence of the comprehensive and definite relations now being established with the associate companies. The work in connection with the new relations will of itself be of very great and far-reaching importance and will entail upon part of the department a great deal of labor. While all of the various things constituting this new work have been duly considered with you and while they are beyond question clearly within our functions, it is necessary that we should plan for these new duties on such a scale that

we will not be obliged to neglect the very important classes of work which we are already engaged upon.

Before speaking of the work we now have in hand or which, under the arrangement heretofore existing pertained to this department, I will briefly outline the nature of the new duties as far as they have yet been developed. First, the department is being reorganized so as to have three principal divisions reporting to the Chief Engineer:

1. A department relating to legal, protection, and railroad and power and electric light interference matters. Work under these headings has for a year or more been going on, but under the new scheme a more effective contact with the associate companies will be possible and required.

* * * *

2. A department under the Plant Engineer, establishing standards of plant construction and maintenance, issuing them to the various Plant Superintendents as fast as such officials are created. These things include not only the most economical and efficient methods of construction, but also the most economical and efficient methods of maintenance.

* * * *

3. A department under the Traffic Engineer, establishing standard methods of operating, standard traffic engineering methods, standard traffic records and reports.

* * * *

While it will be seen that that portion of the work above outlined which is new is of great magnitude and importance, the department must be strengthened in order that, in taking care of the new duties, the existing work will not suffer. In judging then of the necessity for the proposed increase in the strength of the department, it will be helpful if we consider a few of the items of work upon which we are already engaged, or which, even under the old scheme of organization, we should undertake. After so doing, I think it will be seen that when we compare the increased payroll with the increased results which will be obtained, the additional sum of money to be expended is a moderate one.

Some of the work already in hand or which would be undertaken in due course is described below and an estimate of the savings which would be accomplished by its successful performance is given.

* * * *

PHANTOM CIRCUITS AND DUPLEX CABLES:

At the present time phantom circuits can be established only upon overhead lines and even there, where their value has been so abundantly demon-

strated, they are restricted to overhead lines which are not loaded. Inasmuch as a great and increasing number of overhead lines must pass through cables at some point or points, the necessity for a satisfactory method of carrying phantom circuits through cables has assumed great importance. Inasmuch as the loading of overhead lines has already reached large proportions and as, in view of our recent work, this importance will rapidly increase, we must, unless some method is devised of loading phantom circuits, sacrifice the advantage either of loaded circuits or of phantom circuits. Neither of these alternatives should be tolerated. We believe and we confidentially predict that if we can place the organization of this department upon a proper basis, we will develop new types of cable for carrying phantom circuits and will devise and standardize methods for loading phantom circuits, both in cables and overhead. Aside from the relation which the phantom or duplex cable has to overhead phantom circuits, it is important of itself, even where these overhead circuits are not to be considered. This is because we are entering the period of long distance underground cables and the savings which may be made in these cables themselves, growing out of working them duplex, reach into very large figures. As soon as the department is strengthened so as to respond to the extension of its duties now projected, a special study will be made of physics and economics of long underground cables, say between Boston and New York and New York and Washington. We will also make a similar study with respect to the possibilities of universal long distance service throughout the United States. While we can already see as an outcome of these proposed studies matters of very great importance and promise, and while we can also see that in these directions great use will be made of loaded phantom circuits and of phantom circuits in cables, we have thought it best in the present memorandum not to count upon any savings which might lie in these directions, but have restricted ourselves to those savings which would follow in the natural course of events, even if our larger ideas were not realized. Thus, counting upon the new toll cable construction which will be done from year to year, we figure that the use of a successful duplex cable would reduce our construction costs for this class of work at the rate of about \$180,000 a year.

We are confident that if we are provided with the necessary ways and means, we can devise a method whereby we can phantom loaded circuits. By such an achievement the circuit capacity of the existing plant of the American Telephone and Telegraph Company and the associate companies could be increased to an extent which would be attainable under the present state of the art only by an additional plant investment of \$2,500,000, on which the annual charges would be \$250,000. Assuming that we have thus increased the carrying capacity of the plant, additional circuits, by means

of loaded phantom circuits and duplex cables, could be obtained at a much cheaper rate, which we estimate, with the amount of construction running along about as at present, would represent a saving in construction costs of \$500,000 each year, on which the annual charges would be \$50,000.

As has already been stated, a large use of phantom circuits has already been accomplished where unloaded lines and open wires are used. Even in those cases the phantom circuit work has not attained anything like a full measure of success. Take just the territory of one associate company, it was expected that by the use of phantom circuits, a saving in plant investment of \$1,000,000 would be made. It was found upon investigation by this department that owing to faulty location with respect to electric power wires, nearly half of these circuits were too noisy to be operative, so that only part of this saving was realized. We believe it is possible to remove the cause of this failure in such cases. This would require special and detail studies conducted by the plant department of each associate company, such as were recently made by this department in the territory of the Pacific Telephone and Telegraph Company. The savings which could be made in this way throughout the country cannot be estimated, as we have no proper system of reports from, or relations with the plant departments of these various companies, and because in many of these companies such a thing as a plant department does not exist. It is clear that an enormous saving can be made in this way, but it could be accomplished only through improvement in organization all along the line and not by laboratory work or the mere issuing of standards.

FURTHER DEVELOPMENT OF PUPIN INVENTION:

Although most extraordinary savings have already been obtained from the use of the Pupin invention for loaded circuits, resulting in the case of the New York and the New York & New Jersey Companies in a reduced cost of construction of more than \$7,000,000, we have not yet exhausted the possibilities of this class of circuits. Up to the present time it has not been practicable to load circuits as large as the No. 8 wires composing the New York-Chicago circuit. I was greatly impressed during my trip to the Pacific Coast with the great advantage which would accrue to the company operating there if loaded No. 8 circuits were practicable. Troublesome opposition companies exist in the southern and the northern part of the territory, and a vigorous opposition is opening in San Francisco, the heart of the territory. Long toll lines are being projected and some of them are being built by the independent companies. If loaded No. 8's were available, a first-class talk could be given between the most widely separated of the important places, say between Seattle and Los Angeles, a distance of about 1500 miles. Over this distance, loaded No. 8's would give excellent

transmission and for a distance of 2000 miles or a little more, practicable talk could be given over them. Loaded No. 8's would tie this scattered territory together, would greatly redound to the prestige of the Pacific Company and would place the opposition concerns at a very great disadvantage. Even if no use were made elsewhere of such circuits, their importance in this territory would be so great, looked at from every point of view, that the cost of the work which we propose to do upon this subject becomes an utterly insignificant factor. But it is not only in the territory of the Pacific Coast that these loaded No. 8's would be of importance. We already have wires extending as far west as Denver and with the best data which we have before us now, we are warranted in the strong expectation that by loading No. 8's we could give a talk from New York to Denver which would at least be useful for advertising purposes, if indeed it would not have substantial direct commercial value. But leaving out this long talk to Denver, we have before us the problem of improving service to Chicago to about the same extent as the Boston-New York service has been improved. Over loaded No. 8's between New York and Chicago we should be able to give as good a talk as is now obtained between New York and Pittsburgh on unloaded No. 8's. This would be a very great uplift and would have a most satisfactory effect not only upon the long distance service between New York and Chicago, but upon service conditions in those two cities. What is true about New York and Chicago is also true between New York, Minneapolis, Omaha, St. Louis, Kansas City, New Orleans, Atlanta on the west and south, and Montreal and distant New England points on the north and east. Even Pittsburgh and Buffalo, while having relatively good transmission, would have the conditions improved to a sensational extent.

We have good reason to promise you that we will be able to accomplish these results in a reasonable time after the department is strengthened so that in the handling of the more comprehensive work now being undertaken, we will not be obliged to neglect matters such as these.

One motive which makes powerfully for the full development of the capabilities of the Pupin invention should be stated. Its importance is so great that it is worthy of your most careful consideration. It is this: the Pupin patent has already run eight years of its life. Nine years remain. The sooner we perfect the loading of No. 8 lines and possibly lines of even larger gauge, the longer will be the period during which we will have a monopoly of circuits constructed in this manner. If we accomplish this result within two years (I hope we will do it in one), we would have seven years during which we would be sure to exert a dominating influence in the long distance field and during which, by means of this influence and the

incalculable prestige which we would thereby obtain, we could make successful headway against competing companies and entrench ourselves against the time when the Pupin patent will have expired.

THE PROBLEM OF THE TELEPHONE REPEATER:

At the present time we have in service in the long distance lines a number of telephone repeaters. When these instruments are working in the manner intended, they accomplish a substantial improvement in extending the range of telephone transmission. When working satisfactorily, and interposed at the half-way point in the New York-Chicago line, they cause an improvement in the transmission on that line by making it talk as well as though it were 300 miles shorter. They are not uniform in their action, but the chances of our making them so are so good that a strong effort in this direction is justified. This lack of uniformity of action is not the only difficulty with these repeaters. For reasons which need not be discussed here, they are not operative upon loaded lines. This constitutes a serious defect in the repeater situation, not only with respect to loaded overhead lines, but also with respect to loaded underground lines. Naturally it is difficult to forecast the saving in our future construction which would be accomplished by the use of a repeater having uniform action, but otherwise no more efficient than the present one. Some idea of this saving, however, may be obtained from results of a study which we have made with respect to plant which we have already constructed. This study shows that if such a repeater were available when the present loaded circuits were constructed, the first cost of these circuits would have been reduced by \$7,000,000. The annual charges on this figure are \$700,000. This, it should be borne in mind, does not count upon a repeater having greater power than the present one, nor does it count upon the saving which has been accomplished by the use of the repeater in non-loaded circuits. So important do we regard this repeater matter that we are satisfied that we should attempt to develop one having much greater power. There is nothing in the nature of the case to discourage us in this line of work and the art seems to have so many possibilities and the results to be obtained from a more powerful repeater are so far-reaching that work upon this line should be pushed vigorously. If we successfully load the Denver line and thereby accomplish speech between New York and Denver, the development of a successful repeater would enable us to accomplish speech between San Francisco and New York. The achievement of this result would mean universal telephony throughout the United States and its importance is so apparent that no argument is needed to demonstrate it.

I do not think it can be said that we are looking too far ahead in talking

of a New York to San Francisco circuit, for our Operating Department is already studying the subject of connecting the telegraph system of the Pacific Telephone and Telegraph Company with our own, with a view of making a very important leased telegraph line contract. The officials of the American Telephone and Telegraph Company and the Pacific Telephone Company are actually at this time in consultation about this matter.

In addition to this, I have been advised unofficially that a most important customer of ours, a large brokerage firm of Boston, Messrs. Hornblower and Weeks, who pay to us and our associate companies as much as \$65,000 a year for telephone and telegraph service, have written us asking whether within the next two years we will be able to furnish to them in addition to the extensive telegraph service which they now have, extending from New York to Boston, Chicago and Duluth and many other places, a much more extended telegraph service. The new places which they wish to reach are Butte, Montana; Portland, Oregon; Seattle, Washington; Tacoma, Washington; San Francisco, California; Los Angeles, California; and Kansas City, Missouri. While this proposed contract and this inquiry relate to telegraph circuits, it should be borne in mind that our telegraph service is almost uniformly conducted over telephone lines, over which speech is being transmitted at the same time telegraph messages are sent, so that the construction of lines to these most distant points for telegraph purposes would be rendered more economical if they could be also used for transmitting speech.

It looks to us now as though the possibilities of loaded No. 8's for transmitting speech would be exhausted when we reach Denver and that to extend our service beyond, we must have the repeater. As it now appears, we think we may soon know how to accomplish speech without the aid of the repeater as far as Denver, and having done this, all that can stand in the way of a New York-San Francisco talk and of a talk to all parts of the United States is the application of the improved repeater to loaded circuits.

One additional argument making for vigorous work upon the development of a more powerful repeater I call to your particular attention. At the present time scientists in Germany, France and Italy and a number of able experimenters in America are at work upon the problem of wireless telephony. While this branch of the art seems at present to be rather remote in its prospects of success, a most powerful impetus would be given to it if a suitable telephone repeater were available. Whoever can supply and control the necessary telephone repeater will exert a dominating influence in the art of wireless telephony when it is developed. The lack of such a repeater for the art of wireless telephony and the number of able people at work upon that art create a situation which may result in some of these outsiders developing a telephone repeater before we have obtained one ourselves, unless we adopt vigorous measures from now on. A successful

telephone repeater, therefore, would not only react most favorably upon our service where wires are used, but might put us in a position of control with respect to the art of wireless telephony should it turn out to be a factor of importance.

