ADVANCES IN ELECTRONICS

VOLUME II

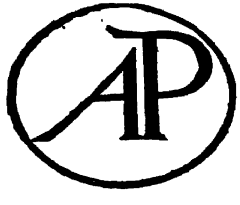
ADVANCES IN ELECTRONICS

Edited by

L. MARTON

National Bureau of Standards, Washington, D. C.

VOLUME II



1950

ACADEMIC PRESS INC., PUBLISHERS
NEW YORK, N. Y.

PRINTED IN THE UNITED STATES OF AMERICA

CONTRIBUTORS TO VOLUME II

DONALD K. COLES, *Westinghouse Research Laboratories, East Pittsburgh, Pennsylvania*

H. FRÖHLICH, *University of Liverpool, Liverpool