* * * *

(Signed) J. J. CARTY,

Chief Engineer.

Some Aspects of Powder Metallurgy

By EARLE E. SCHUMACHER and ALEXANDER G. SOUDEN

INTRODUCTION

THIS correlated review is an attempt to present some of the more common aspects of the powder metallurgy process in order to acquaint telephone engineers with an increasingly important production method, and to provide an outline of topic references that could otherwise be obtained only from many different sources.

Basically, the art of powder metallurgy deals with the preparation of metal powders and their utilization. This is a general description, however, and covers not only the metallurgical field, but also the paint and pigment and other more strictly chemical industries. As a more pertinent definition, the following has been suggested: "Powder metallurgy is the art of producing metal powders and shaped objects from individual, mixed, or alloyed metal powders, with or without the inclusion of non-metallic constituents, by pressing or forming objects which are simultaneously or subsequently heated to produce a coalesced, sintered, alloyed, brazed, or welded mass, characterized by the absence of fusion, or the fusion of a minor component only"¹.

In the past few years, powder metallurgy has received considerable attention, not only in technical publications, but also in the newspapers and popular periodicals, the general implication of the latter being that a completely new and revolutionary field of metallurgical endeavor has been uncovered. Actually, however, instead of something new, we are dealing with an art that had its inception at the time man first started using metals; numerous examples exist today of the early attempts to produce solid articles from metal powders. It is not surprising that early investigators and workers dealt with powders rather than massive structures of metals. With the exception of a few low melting metals such as tin and lead, most of the metals available melted at temperatures above those which could be attained at the time with crude furnace equipment. It was possible, however, to prepare powders of many metals by rather simple means without extensive furnace equipment, and a number of such powders were produced. Iron, for example, was reduced from its ores and worked to solid form at least 5,000 years ago, long before furnaces were devised which could even approach the melting point of the metal. The resulting reduced product was not, of course, massive iron, but was a sponge powder material which

could be compacted, heated, sintered, and forged in much the same manner that metal powders are treated today. An outstanding example of the massive pieces produced by such methods is the 6½ ton Delhi pillar made about 1,600 years ago².

HISTORY OF DEVELOPMENT

The ancient Egyptians and probably other early civilizations discovered how to make powders of gold, silver, copper, bronze, iron, lead, and to a limited extent, tin, antimony, and platinum³, but it was necessity rather than desire which led these early workers to produce their massive metal tools, ornaments, and weapons by powder methods. It is interesting to note that as furnaces were devised to obtain higher temperatures, the list of metals prepared from powders decreased. The lower melting metals, of course, were the first to be prepared by melting and casting methods, and as higher temperatures were attained, only the more highly refractory metals remained on the powder preparation list.

Although iron had been known in prehistoric days, it remained a scarce, precious metal for several thousand years, and did not come into general use until introduced by the Hittites around 1300 B.C. The Hittites presumably mined iron ore in the iron region along the Black Sea in Asia Minor and worked the material to metal form⁴. By 100 B.C., the use of iron had spread westward to include many of the countries bordering the Aegean and the Mediterranean. The primitive methods of iron working probably consisted in heating the iron ore in a charcoal fire fanned by an air blast from a bellows until reduction of the oxide was attained. The spongy mass was then pressed, heated, and forged to the desired shape.

That this was the general practice followed in many countries in the production of metal objects has been observed from articles unearthed from earlier civilizations. Somewhat similar methods of working other metals have been observed, and where difficulty was experienced in obtaining sintering, other metal powders were added that were lower melting themselves, or that formed lower melting alloys which wet and welded together the particles of metal being worked to form a lump that could be shaped. The Incas in South America used such a method in fabricating many small articles of platinum⁵. The grains of native platinum were mixed with some gold and silver, and, by means of a blow-pipe, were fritted together by the lower melting alloy of gold and silver. The resulting mass could then be forged to the desired shape.

During the eighteenth century there was a fair amount of activity in the production of metal powders, and in studies of the fabrication of metal parts from the powders. Platinum was introduced into England in 1741 and attempts were made to produce the metal in compact usable form³.

Various expedients were used, and one which utilized a unique principle is worthy of note.

It was observed about the middle of the century that platinum would fuse at relatively low temperatures in the presence of arsenic³, and that, on prolonged heating, the arsenic could be volatilized out of the fused lump to leave behind a sponge of metallic platinum. This sponge could then be heated and forged to solid form. Similar results were obtained using mercury* or sulphur in place of arsenic; and the success of the forging methods led other investigators to study the welding of grains of native platinum or platinum scraps without the use of added elements to lower the fusion point.

Such was the situation in the early part of the nineteenth century when Wollaston⁷ developed his method for the preparation of platinum ware. Numerous other investigators^{3,6,8} had produced articles of platinum by treatment of finely divided platinum or sponge, but by careful refinements in the process with control of particle size, purity, compacting pressure and sintering treatment, Wollaston obtained a superior product. Precautions were taken to use only the more finely divided platinum particles, and to press the powder carefully in a mold while wet. This pressing of wet powder is claimed to have been one of the main contributions made by Wollaston since a much lower compacting pressure was allowable, and the particles were not work hardened. The resulting cake was then slowly dried to remove volatile matter and adsorbed gases before sintering at 800°–1000° C. The material was forged while still hot, and gave the first really pure, blister-free platinum sheet. That the process developed by Wollaston was sound is shown by the fact that the platinum produced by powder metallurgy at present in England is made by essentially the same procedure²¹. The careful studies made by Wollaston in fabricating platinum ware of high purity thus led to the basic principles utilized in successfully producing massive metal parts from metal powder.

During the nineteenth century, many metals were produced in powder form, but there seems to have been no correlated effort to convert the powders into coherent form. This may have been due to the development of better melting furnace equipment that allowed ordinary melting and casting techniques to be employed for most metals and alloys. On the other hand, there remained some of the more refractory metals such as tungsten, tanta-

* As an example of how new methods introduced can often be traced back to earlier sources, the use of mercury to form an amalgam which could then be heated to leave a powder sponge material, has been attributed to the monk Theophilus in the 11th century⁶. In this case, the amalgam process was used with gold, and the end product sought was gold powder which could be used as a pigment in inks for illuminating manuscripts. There was no attempt, however, to carry the process further to make solid metal parts as was the case with platinum as cited above.

lum, molybdenum, osmium, and iridium which could have been treated in much the same manner as in Wollaston's process for platinum.

There were, however, instances where real effort was made to develop useful products by means of powder metallurgy. As early as 1870, the fundamental idea of a self-lubricating bearing was disclosed in a patent by Gwynn⁹ and was the prototype for a large number of later developments in the field. To 99 parts of tin prepared by rasping or filing, one part of petroleum still residue was added, and the mass heated and intimately mixed. The mixture was then pressed to give the shape and solidity desired. It was specifically stated by Gwynn that journal boxes made by this method or lined with the material would allow shafts to run at high speed without other lubrication¹⁰.

There were a number of metal powder producers in the nineteenth century, most of them producing flake powders, but a virtual monopoly in the field was held by Sir Henry Bessemer from about 1840 to 1885, when he retired from the business¹¹. The process was a secret one and remained so for almost his entire business career, and the profits were so large that they financed the development of the Bessemer process for making steel. Essentially, the method was one of machining very fine filaments from solid metal bars and passing the filaments through rolls to flatten and break them into flat tabular particles. Precautions were taken to prevent sticking and give a high polish to the powder by adding a very small amount of olive oil. The powder was graded by means of an air blast in a tunnel about 40 feet long and 2½ feet wide, the finest powder fraction being collected in silk bags attached to the end of the tunnel. Bessemer's powder metals included copper, and most of the common alloys of copper.

Even with the relatively large scale production of flake metal powders by Bessemer up to 1885, and the subsequent preparation of powder metals by stamp mills which pulverized the metal by severe working, there was very little actual commercial manufacture of solid compacts from powder metals.

The electric lamp industry provided the stimulus for further study in the search for a metallic filament to replace the carbon filament first used. This culminated in the production of the tungsten filament^{10,12} and indicated the technique to be applied in the development of the other refractory metals as well as the production of cemented carbides, electrical contacts, and electrode materials.

Even with the promise shown by this development and the production of other ductile heavy metals, there was little other commercial activity in powder metallurgy as late as 1915-1920.

Various types of porous bearings had received sporadic attention, and, in 1921, a new porous bronze bearing was described¹³. The material was

a bronze having finely divided graphite uniformly distributed throughout the mass. It was prepared by mixing the oxides of tin and copper with graphite, compressing the mass and heating. There was reduction of the oxides by the graphite and partial diffusion of the copper and tin to give a porous bronze structure in which excess graphite was uniformly distributed in amounts as high as 40 per cent by volume. In addition, there was sufficient porosity for the introduction of 2 to 3 per cent of oil. Later developments utilized the metal powders rather than the oxides¹⁴, and porous bearings in a variety of compositions and forms have constituted a large part of the total production of powder metallurgy products over the years. Of considerable influence on the design and utilization of this type of bearing has been the demand by the automotive industry for large quantities of small bearing parts. Many of these parts are at inaccessible places, and the value of a self-lubricating surface is apparent. As suggested previously, these bearings are not all of the simple pressed porous alloy structure described, but many are complicated such as those having a steel backing coated with a porous sponge alloy of copper-nickel in which the voids are impregnated with Babbitt metal¹⁵.

A later development, and one which has had tremendous industrial significance, was the production of cemented carbides^{17,18,19} and their use in cutting tools, dies, and hard surfaced parts of many types. Essentially these consist of finely divided tungsten carbide particles bonded by cobalt, or in some few instances, nickel or iron. Other carbides such as those of tantalum, titanium, or columbium may be added to impart special properties.

Powder metallurgy is admittedly an art that has progressed more rapidly than the science, but the gap is being closed by investigations of a fundamental nature. Much of the lack of correlated information in the field has been due, in part, to an understandable reluctance of the manufacturers to divulge information on their processes to competitors, and largely, as well, to the narrow specialized uses that apparently discouraged a general systematic investigation of the problems involved. Within the past ten or fifteen years, mainly through the efforts of producers of metal powders, research of a fundamental nature has been stimulated. Another factor has been the large scale adoption of the powder metallurgy process by the automobile industry for use in the preparation of many different parts. The field is still narrow and specialized, but the art has progressed to the point where powder metal parts are competing, in some instances, with parts made by the standard melting, casting, and machining procedures.

As in many similar situations where rapid expansion has occurred, there has been a tendency, not as yet based on actual performance, to oversell the product. This is a sign of healthy activity on the part of the exploiters

in the field, but a somewhat unwise course for industry as a whole to pursue. That there are limitations to powder metallurgy and many serious problems unsolved, is generally now recognized, and there is a tendency toward more conservative evaluation of the potentialities of the process.

It is the purpose of the remainder of this article to describe some of the common methods of preparing metal powders, to explain the fundamental principles involved in powder metallurgy, to describe the advantages and limitations of the process, and to indicate the type of product that may be expected.

MANUFACTURE OF METAL POWDERS

Metal powders are made in a variety of ways, each method of preparation being suited to the metal being treated or to the end product desired. Experience has shown that no one type of metal powder can serve all the projected uses in industry, so it is not surprising that there have been developed numerous methods for the preparation of metal powders, each of which has advantages for certain types of work, and which may or may not be suited for other uses^{11,16}. Listed below are some of the common methods which have been developed for producing metals and alloys in powder form. No attempt is made here to discuss these methods in detail or to point out the relative hazards²⁰ involved in the various processes. It is worthy of note, however, that many metal powders in a finely divided state have such a large surface area in proportion to their bulk that they are usually subject to rapid oxidation, so rapid in many instances that they constitute an explosion hazard. Care must therefore be exercised throughout in the preparation of these powders, and many must be prepared and stored in inert atmospheres.

1. *Machining*

Machining of metals to produce powder has been mentioned above in connection with the process of Bessemer. A relatively coarse powder is produced. The cost of production is usually high, and the powder use is limited to a few special applications such as dental alloys where no fines or dust are allowable, and where the high cost of the alloy itself justifies the extra cost of this method.

2. *Milling*

By the use of various types of mills such as stamp mills, jaw crushers, gyratory crushers, impact, and ball mills, both brittle and malleable metals can be reduced to powder. The friable metals tend to produce angular, jagged, particles of irregular shape while the malleable metals usually produce flakes. Because of the lubricant necessary with malleable metals

to prevent the flakes from welding together, this type of powder is not greatly used for molding metal parts. The grease or other lubricant interferes with proper sintering, and there is an additional disadvantage of flakes in that low strength laminated or layered structures result in the pressing operation. The flake powders are more generally used as pigments in the paint industry where their flat surface is an asset for good coverage.

A special type of mill, the Eddy Mill, can be used for malleable metals to give particles of suitable shape, fineness, and purity for the manufacture of sintered briquettes. Essentially, the mill consists of a chamber wherein are mounted two fans facing one another and operating at high speeds in opposite directions. The metal is introduced into the chamber in relatively small pieces, (e.g. $\frac{1}{4}$ - $\frac{1}{2}$ inch lengths of 0.05 inch diameter wire) which, by collision with one another in the fan blasts, become very finely pulverized. The process can be accurately controlled and a variety of shapes, angular, flake, or pebble, can be produced as desired.

3. *Shotting*

Metal shot can be prepared by dropping the molten metal from a small opening through air or an inert atmosphere into water. If the method is controlled properly, a fairly fine shot can be produced. On the whole, however, this process in powder metallurgical work is confined largely to preparing intermediate size particles for further reduction by other methods.

4. *Atomization*

For metals having relatively low melting points, atomization provides a convenient method of producing fine particles. The molten metal is forced through a small nozzle orifice and broken up by a stream of compressed air, steam, or inert gas. The process can be controlled rather closely by proper choice of nozzle, pressure and temperature of the gas used, and the rate of metal flow. As a rule, it is applied to metals melting below 700° C. such as lead, lead alloys, zinc, and aluminum; but copper, having a much higher melting point, has also been successfully treated in this manner. The product can be drawn off and collected in standard dust collector systems, and is suitable for many types of powder compacting.

5. *Carbonyl Process*

Both nickel and iron under suitable temperature and pressure conditions will react with carbon monoxide to form the respective carbonyls²². From these carbonyls, the metals can be obtained by a reverse of the process, decomposing the compound to the metal and the monoxide. The virtue of the process lies in the shape of particle, which appears to be almost

spherical, the purity, and the control which can be exercised in particle size. . The method has been used for years in the Mond process for making nickel shot, but, until recently, foreign producers exercised almost a complete monopoly on the manufacture of fine powders from carbonyl. Within the last few years, iron carbonyl powder has been produced on a large scale in this country in several different grades suited to industrial needs. The iron powder is a specialty product commanding a higher price than that produced by most other methods, but because of superior properties it has been used extensively in the electrical industry, particularly in the communications field for various types of magnetic cores.

6. *Condensation of Vapor*

Metals which have low boiling points can be vaporized and the vapor then condensed in powder form. These include zinc, magnesium, and cadmium. The powders so produced are used mainly in the chemical industry.

7. *Reduction of Chemical Compounds*

Metal powders whose characteristics can be varied over a wide range are prepared in large quantities by reduction of compounds of the metal with hydrogen or other reducing gases at temperatures below the melting point. The oxide of the metal is most generally utilized for the purpose, and among the metals produced are copper, nickel, iron, cobalt, molybdenum, and tungsten. The type and shape of the metal powder is governed somewhat by the compound from which it is reduced, so that, within limits, these factors are controllable by proper choice of compound.

8. *Electrolytic Deposition*

Metals can be electrodeposited in several ways to obtain powder depending upon the plating conditions. A hard, brittle deposit may be obtained which can be further crushed or ground to small particles, or a soft sponge, or even the metal in powder form can be produced. The powder is usually dendritic in shape and requires further treatment for use in molding. This generally comprises some sort of milling or grinding operation, and an annealing treatment to eliminate hydrogen and soften the powder.

9. *Other Methods*

Other methods for the preparation of metal powders include chemical precipitation, granulation, alloy formation and removal of an alloying constituent (such as platinum-arsenic, platinum-mercury, and gold-sulphur previously discussed), and the hydride process⁶⁷. The last mentioned method is probably the only one of these which is of more than academic interest for powder metallurgy uses.

Hydrides can be formed of many metals, those of titanium, zirconium, thorium, hafnium, columbium, and tantalum being of particular interest since they are reported to be stable at room temperature. They are produced in 300 mesh size or finer, have the appearance of metal, and begin to dissociate into hydrogen and the pure metal in vacuum or non-oxidizing atmospheres above 350° C. The hydrides can be mixed with other metal powders, and, when compacted and sintered, slowly release hydrogen which creates a protective atmosphere around the metal particles and sometimes acts to remove oxide films already present.

Despite the number of methods known for producing metal powders, the bulk of the powders used on a large scale are produced by only three methods²³: electrolytic deposition, atomization, and reduction of metal salts by gases. The carbonyl process produces a specialty product as does the hydride process, and, while both have their uses, the amount consumed is probably small in relation to that prepared by the other methods.

THE POWDER METALLURGY PROCESS

As has been indicated in the introduction, there are a number of definite steps in the powder metallurgy process which may be summarized as follows:

1. Selection of the powder or powders best suited for production of the part under consideration.
2. Proper mixing. (If more than one type of powder is being used)
3. Pressing. (Sometimes followed by pre-sintering)
4. Sintering. (Sometimes followed by an impregnating operation)
5. Coining or Sizing operation if necessary.

Each of these important operations is discussed in somewhat more detail below:

1. *Selection of Powder*

When the actual metal or alloy composition has been decided upon, there are a number of factors which must be considered in the selection of the type of powder itself. An essential characteristic is purity²³ because in the powder metallurgy process impurities cannot be slagged off as in most melting processes, and may interfere with pressing and sintering operations. Oxide films, for example, may prevent good contact between metal particles. Clean surfaces are essential if ductility, and high tensile and shear strength are required in the finished article. In most cases, there is a definite limit set for objectionable impurities in a given powder, but in some instances materials normally classed as impurities are deliberately added to obtain a desired result. An example is the addition of thorium

dioxide to tungsten as later described in the section on types of metal powder products.

The physical properties of the metal powders are also determining factors in their selection. These include particle shape, size, hardness, particle size distribution, flow characteristics, apparent density of loose powder, and particle grain structure.

Particle shape and size are governed largely by the method of production of the powder as has been suggested previously. The carbonyl process yields spherical particles, for example, while other methods produce particles that are angular, acicular, spongy, flat, rounded, granular, dendritic or otherwise irregular.

The hardness depends largely upon the metal itself, its purity, and the method of preparation. Hardness, in addition to shape of the particle, will be reflected in the amount of pressure required to obtain a given density in a finished part, and is a factor in the economics of die cost because of its influence on die life.

Particle size distribution in a metal powder is of great importance although no particular specification can be set up at present. The problem of size distribution and shape has been treated in some detail by W. D. Jones²⁴ and others, especially as concerned with interstitial volume or porosity. If all particles were cubes of the same size and could be placed in perfect order with the cube faces matching identically, there would be a minimum of porosity in the powder and in the pressed part. This is obviously impossible of attainment. In practice, packing is not systematic, but random, and even if identically sized cubes could be obtained, the voids between particles would be appreciable. In addition to the porosity resulting from the random packing, there are cavities which are due to bridging action of the particles themselves. This bridging is not due to irregular or angular particle shape, but can occur quite easily with spherical particles. Shaking or compressing the powder tends to destroy the bridges or arches and allow denser packing. As the powder is shaken down there is rotation of particles until corresponding surfaces come in contact and relatively dense packing is obtained. Such a rotation may not be present, however, during the rapid stroke in a die, and the particles cannot seek corresponding surfaces. In this case, there is a deformation of the particles pressed against one another so that there may be an actual keying, and the smaller particles may be pressed into the voids to produce the same result of denser packing. With a distribution of particle size, the voids between larger particles can be filled with smaller particles and, in practice, that is what is sought. The problem of setting up specified sizes or particle size distribution for powder metallurgy methods is not easy, however, because of practical complications arising in the pressing and sintering operations. Pore size

rather than total porosity then becomes the problem, since, in sintering, only the smaller pores may become closed. At present, the manufacturer of metal powders cannot guarantee his particle size distribution, nor can the user determine and specify exactly what he needs. The grades can be approximated, only, and the types required must be determined in an empirical manner²³.

The apparent density (or loading weight) is the ratio of weight in grams to volume in cubic centimeters of powder, measured according to some specified method of filling a designated receptacle. It is of considerable practical importance since it has effect on several of the operations of powder metallurgy, especially that of pressing the compact. The lower the apparent density of a powder as compared with the actual density of the solid metal, the greater will be the volume of powder required to produce a briquette of given size. This necessitates deeper dies and longer plungers than for denser materials, and for very low apparent densities may become a serious design problem. Powders can usually be supplied in a range of densities, and the proper powder selected for use. For proper blending and mixing of different metal powders for producing solid metal parts, it is advisable to select grades having comparable apparent densities. An example of the use of a low-density copper powder may be cited. For the manufacture of starting brushes in the electrical industry, copper powder and carbon powder are mixed together and compressed. By using copper powder of a low apparent density, approaching that of the carbon (1.2), good blending is assured and the danger of segregation eliminated²³.

Low rate of flow of metal powders interferes with automatic pressing operations and may make it necessary to install vibrating equipment on the feeder hopper or even on the die itself. Rate of flow is influenced by particle size distribution, particle shape, and amount of absorbed moisture.

2. *Mixing*

When only one metal is to be pressed and sintered, there is usually no necessity for mixing since the powder as received from the manufacturer is generally well blended. Where several batches of the same metal of different particle size distribution are to be added, or where different metal powders are to be used, it is necessary to mix them thoroughly prior to pressing and sintering. This may be done in any of the standard type mixers with the precaution, in some instances, of providing against oxidation of the powders.

3. *Pressing*

For preparation of the compacts, the pressing operation may be done at either ordinary or elevated temperatures. The majority of parts pro-

duced, however, are pressed at room temperature. The presses^{25,26} which now are designed primarily for this type of work may be of the mechanical or hydraulic types for high production rates with modifications for rapid plunger strokes as required.

The dies are generally of hardened steel having the inner surfaces highly polished by lapping with polishing rouge in the direction of the plunger stroke so that any fine scratches that remain are in the direction of ejection of the pressed part²⁶. In some instances where parts are made from highly abrasive particles, the dies are made of or lined with hard carbide materials. Die depth depends upon the apparent density of the powder being pressed, but the usual ratio of depth to part thickness is approximately 3 to 1. The greater die depth required for powders of lower density introduces the complications of friction at the die sides, unevenness of pressure distribution, and internal friction of the powder itself. There is almost no lateral flow in the powder mass, a condition which limits the shapes that can be pressed.

Pressure used varies from 5 to over 100 tons per square inch, in general, and is an important factor in limiting the size of parts that can be made by the powder process.

Following pressing, a powder compact may sometimes be given a pre-sintering treatment below the normal sintering temperature in order to increase its strength to facilitate handling, or to remove lubricants or binders which might cause difficulties later.

4. Sintering^{24,27,28}

Sintering is the fundamental process in powder metallurgy whereby solid bodies are bonded by atomic forces.

Theoretically it is possible to obtain bonding by bringing the powder particles into so close contact with one another that the atomic forces of cohesion may become operative. But this occurs only when the respective atoms of such adjacent particles are distant in the order of magnitude of the crystal interatomic spacings; this is a condition against which there are many obstructions. Visually and even microscopically smooth particles have surfaces which are extremely jagged with respect to interatomic spacings and crystal planes. Then not large, flat areas representing large numbers of atoms, but only successive points representing relatively very small groups of atoms can be brought into sufficiently intimate contact. Moreover, even this small contact may be reduced by the presence of oxide films.

An increase of pressure will improve the bonding of such powders since the particles are deformed and pressed against one another to give increased surface contact. At the same time, rupture of the oxide films may occur with subsequent closer contact of the metal particles. This is the

general case for pressed powder compacts, or "green compacts" as they are designated. There is frequently a surprising strength associated with such pressed parts but, on the whole, a heat treatment is necessary to produce a material approaching the strength and solidity of a cast or wrought metal part.

The heating of pressed powder briquettes is usually done in an inert, reducing, or neutral atmosphere, or in vacuum. The temperature used is determined by the metal powders comprising the compact, and by the properties desired in the final product. The melting point is not exceeded for any of the components of the mixture except in those instances where such fusion of a minor constituent is desired, as, for example, in the production of cemented carbides. No definite temperature may be set for the heat treatment, but general practice is to treat at a temperature about two-thirds that of the melting point of the metal or alloy being fabricated. Higher temperatures are frequently used, however, and may be only slightly below the melting point.

The effect of heat is possibly that of causing increased surface diffusion and plasticity. The atoms on the surface of metal particles possess considerable mobility far below the melting point, and the surface energy at elevated temperatures may be appreciable. Where particles are in contact surrounding a void, flow of metal is in such a direction as to increase the area of contact.

When the sintering temperature is within the recrystallization range of the metal or metal alloy powder being treated, marked structural changes may occur. Recrystallization takes place at sites of plastic strain. Since these sites are regions of contact between particles, new crystallites form and grow into the adjacent particles so that a new series of grain boundaries is formed. The numerous cavities or voids present in the structure are not completely filled in or sealed in this operation. This could not occur without change of overall dimensions of the compressed mass. The voids may be present at the new boundaries or even enclosed in the crystallites, and produce a non-homogeneous sintered metal of relatively weak structure susceptible to sudden shock. By a high temperature treatment just below the melting point, or by alternate working and annealing, the voids can be closed and the metal consolidated to a dense, strong mass.

Surface oxide films which interfere with the sintering operation may sometimes be destroyed by treatment of the powder compact in a reducing atmosphere. If the oxide cannot be reduced in this manner, the pure metal can only be obtained by sintering operations if the oxide has a higher vapor pressure than the metal²⁹.

Gases, either adsorbed, dissolved, entrapped, chemically bound, or

resulting from chemical action, may interfere with sintering and the general rule is to avoid them if possible in attempting to produce solid metal.

Following sintering, there is sometimes a treatment for impregnating a porous structure with some material designed to confer special properties on the compact. Pressed and sintered bearings may, for example, be impregnated with oil, and a strong, porous network of tungsten may be impregnated with copper by suitable means to produce spot and line welding electrode material having high compressive strength associated with good heat and electrical conductivity.

5. *Coining or Sizing*

Although the dimensional tolerances of sintered metal parts can be rather closely controlled, it may be advantageous to control final size and improve surface structure by a coining operation consisting in re-pressing the compact in a die of suitable size.

THE MODERN FIELD OF POWDER METALLURGY

Most of the developments and uses of metal powders described thus far, it should be noted, have been concerned with products which could not be produced in any other way than by powder metallurgy processes. This, in fact, has been the principal field of powder metallurgy. Porous bearings with uniformly distributed porosity could not possibly be fabricated by any of the standard melting and casting techniques, nor could the carbide cutting tools be likewise manufactured.

In general, the powder metallurgy process has been applied under conditions as outlined below^{30,31,32}:

1. Production of refractory metals such as tungsten, tantalum, columbium, and molybdenum.
2. Development of structures not practical by other methods. These include telephone and radio cores, and articles requiring uniform or controlled porosity such as porous bearings and metallic filters.
3. Preparation of metals to include uniformly distributed non-metals.
4. Preparation of samples comprising a metal with another metal or metals which would be immiscible in the molten state, or which do not form alloys.
5. Preparation of samples of two or more metals where one component has a low boiling point.
6. Fabrication of products that can be made more economically by the powder process than by other methods¹⁴.

Considerable work has been done by the automotive industry and others

in developing products from powder metals that fall into class 6 above. There are many instances where automatic pressing and continuous annealing operations on small parts in quantity have made the process economically feasible for competition with the standard casting method. There are many factors involved in determining whether parts should be thus fabricated, and these will be described at greater length in the section on limitations of the powder method.

With the advent of increased production for war purposes, the powder process has, in many instances, been utilized to insure a steady supply of many small parts needed for ordnance. The use of powder metallurgy has released machines and mechanics for other types of work, and because of the speed and ease of setting up for production, it has often been possible for suppliers of small parts to adhere to schedules they could not otherwise meet³³. In addition, because of the low metal loss connected with the powder process, there is considerable saving of scarce or strategic material.

To the six general classes of materials listed above, can then be added another class that can best be described as utilitarian. The powder method has been used as an expedient to supplement and extend normal production methods without regard to cost. However, it has often proved itself to be economically competitive, and in many cases, has effected considerable savings over normal production methods³³.

The intensified war production schedules have opened the larger field that has been long predicted by powder metallurgists, that of using the powder method to displace the conventional methods of making many parts not in the classification of specialty products. Even under the abnormal war conditions, however, there are indications that progress along these lines will not be rapid and the early promise shown has not been completely realized. Progress has been made, nevertheless, but many of the developments and products are known only to those workers actually engaged in producing parts for the wartime program, and only when the story of the progress made can be told, will complete evaluation of the process be possible.

It is the belief of some metallurgists, as yet realized commercially with only a few special items, that parts can eventually be prepared by powder methods with properties superior to those obtained by melting, casting, and working techniques. At least one investigator reasoned that, since sintered tungsten is stronger than fused tungsten, iron or steel prepared similarly should show the same superiority³⁴. Actual studies conducted using relatively high compacting pressures indicate that both iron and steel can be prepared by powder methods with tensile properties better than those obtained on the some materials made by fusion processes.

TYPICAL POWDER METALLURGY PRODUCTS

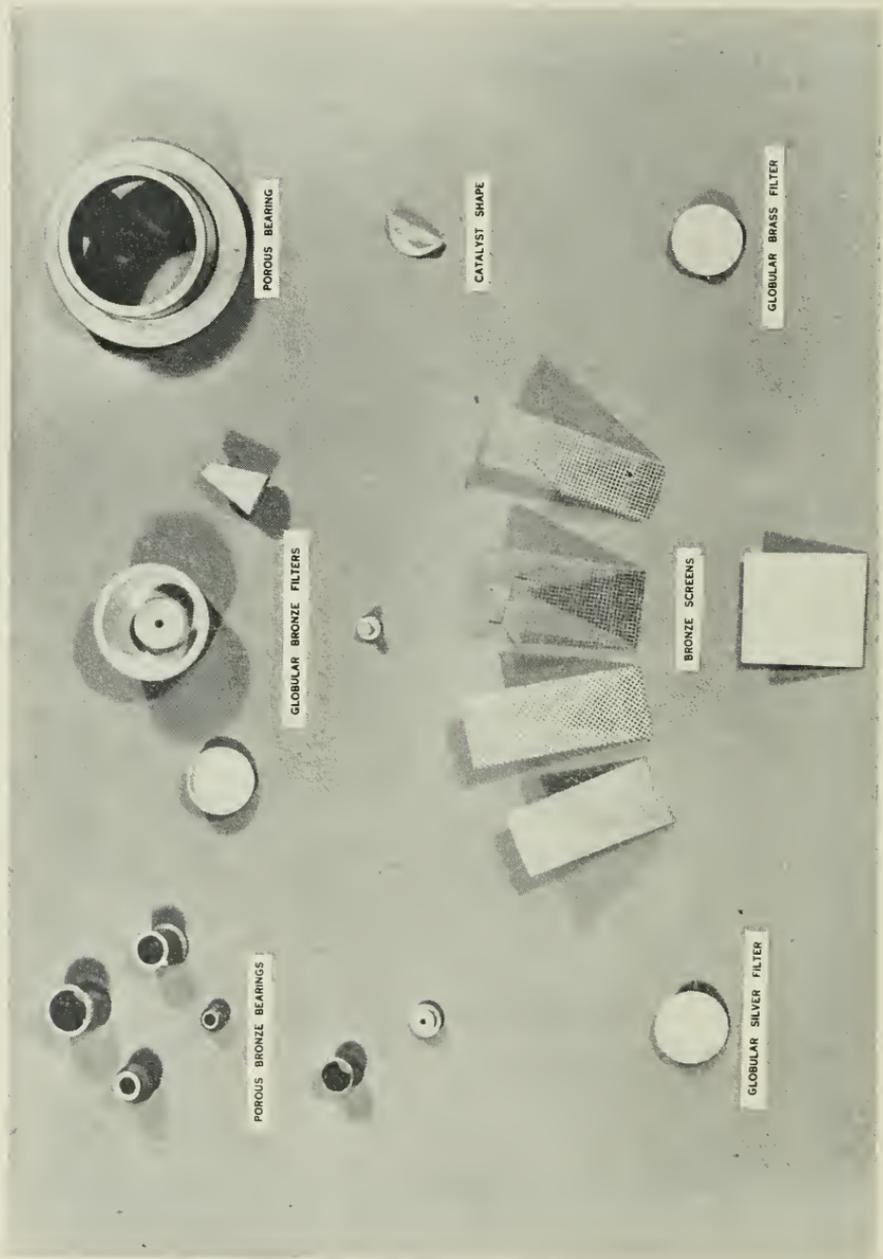
Most of the materials produced by powder metallurgy prior to about 1940 are well known; some have already been mentioned in this article, but for convenience are included in the following descriptions of typical products. Others of more recent development owe their immediate existence to the demands of wartime production, and, while some have been described in some detail in the technical literature, many have had only brief mention. Some typical parts made by powder methods are shown in Figures 1, 2, and 3.

1. *Cemented Carbides*^{17,18,19,35}

Although tungsten carbide was produced many years ago and was found to be extremely hard, it was so brittle and low in strength that its use commercially where advantage could be taken of the high degree of hardness was not possible. About 20 years ago, it was discovered that the addition of a small amount of metallic constituent, such as cobalt, to the tungsten carbide powder would yield a hard, relatively strong compact after sintering. During the heating operation, there is partial melting with some solution of the carbide by the cobalt; and on cooling the cementing material produces the required strength.

The method of preparing the powders, compacting, and sintering has undergone considerable improvement since the first carbide materials were made. Essentially, in outline, the process consists of first preparing the tungsten carbide powder, mixing it with cobalt powder and ball milling the mixture until proper grain size is obtained and the carbide particles are coated with a thin layer of the cementing metal. In this treatment, other carbides are added as required. Following the milling operation, the mixed powders are pressed in suitable molds and given a pre-sintering heat treatment to increase the strength for handling and to remove, by volatilization, lubricants which may have been used to facilitate pressing. After the pre-sintering operation, the compact can be cut to desired shapes quite readily. The sintering treatment which follows is carried out at about 1400°–1500° C. with the pressed parts placed in carbon boats or on carbon slabs and heated in a suitable neutral or reducing atmosphere. There is considerable shrinkage in dimensions in this sintering treatment which gives a product that is hard, dense, sound, and strong. Any further shaping is done by grinding or lapping operations.

The cemented carbides have many uses usually falling into the three general classes of die materials, cutting tool materials, and wear and corrosion resistant materials.

Fig. 1—Some porous parts made by powder metallurgy.²³

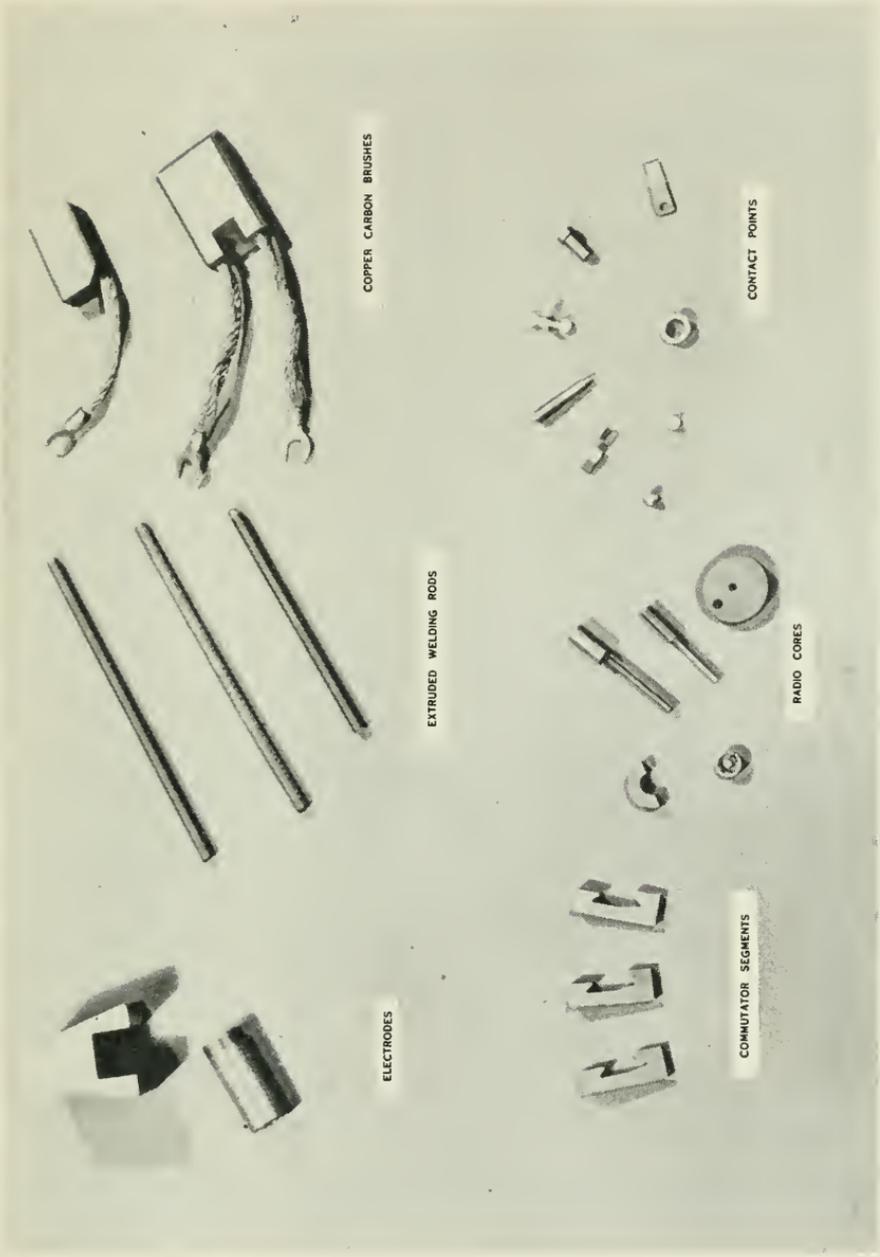


Fig. 2—Some electrical parts made by powder metallurgy.²³

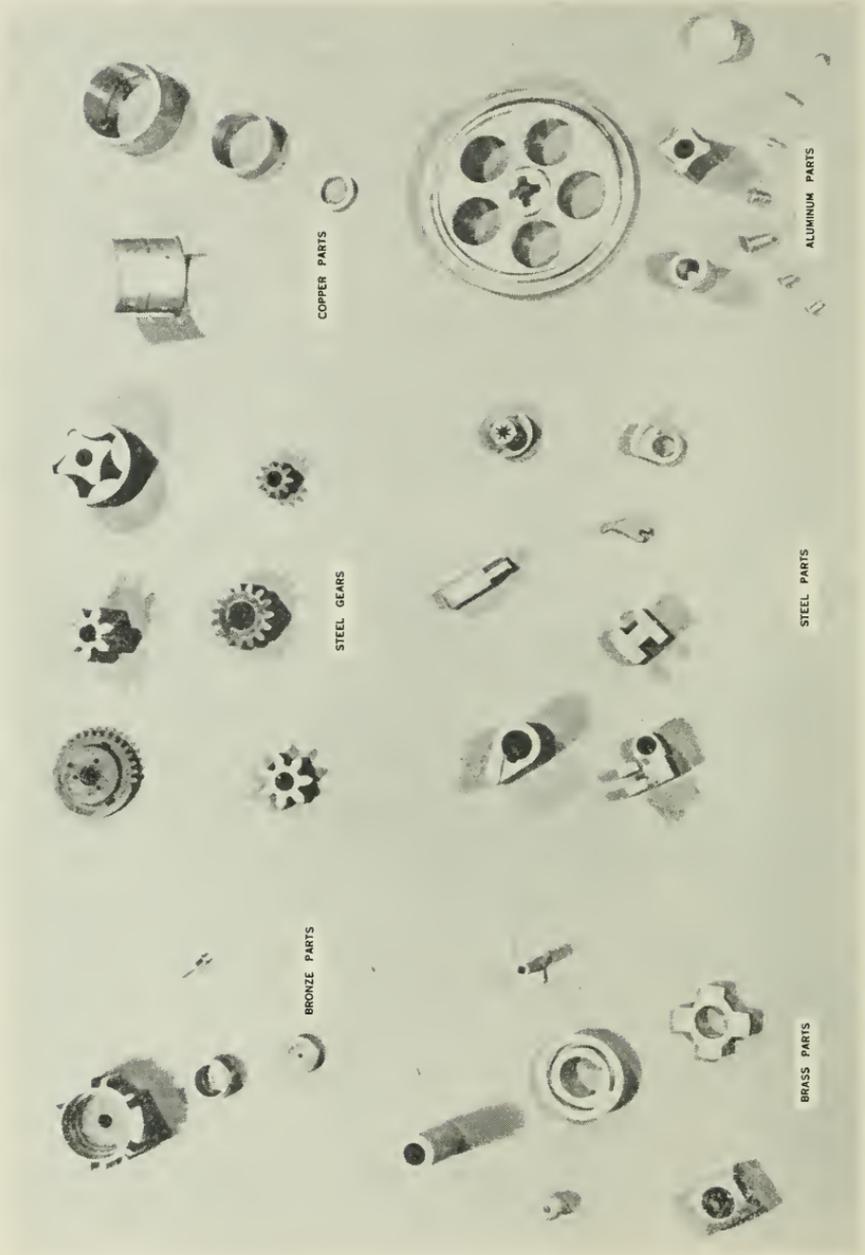


Fig. 3—Some machine parts made by powder metallurgy.²³

The die materials are usually the simpler tungsten carbide compositions and can be used to advantage for extruding, drawing, sizing, and other operations where the shape or dimensions of the article being worked is changed but where no metal is removed in the operation. The tungsten carbide can be used in this way for shaping many types of metals and alloys and this has been a major use of the product.

Cemented carbides, either the simple type or the mixtures, depending on the application, have been successfully used as cutting tips on a variety of tools, and for a number of different materials. This use has increased steadily due to the remarkable increase in production achieved. Decrease in cost of the tips and parts during recent years has further stimulated use.

Wear and corrosion resistant parts include gauges, guides of many types, pump valves for abrasive materials, sandblast nozzles, burnishing tools and dies, and many others where utilization of the superior properties is indicated.

One use recently reported³⁶ has been that of cemented tungsten carbide for bullet cores in ammunition for anti-tank weapons used by the enemy in the desert warfare in Africa. The material has about twice the density of steel and is much harder, and, although not greatly resistant to shock under normal conditions, becomes quite effective under the high pressures attained during striking and penetrating armor plate.

2. Porous Bearings

Porous bearings, always a large runner in the powder metallurgy field, have been described in the section on the historical development. Where the bearings are impregnated with oil, there is usually sufficient to last the lifetime of the assembly, but provision can readily be made for supplying additional oil if needed by utilizing the capillary action of the interconnecting pores to draw oil from a reservoir in contact with the bearing wall. In such assemblies, there is always a film of oil for the shaft to run on in contrast to normal bearings where an oil film does not coat the shaft until run for some time.

3. Motor Brushes and Commutator Segments

Numerous types of current collector brushes are now made by powder methods. Copper powder can be added to the graphite mixture, and the desired part pressed and sintered below the melting point of copper to develop a strong, high conductivity brush of longer life for use against copper surfaces. Greater wear resistance may be obtained by adding zinc, tin, or nickel to the mixture. Improvement in operating smoothness may be attained by the incorporation of lead^{14,23}.

The brushes can be pressed around pigtail conductor wire inserts to insure good contact for the lead wire and eliminate attachment problems.

Commutator segments, resistance rings, and rotor bars in squirrel-cage motors, have been successfully produced from copper by powder metallurgy methods³⁷.

4. *Refractory Metals*

Because of high melting points, the refractory metals, of which tungsten, molybdenum and tantalum are the most important, are prepared by powder methods. The preparation of each is similar, with the technique differing only in certain details where the characteristics of the individual metals require it³⁸.

With tungsten³⁹, the ore is treated by chemical methods to yield pure tungstic oxide which is then reduced by hydrogen at 650°–950° C. to give tungsten powder with a particle size range from 0.5 to 8 microns. As with other metal powders, care is exercised throughout to maintain high purity. After proper mixing and blending, the powder is compressed and the briquette given a pre-sintering treatment at 1000°–1200° C. to give sufficient strength for further handling. The resulting bar is then clamped in electrodes in a suitably designed hydrogen chamber, where acting as a resistance heater, heavy electric current is passed through it. The compact shrinks, the density increases, and a relatively solid bar results which can then be hot-worked. During the swaging or rolling, the working temperature can be gradually decreased until there is sufficient ductility by control of grain size to draw the material cold.

For tungsten used in lamp filaments, certain additions such as thorium oxide, or compounds of sodium or potassium mixed with such relatively non-volatile materials as SiO_2 , Al_2O_3 , or ThO_2 are intimately mixed with the tungstic oxide prior to reduction. These additions are effective in controlling grain growth and insuring proper grain boundary orientation for producing "non-sag" coiled filament. Essentially, the sodium and potassium compounds promote large grain growth while the others, such as thorium oxide, inhibit grain growth under the conditions of wire fabrication. When the material is drawn in wire form, the thoria particles form elongated stringers in the direction of drawing and tend to prevent grain growth across the wire while allowing exaggerated growth along the axis of the wire. The resulting structure of long grains with boundaries forming acute angles with the longitudinal axis of the wire is ideally suited for the coil type of lamp filament.

Molybdenum and tantalum are prepared in much the same manner as tungsten, although tantalum sintering and annealing must be conducted in high vacuum because of the ability of the metal to absorb and retain gases at high temperatures.

5. Heavy Alloy

"Heavy alloy" is the name applied to a group of alloys composed of tungsten, copper, and nickel having a density of 16 grams per cc. or greater^{40,41}. They were originally developed for fabricating the containers and nozzles for radium units, but have such interesting properties that a number of other uses have become evident. Specific properties depend upon the composition, but generally the tungsten comprises about 90 per cent of the alloy.

One of the best compositions claimed is that of 90 tungsten-7.5 nickel-2.5 copper which has properties as listed below:

Tensile Strength.....	90,000 psi
Yield Point.....	83,000 psi
Elongation in 1 inch.....	4%
Elastic modulus.....	32×10^6 psi
Brinell hardness.....	250-290
Density.....	16.3-17.0 gms. per cc.
Coefficient of expansion.....	5.6×10^{-6}
Thermal conductivity.....	0.25 c.g.s. units

The alloys are prepared by mixing the metal powders dry, adding a small amount of wax, in benzol solution, mixing until the solvent has evaporated, and then pressing to shape. The compact is heated slowly to about 1000° C. and then sintered at a higher temperature at which the nickel and copper particles fuse, and the tungsten is not only wet by the liquid, but actually dissolved. The fine particles are thus dissolved, but tungsten is reprecipitated on certain nuclei to develop large rounded grains. The solution and redeposition continue until the original fine tungsten particles are replaced by grains approximately 100 times the original particle diameter. The alloy thus consists of tungsten particles in a cementing phase of copper-nickel-tungsten.

There is a shrinkage of up to 20 per cent, and the resulting compact is relatively free of porosity.

The alloy has good machining properties and can be treated much like many cast alloys. It has good corrosion resistance and can take a variety of surface finishes.

In addition to its use in X-ray and radium work, its high density and strength make it attractive for use as a counterweight material in high-speed motor setups of many types.

6. Electrical contacts and electrode materials

Powder metallurgy can be utilized to fabricate material composed of two or more metals without any appreciable alloying so that the characteristics of each of the components may be retained to a large degree. This has opened a large field for electrical contacts and welding electrodes made by using compositions where the refractory nature of materials such as

tungsten, molybdenum, nickel, or graphite can be retained, while good electrical conductivity may be obtained with copper and silver^{14,32,42}.

Another type of material with good spark quenching properties is the combination of silver and cadmium oxide, which, because no alloying results, also has high electrical conductivity⁴³.

The contact materials may be made by any of the suitable powder techniques. One method is to press and sinter the powder composition sought, with or without final sizing or shaping of the part. Another method that is utilized for making tungsten-copper compositions consists in pressing a bar from tungsten powder and sintering at 1300° C. in hydrogen. The tungsten thus forms a strong porous structure which can then be impregnated with copper. This may be accomplished by placing the part in a graphite boat with copper, heating above the melting point of the latter, and allowing the voids to be filled by capillary action.⁴⁴

No single contact material is satisfactory for all purposes, and a number of different combinations have been developed. These include silver-tungsten, copper-tungsten, silver-graphite, silver-molybdenum, cemented tungsten carbide, and copper-nickel-tungsten. They are used in many installations such as circuit breakers, welding machines, relays, and many types of industrial control equipment.

7. Alnico magnets

Many Alnico magnets of small size have been produced commercially by powder methods^{45,46}. Magnets made in this manner are fine grained in contrast to the relatively coarse grained material obtained by casting methods. The material is uniform throughout with no cold shuts, cracks, blow holes or grain boundary segregation so that a more uniform flux density is obtained. Of particular interest are the close dimensional tolerances which can be maintained in the powder method and the small amount of grinding required in finishing. The composition can be held much more closely than for the cast alloy.

The process is limited economically to the production of small samples. Large samples can be prepared by conventional methods at a cost that would not allow sintered products to compete.

The presence of a highly oxidizable element (9-13 per cent of aluminum) presented difficulties when attempts were first made to prepare Alnico by sintering pressed compacts. To overcome this oxidation, the aluminum is added in the form of alloy powder of 50 aluminum-50 iron composition prepared by crushing and ball milling a casting of the brittle material⁴⁷. In such form, there is practically no oxidation of the aluminum under the sintering conditions which prevail.

Another method to minimize or eliminate oxidation in sintering operations utilizes, in addition to the 50 aluminum-50 iron powder, approximately 2 per cent of titanium hydride incorporated in a powder mixture of aluminum-iron-nickel⁴³. Decomposition of the hydride commences at about 450° C. with release of nascent hydrogen so that during the sintering operation, oxidation is prevented and part or all of any oxide already present may be reduced.*

8. *Metal filters and screens*

Related to the porous metal type of bearing and prepared in much the same way are the metal filters and screens made by powder methods^{3,49}. Bronze, copper-nickel alloys, or pure nickel may be utilized, and porosities up to 80% by volume may be obtained. These filters have been used to advantage in the chemical industry for filtering strong alkaline solutions and other liquids of many kinds. One reported application is as a fuel filter 5 inches long and 2 inches in diameter for a Diesel engine³.

Generally, the filter part can be bonded to steel or copper and made an integral part of the apparatus in which it is to function.

In the manufacture of the filters, the porosity can be accurately controlled. In addition to the methods of producing porous parts as previously described, a highly porous metallic mass can be prepared by sintering the component metal powders (sometimes with volatile additions) in the uncompacted condition using a protective atmosphere and a temperature determined by the type of powders used⁵⁰.

9. *Alloys having special properties dependent on close control of composition*

There are some alloys for special purposes where accurate control of composition and reproducibility of composition are of primary importance. Two such materials are: low-expansion alloys for metal to glass seals, and thermocouple wire for temperature measurement.

An alloy of 54% iron-28% nickel-18% cobalt having approximately the same coefficient of expansion as certain grades of glass is normally prepared by melting and casting procedures. This alloy can be prepared by sintering methods, however, with the same physical characteristics, but with closer composition control and less contamination⁴⁴.

Alloys of nickel-molybdenum and nickel-tungsten have been prepared

* The need for titanium hydride in the preparation of alloys of this type, and the effect of the hydride in controlling oxidation has been the subject of some discussion⁴⁵. Its use is mentioned here only as a variation of the method described above and apart from any effects it may have on the magnetic properties of the alloys to which it is added.

by powder methods for use as thermocouple elements⁴⁴. When these are used with nickel wire as the second element, the couples can operate at temperatures up to 1300° (Ni-Mo) and 1400° C. (Ni-W). Ease of preparation and not reproducibility of composition was probably the main factor in the fabrication of these two types of thermocouple elements since the compositions reported are in the range where relatively large changes in composition produce little variation in thermoelectric voltage.

10. *Parts for Ordnance*

As has already been mentioned, many powder metal parts are being manufactured for use in equipment of the Armed Services, and, while in some instances the parts are made by powder methods only because of expedience, it should be noted that, in all cases definite specifications must be met before acceptance, and a powder metal part that does not meet the rigid requirements has no more chance of acceptance than has an inferior part made by other methods.

Among the parts which have been successfully produced are copper and brass rotating bands for projectiles³⁶. While the cost of the powder metal bands is greater than that of bands made in the normal manner from copper or brass tubing, they compare favorably in actual performance in firing tests both as to behavior on the projectile and wear on the gun barrel.

Improvement of the strength of porous metal bearings has been a factor in their adoption for use in anti-aircraft guns where they may operate under severe conditions. It has been reported that 100 parts are thus utilized in a single gun installation⁵¹.

Another item reported to be in production is an iron powder part of an elevating hand mechanism for both the .30 and .50 caliber anti-aircraft machine guns³⁶. Knurling of the outer surface of the ring part and the marking off of degrees are performed on the part in a coining blow.

11. *Sintered Iron Parts*

Prior to the wartime demands for sintered iron parts, there had been developed a fairly extensive field for peacetime uses particularly in the automotive industry. Bearings had been manufactured for some time and, following this, production had extended to the fabrication of oil pump gears, door catches, cams, and other parts where very high strength is not essential. In general, these sintered iron parts have mechanical properties similar to those of cast iron, but considerable range in properties may be obtained by proper selection of raw material and treatment. Grad-

ing of parts from iron powder into three classes according to the type of product and properties has been outlined as follows^{52,53}:

- Type A* Materials having mechanical properties similar to ordinary cast iron suitable for applications where stresses are very low.
- Type B* Materials similar to Type A but having improved tensile strength, a definite yield point, and a noticeable elongation.
- Type C* Materials having mechanical properties approaching ordinary malleable iron, suitable for applications where stresses, including impact, are moderate.

Prior to 1941, the iron powder used commercially for pressed and sintered parts was of Swedish origin because that was the only powder available in quantity, quality, and at a price which allowed competition economically with established methods of production. Domestic iron powders are now available, however, that are superior to those formerly imported.

Of the sintered iron products manufactured in this country, an interesting example is a small gear for automobile oil pumps²⁷. This gear was formerly made by machining cast iron blanks but was adapted for powder metal production because of greater ease in fabrication at less cost and more satisfactory operation. The gear teeth must be true involute curves with surfaces such that noisy operation and binding are prevented. All of these characteristics can be readily obtained by pressing and sintering, while more difficulty is encountered with cast gears because of the intricate machining work involved. The sintered gear avoids these expensive machining operations, and the teeth have so much better surface finish, and mesh so accurately, that noisy operation is avoided. In addition, the associated porosity is helpful in that oil impregnation assists in smoother and quieter operation.

The pressed gears are lighter in weight than the cast gears, and while the mechanical properties are not of high order, they are satisfactory for the use.

12. Cladding and Duplexing

Powder methods are useful in cladding, duplexing, or any of the processes whereby one metal or alloy may be coated with another for protective purposes, to obtain special properties as in bimetal strip, to obtain hard surface layers on strong, tough backing material, or to obtain a thin layer of relatively high-cost metal of desirable properties on a suitable low-cost backing strip.

For fabrication of bimetal, layers of the respective component metal powders may be placed in the die in the desired proportions and compacted. Upon sintering, an alloy bond is formed between the layers, and the briquette

may be rolled or otherwise worked to the desired thickness^{50,54}. An advantage of this type of bimetal fabrication is the use of alloy bonding at the interface instead of a solder which might limit the operating temperature of the material¹⁴.

13. *Metallic Friction Materials*^{14,55}

The ordinary type of friction material for brake linings, clutch facings, and similar uses is generally composed of asbestos with an organic type of binder. Under normal operating conditions, this type of material is quite satisfactory, but where severe conditions of operation are encountered, the heat generated at the braking surfaces may be sufficient to decompose the binder and cause rapid wearing of the friction facing.

By powder methods, however, a metallic matrix can be formed with admixtures of friction producing ingredients to give a facing that is capable of withstanding the high temperatures generated under severe operating conditions. The exact composition of the facing is determined by the requirements, and a number of different metallic and non-metallic materials are used. Generally, however, the basic ingredient of the matrix is copper to which may be added such modifying metals as tin, lead, zinc or iron. The friction-producing powder is generally an abrasive such as silica or emery which is varied in amount according to the coefficient of friction that is desired.

The metallic elements may constitute only about 50 per cent of the part by volume with the other 50 per cent represented by non-metallic ingredients and pores. In consequence, therefore, the facing is weak and brittle and is usually bonded to a strong backing plate.

The friction materials are prepared in the normal manner by mixing suitable powders, compressing in suitable form, (usually as thin annular rings) and sintering. The sintering operation is generally performed so that the part is bonded to the backing plate at the same time. Finishing operations are then performed to adjust the facing to size and proper shape for use.

14. *Cores for Inductance Coils for Telephone and Radio*^{56,57}

Although the manufacture of cores for induction coils for telephone and radio use does not fit into the field of powder metallurgy as more strictly defined in the Introduction, the procedure is in many ways so similar to the processes described above, and the product of such interest, that a brief description is included in this review.

The coils in communication circuits may operate over a wide range of frequencies from voice frequencies up to millions of cycles per second.

By the use of a finely divided magnetic powder, the particles of which are insulated from one another, the eddy current losses in the cores can be reduced to a level low enough for satisfactory use.

The first metal powder used for cores in the telephone industry was electrolytic iron. This was later superseded by more suitable magnetic materials such as the permalloys.

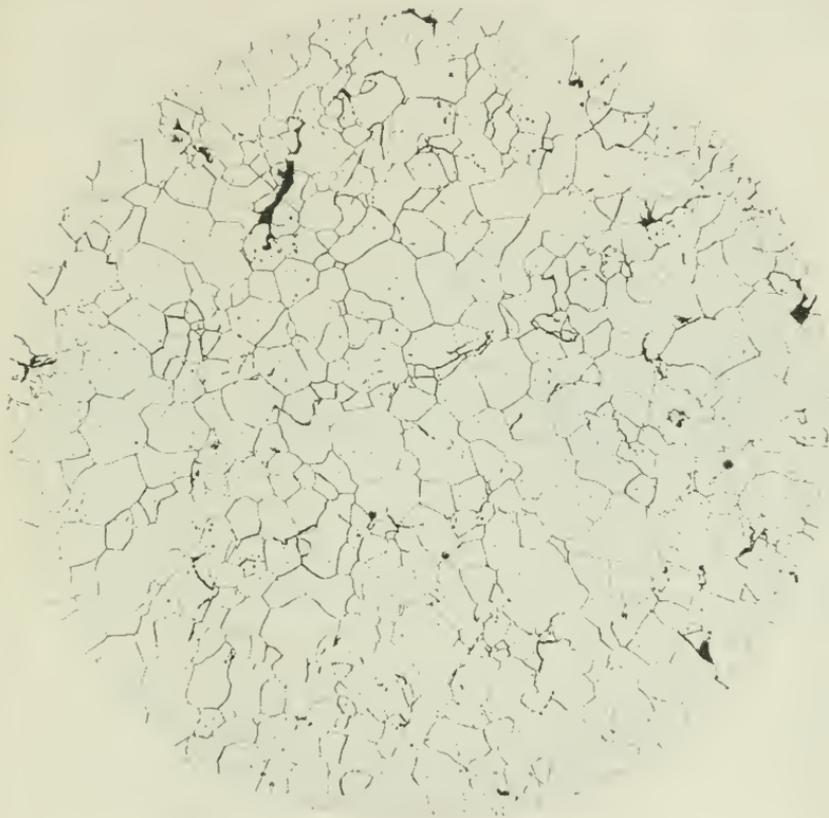


Fig. 4—Brittle molybdenum-permalloy, as rolled to produce fine, equiaxed grain. Magnified 130 diameters.⁵⁷

The procedure utilized to obtain the permalloy powder is worthy of note. Ingots of the desired composition are prepared by melting and casting in the normal manner with, however, the addition to the melt of a small amount of sulphur which acts as an embrittling agent to facilitate pulverization. The sulphur exists as microscopic films of complex sulphides at the crystallite boundaries. At normal temperatures these films are brittle, but at elevated temperatures are either malleable or dissolve in

the iron-nickel solid solution. The alloy can therefore be hot-rolled to small section under controlled conditions to develop a desired grain size, and then cold-worked to separate the individual crystals. Grain size depends upon the degree of refinement in the hot-rolling operation and upon the distribution of the sulphide film around the grain boundaries. Final pulverization is accomplished in an attrition mill, and the product is generally annealed to soften the particles.

The powder is then treated to cover each of the particles with an insulating film that is generally of the ceramic type. The cores are then pressed at about 100 tons per square inch to develop proper density and strength. There is no sintering treatment performed on this type of material after pressing, but the cores are generally annealed to remove pressing strains and restore magnetic quality.

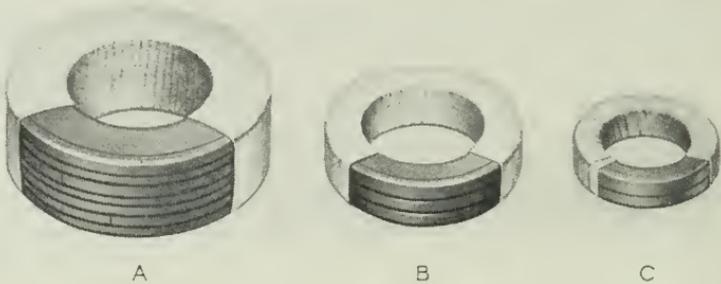


Fig. 5—Relative size of cores for a given duty in a telephone circuit. A is a core made of electrolytic iron powder; B is made of 80% Ni-permalloy powder; C is 2-81 Mo-permalloy.⁶⁸

This powder metal process is thus different from those previously described and is one of the specialty group of materials which cannot be prepared by any other methods. Except for the deliberate coating of the metal particles with an insulating film, and the avoidance of a sintering operation, however, the procedure is that normally followed in preparing powder metal articles.

In addition to the permalloys described, a number of other magnetic powders are used in pressed core form for various applications. These include electrolytic iron, carbonyl iron, and several types of magnetic oxides. Carbonyl iron, in particular, has been used extensively for radio cores where the spherical shape and small size of the particles have been factors in their successful utilization.

ADVANTAGES OF THE POWDER METALLURGY PROCESS

Throughout this paper, numerous advantages of the powder metallurgy process have been indicated as well as some of the limitations. Outlined

below in somewhat more detail are these considerations and others which enter into an evaluation of the process as a whole. In those instances where a product cannot be made in any way except by powder methods, evaluation is easy. But for those products which must compete with standard methods of fabrication, the problem is more complex, and generalizations cannot always be applied.

The following are some of the advantages of the powder metallurgy process:

1. High purity of the metal content of the finished product can be maintained. Control of the manufacture of the powders enables producers to supply metals that generally run well above 99 per cent purity, and often as high as 99.99 per cent for some metals such as tungsten, tantalum, and zirconium²³. Opportunity for additional impurity pickup is slight under the conditions prevailing in the pressing and sintering operations, so that the original metal purity is retained, and may even be improved by oxide reduction or removal of volatile impurities.
2. Composition of the product can be accurately controlled and reproduced^{30,44}. There are no losses due to oxidation or slagging as in melting processes so that the metal content can be quite readily fixed.
3. Structures, alloys, or materials not possible of fabrication by any other method can be produced by powder methods^{29,30,49,58}. These have been adequately described and include porous bearings, sintered carbides, refractory metals such as tungsten and tantalum, and combinations of metals, and of metals and non-metals that do not alloy.
4. High production rates^{58,61,65}, especially on small parts, can be attained by use of automatic presses of the pill tableting type and of continuous type sintering furnaces. One order of forty million, small parts required by the Navy was produced at the rate of 520 pieces per minute by powder methods⁶⁴.

Larger size articles cannot be produced at any such rate because of press limitations which may necessitate hand operation, but with pressed iron parts, high rate of production is one of the factors that allows the process to compete with other standard methods of manufacture.

5. A wide range of certain physical properties can be obtained for any particular material being fabricated^{58,62}. Control can be exercised over such properties as density, porosity, grain size, and strength by variation of the type and size of powder particles, die pressure, and sintering time and temperature.

In some instances such as small Alnico magnets, structures devel-

- oped may have better mechanical properties than the same material in cast form^{44,45,46,47}. The same type of fine-grain structure developed in laboratory samples of iron parts compacted at high pressures and sintered at relatively low temperatures also exhibit superior tensile properties³⁴.
6. The powder method of manufacture may be more economical in many instances due to factors such as rapid quantity production, lower labor costs, ease of setting up for manufacture, conservation of material, and elimination of machining operations^{30,62}. A reported instance of analysis of the normal cost of producing approximately one hundred different units used in a piece of Ordnance equipment revealed that powder metal parts effected a saving of about 70 per cent³³.
 7. Rather close dimensional tolerances^{30,58,61} on small or medium size parts up to about two inches major dimension can be secured, averaging ± 0.001 inch. Closer tolerances of ± 0.0005 inch are attainable and may be even smaller on special production jobs. On larger parts, the tolerance may be in the order of ± 0.002 inch. Frequently, however, accuracy of dimensions is attained only through a coining or re-pressing operation of the sintered part.
 8. There is usually very little material waste associated with powder metal parts manufacture since there is little or no scrap loss^{23,58,62}. Powder losses generally run below 0.5 per cent⁵⁹. In melting and casting operations on small parts, on the other hand, the sprues and risers may be several times the weight of the finished casting. In addition, machining operations on cast parts may remove from 10 to 50 per cent of the metal, and while most of it is recoverable as scrap, it represents a loss in the manufacturing process⁶².
 9. Highly skilled labor is not required for most operations in the powder method^{33,59}. Except for the construction of the necessary dies and die parts, semi-skilled labor may be used. This is of value in industrial plants producing parts for Ordnance because skilled mechanics who would normally be required for machining operations can be made available for other work.
 10. Tooling costs are relatively low in comparison with other high-production methods, and less time is usually required to set up for production³³. Secondary operations such as machining of the sintered products may be eliminated or greatly reduced.

LIMITATIONS AND PROBLEMS OF THE POWDER PROCESS

As has been indicated in several sections of this review, there has been a recent shift in emphasis in the type of product made by powder methods,

and in addition to those materials that are difficult or impossible to make by other methods, parts are now being manufactured in direct competition with those made by conventional, established procedures.

Under these circumstances, economy of production, in addition to technical feasibility, becomes a major factor in the utilization of the process. Of the numerous limitations of the process, some are inherent and definitely limit its application while others are incidental and susceptible to certain measures of control. The more important of these limitations and problems are outlined below:

1. The cost of metal powders is high in comparison with metal for other methods of producing similar parts, and availability of suitable powders is another problem^{44,63}. Both cost reduction and availability have received considerable attention in recent years, and with increased use of metal powders and the large-scale powder production entailed, substantial price reductions have been effected and a wider variety of types of powder have been made available. The development of domestic sources of supply of a satisfactory low-cost iron powder to replace Swedish sponge iron is an example of a successful attempt to overcome a limitation of the process⁵³.

In a final analysis, metal powder costs must be balanced against overall costs before a raw material cost standard can be set up.

2. Die expense^{14,44,61,63} is high, especially for large and complicated parts and for high pressures. New dies are required for each part of different shape and size, and each die must be installed and carefully adjusted for operation. With the entry of the powder metallurgist to the low cost part field, there will be need for more complicated dies to meet the competition of intricately shaped parts produced by casting methods. The tool cost, however, for the powder process is generally lower for a given part than with other processes. Die cost may range from about \$150 for small simple parts up to \$1800 or more for large parts or complicated shapes³³.
3. Sintering furnaces pose many problems in the production of powder metal parts¹⁴. Close temperature control and uniformity are essential for control of dimensional changes in compacts. The fabrication of iron or alloy steel parts requiring higher temperatures than have been previously utilized in the industry add to the difficulties of furnace design.
4. The size and form of powder metal products is limited^{44,62,63}. Large samples require huge presses to obtain the desired compacting pressures and both tool and press costs increase. Increase in size of compacts leads to a non-uniform distribution of pressure and may adversely affect the shape and dimensions of the article in the sintering

- operation. Large presses are usually not of the automatic type, which means hand operation with lower production rates and increased cost. Low apparent density of most metal powders affects die design and limits the thickness of parts produced. A compression ratio of about 3 to 1 is generally assumed, which means mold depth must be at least 3 times the thickness of the finished compact. Other factors of die design are noted under item 7 of this section.
5. The powder process is essentially one of mass production, and a reasonable number of parts must, in general, be fabricated or the costs per unit will be excessive.
 6. On a production basis, powder metal structural parts generally have relatively low elongation, tensile strength and impact strength^{14,44,63}. The mechanical properties of a sintered part depend to some extent on its density, which itself is a function of the type of powder used, the compacting pressure, and the sintering treatment. Because of the voids normally present in powder metal parts, the ultimate properties cannot be expected to be as good as those obtained on cast and wrought materials⁶².
 7. There are a number of design limitations for powder metallurgy parts^{11,44,61,62}.
 - a. Sharp corners should be avoided and internal angles should have fillets.
 - b. Large and abrupt changes in thickness of parts should be avoided, as should uneven cross sections.
 - c. Re-entrant angles, grooves, and undercuts cannot be molded, and if required, must be machined in an extra operation. Internal and external threads, and holes at right angles to the central hole or perpendicular to the axis of pressing, likewise cannot be pressed, and must be machined.
 - d. Length of pressed parts must be comparable to the cross-section area because of pressing limitations. A long section may have a soft central portion of low density.
 - e. There is almost no flow of metal powders during compacting because of friction between particles, and between particles and the die walls^{60,62}.
 8. Although powder parts can be produced to close dimensions by careful control of the compacting and sintering operations and by coining or re-pressing the sintered pieces, tolerances should, in general, be fairly liberal if costs are to be kept down^{14,61}. Close dimensional tolerances may necessitate machining operations to meet specifications. Eccentricity of cylindrical parts may be controlled fairly closely, but concentricity may be troublesome because there must be clear-

ance between die parts and plungers, and the clearances may be cumulative⁶⁰.

9. There is a lack of technical information available for engineers and designers. Tests on metal powders and finished parts have not been standardized, and until such standardization has been achieved the metal powder consumer and the ultimate user of the sintered parts have no check on the respective products. This situation is now being remedied.
10. There are some thermal limitations that may cause difficulties in the sintering process in certain instances⁶³. Some oxides can be reduced only at temperatures above the melting point of the metal itself and prevent effective welding of the powder particles.
11. Metal powders in a fine state of subdivision are readily combustible and must be treated as potential fire and explosion hazards^{20,62}. Zirconium, magnesium, aluminum, and titanium are the most inflammable with iron, manganese, zinc, silicon, tin, and antimony, moderately inflammable. Precautions must be taken to keep dust out of the air in the mixing and pressing rooms, not only because of the explosion hazard, but also because of possible toxic effect on workers.
12. Deterioration of metal powders may occur in storage due to oxidation or absorption of moisture with subsequent chemical reaction to change the composition⁶².

CONCLUSION

This correlated review of some of the more common aspects of powder metallurgy is presented to provide information on an increasingly important production method. The review makes no pretense of complete coverage of the subject, and many important topics such as hot pressing, press and furnace design and operation, sintering atmospheres, and die design and operation have not been described. These and other more specialized topics that are beyond the scope of this paper may be found in the appended list of references.

ACKNOWLEDGMENT

The authors have drawn freely from many of the articles on powder metallurgy published during the past ten years. Wherever possible, reference is made to the original source of the topic material or to associated articles where more complete information may be found. To these sources listed below, the authors are indebted for much of the material presented in this review.

The authors express appreciation to Mr. Charles Hardy for permission to use his pictures of powder metal parts shown in Figs. 1, 2, and 3.

REFERENCES

1. Subcommittee of A.S.M. Metals Handbook Committee. "Terms Used in Powder Metallurgy," *Metal Progress*, 41, 44, 1942.
2. R. Hadfield. "Sinhalese Iron and Steel of Ancient Origin." *Jour. Iron and Steel Inst.*, 85, 134, 1912.
3. H. W. Greenwood. "Powder Metallurgy." *Metal Ind.* (London), 60, 77, 1942.
4. J. H. Robinson, J. H. Breasted, and E. P. Smith, "History of Civilization. Earlier Ages." Ginn and Company, 1937.
5. P. Bergsöe, "Metallurgy of Gold and Platinum Among the Pre-Columbian Indians," *Nature*, 137, 29, 1936.
6. C. S. Smith, "The Early Development of Powder Metallurgy," *Powder Metallurgy, A.S.M.*, 1942, p. 4.
7. W. H. Wollaston, "On a Method of Rendering Platina Malleable," *Phil. Trans. Roy Soc.*, 119, 1, 1829.
8. R. Knight, "A New and Expeditious Process for Rendering Platina Malleable," *Phil. Mag.*, 6, 1, 1800.
A. Tilloch, "A New Process of Rendering Platina Malleable," *Phil. Mag.*, 21, 188, 1805.
- M. Baruel, "Process for Procuring Pure Platinum, Palladium, Rhodium, Indium, and Osmium from the Ores of Platinium," *Quart. Jour. Sci. Lit. Arts*, 12, 246, 1822.
9. S. Gwynn, "Improved Compositions of Matter, called 'Metaline', for Journals, Bearings, Etc.," *U. S. Patent* 101, 863.
10. A. W. Deller, "Patent Survey of Powder Metallurgy," *Powder Metallurgy, A.S.M.*, 1942, p. 551.
11. C. B. Carpenter, "Powder Metallurgy," *Colorado School of Mines Quarterly*, 36 Oct. 1941.
12. W. D. Coolidge, "Ductile Tungsten," *Trans. A.I.E.E.*, 29, 961, 1910.
13. E. G. Gilson, "Genelite," *Gen. Elec. Rev.*, 24, 949, 1921.
14. H. E. Hall, "Developments in Metal Powders and Products," *Powder Metallurgy, A.S.M.*, 1942, p. 18.
15. R. P. Koehring, "Powdered Metal Compact Bonds Babbitt to Steel Back," *Metal Prog.*, 38, 173, 1940.
16. D. O. Noel, J. D. Shaw, and E. B. Gebert, "Production and Some Testing Methods of Metal Powders," *Trans. A.I.M.E.*, 128, 37, 1938.
17. S. L. Hoyt, "Hard Metal Carbides and Cemented Tungsten Carbide," *Trans. A.I.M.E.* 89, 9, 1930.
18. W. P. Sykes, "Cemented Tungsten Carbide Alloys," *Trans. A.I.M.E.*, 128, 76, 1938.
19. A. MacKenzie, "Cemented or Sintered Hard Carbides," *Metals Handbook*, 1939, p. 909.
20. E. Pletsch and T. Edwardsen, "Hazards in the Production and Use of Metal Powders," *Powder Metallurgy, A.S.M.*, 1942, p. 546.
21. R. H. Atkinson and A. R. Raper, "Metals of the Platinum Group," *Jour. Inst. Met.*, 59, 179, 1936.
D. McDonald, "Platinum," *Jour. Soc. Chem. Ind.*, 50, 1031, 1931.
22. W. E. Trout, "The Metal Carbonyls," *Jour. Chem. Ed.*, 15, 113, 1938.
23. C. Hardy, "The Fundamentals Necessary to Apply Powder Metallurgy," Symposium on Powder Metallurgy, *A.S.T.M.*, 1943, p. 1.
24. W. D. Jones, "Principles of Powder Metallurgy," Longmans, Green and Co., 1937.
25. E. V. Crane and A. G. Bureau, "Presses and Processes for Metal Powder Products," *The Electrochem. Soc. Preprint* 85-14, April 1944.
26. L. H. Bailey, "Machinery for Compressing Powdered Metals," *Powder Metallurgy A.S.M.*, 1942, p. 271.
27. W. D. Jones, "Powder Metallurgy," *Chem. and Ind.*, 62, 78, 1943.
28. P. E. Wretblad and J. Wulff, "Sintering," "Powder Metallurgy," *A.S.M.*, 1942, p. 36.
29. J. D. Fast, "The Preparation of Metals in Compact Form by Pressing and Sintering," *Philips Technical Rev.*, 4, 309, 1939.
30. C. Hardy and C. W. Balke, "Powder Metallurgy," *Metals Handbook*, 1939, p. 104.
31. R. H. Leach, "Powder Metallurgy," *Metal Progress*, 36, 350, 1939.
32. G. J. Comstock, "Types of Metal Powder Products—A Classification," *Trans. A.I.M.E.*, 128, 57, 1938.
33. A. J. Langhammer, Symposium on Powder Metallurgy, *A.S.T.M.*, 1943, p. 39.

34. C. W. Balke, "The Effect of Pressure on the Properties of Compacts," Symposium on Powder Metallurgy, *A.S.T.M.*, 1943, p. 11.
35. E. W. Engle, "Cemented Carbides," Powder Metallurgy, *A.S.M.*, 1942, p. 436.
36. F. H. Clark, "Powder Metallurgy," *Mining and Met.*, 25, 81, 1944.
37. C. Hardy, "Powder Metallurgy in the Electrical Field," *Metal Progress*, 36, 57, 1939.
38. H. W. Highriter, "Refractory Metals," Powder Metallurgy, *A.S.M.*, 1942, p. 408.
39. P. E. Wretblad, "Manufacture of Tungsten Metal," Powder Metallurgy, *A.S.M.*, 1942, p. 420.
40. G. H. S. Price, S. V. Williams, and G. J. O. Garrard, "Heavy Alloy—Its Production, Properties and Uses," *Metal Ind.* (London), 59, 354, 1941.
41. H. H. Hausner, "Some Modified 'Heavy Metal' Alloys," *Metals and Alloys*, 18, 1335, 1943.
42. F. R. Hensel, E. I. Larsen, and E. F. Swazy, "Tungsten-Copper for Electrical Contacts," *Metals and Alloys*, 13, 577, 1941.
43. E. M. Wise, "Rare and Precious Metals," *Mining and Met.*, 25, 78, 1944.
44. F. C. Kelley, "Powder Metallurgy," *Electrical Engineering*, 61, 468, 1942.
45. G. H. Howe, "Sintering of Alnico," *Iron Age*, 145, 27, 1940.
46. B. M. Smith, "Alnico—Properties and Equipment for Magnetization and Test," *General Electric Rev.*, 45, 210, 1942.
47. G. H. Howe, "Sintered Alnico," Powder Metallurgy, *A.S.M.*, 1942, p. 530.
48. P. R. Kalischer, "Some Experiments in the Production of Aluminum-Nickel-Iron Alloys by Powder Metallurgy" (also discussion), *Trans. A.I.M.E.*, 145, 369, 1941.
49. W. D. Jones, "The Manufacture of Articles from Powdered Metals," *Mech. World* 111, 91, 1942.
50. J. J. Cordiano, "Copper in Powder Metallurgy," *The Electrochem. Soc. Preprint* 85-4, April 1944.
51. E. E. Thum, *Metal Progress*, 43, 412, 1943.
52. F. V. Lenel, "The Mechanical Properties of the Products of Powder Metallurgy," *Metallurgia*, 28, 189, 1943.
53. A. J. Langhammer, Symposium on Powder Metallurgy, *A.S.T.M.*, 1943, pp. 23 and 24.
54. C. Hardy, "Manufacture and Use of Powdered Metals," *Metal Progress*, 22, 32, 1932.
55. J. F. Kuzmick, "Metal Powder Friction Materials," Symposium on Powder Metallurgy, *A.S.T.M.*, 1943, p. 44.
56. W. C. Ellis and E. E. Schumacher, "A Survey of Magnetic Materials in Relation to Structure," *Metals and Alloys*, 5, 269, 1934; 6, 26, 1935.
57. E. E. Schumacher, "Magnetic Powders," Powder Metallurgy, *A.S.M.*, 1942, p. 166.
58. C. Hardy, "Powder Metallurgy as a Help to National Defense," *Wire and Wire Products*, 17, 247, 1942.
59. C. Hardy and G. D. Cremer, "Production of High Density Parts by Powder Metallurgy Increases," *Mining and Met.*, 23, 509, 1942.
60. E. S. Patch, "Limitations of Powder Metallurgy," *Metal Ind.* (London), 58, 212, 1941.
61. M. T. Victor and C. A. Sorg, "Design of Powder Metallurgy Parts," *Metals and Alloys*, 19, 584, 1944.
62. J. L. Bray, "Powder Metallurgy and the 'Magal' Age," *Industry and Power*, 46, 62, 1944.
63. P. Schwarzkopf and C. G. Goetzel, "Processing Trends in Powder Metallurgy," *Iron Age*, 146, 39, 1940.
64. C. Hardy, Symposium on Powder Metallurgy, *A.S.T.M.*, 1943, p. 54 (Discussion).
65. P. R. Kalischer, "Powder Metallurgy," *Metals and Alloys*, 19, 82, 1944.
66. V. E. Legg, "Survey of Magnetic Materials and Applications in the Telephone System," *Bell System Tech. Jour.*, 18, 438, 1939.
67. P. P. Alexander, "The Hydride Process," *Metals and Alloys*, 8, 263, 1937; 9, 45, 179, 270, 1938.

Abstracts of Technical Articles by Bell System Authors

*Automatic Ticketing of Telephone Calls.*¹ O. A. FRIEND. In January a new arrangement of dial central office equipment was placed in service at Culver City, California, designed to enable subscribers to dial for themselves their short-haul toll calls to other exchanges within the Los Angeles metropolitan area. These calls were formerly placed with an operator, who completed and timed the call and wrote a ticket used for billing. The new equipment controls the automatic completion of the dialed call, identifies the calling line, and prints a ticket showing the calling and called numbers and other information needed for billing. The arrangement applies to the step-by-step switching system and employs senders capable of routing these calls efficiently through a metropolitan trunking network. It affords operating economy together with faster and more convenient service.

*Noise Figures of Radio Receivers.*² H. T. FRIIS. A rigorous definition of the noise figure of radio receivers is given in this paper. The definition is not limited to high-gain receivers, but can be applied to four-terminal networks in general. An analysis is made of the relationship between the noise figure of the receiver as a whole and the noise figures of its components. Mismatch relations between the components of the receiver and methods of measurements of noise figures are discussed briefly.

*Structural Features of Buna S—Relation to Physical Properties.*³ A. R. KEMP and W. G. STRAITIFF. The non-symmetry in the chain structure of Buna S hydrocarbon is discussed in relation to the prevention of crystallization and the impeding of cross linking during vulcanization. This lack of chain symmetry is put forward to account for the poor quality of Buna S vulcanizates in comparison with corresponding vulcanizates prepared from natural rubber. Fractionation data on a regular benzene-soluble crude Buna S indicates the presence of an objectionable broad range of polymer sizes. It is shown that the lowest-molecular-weight polymer fractions in Buna S are not chemically bound in the vulcanizate but remain soluble in chloroform. By removing most of this low polymer from Buna S, the chloroform extract of its vulcanizate decreases accordingly. Vulcanizates were prepared from high- and low-molecular-weight fractions of Buna S. The high fractions were tough, dry, and difficult to handle on the mill; the

¹*Elec. Engg., Transactions Section*, March, 1944.

²*Proc. I. R. E.*, July, 1944.

³*Indus. & Engg. Chem.*, August, 1944.

lower-molecular-weight fractions were soft and sticky. The tensile strength of vulcanizates from the high fraction was somewhat greater than that of the whole polymer, but the modulus was considerably increased. For the low-molecular-weight polymer both tensile and modulus values were much lower. Vulcanizates prepared by mixing natural rubber and gutta-percha hydrocarbons show lower strength than either of the hydrocarbons separately tested in the same formula.

Contributors to this Issue

EARLE 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 Research Metallurgist, Mr. Schumacher is in charge of a group concerned with research and development studies on metals and alloys.

THOMAS SHAW, S.B., Massachusetts Institute of Technology. American Telephone and Telegraph Company, Engineering Department, 1905-19; Department of Development and Research, 1919-33. Bell Telephone Laboratories, 1934-. As Loading Engineer, Mr. Shaw has been chiefly concerned with development problems in loading telephone circuits, including transmission and economic aspects of the loading apparatus.

ALEXANDER G. SOUDEN, Massachusetts Institute of Technology, B.S. 1929; M.S. 1930. Bell Telephone Laboratories, 1930-. Mr. Souden has been engaged principally in metallurgical investigations of non-ferrous alloys, magnetic powder cores, and varistors.

THIS JOURNAL IS
MANUFACTURED UNDER WARTIME CONDITIONS
IN CONFORMITY WITH ALL GOVERNMENT
REGULATIONS CONTROLLING THE USE OF
PAPER AND OTHER MATERIALS



P

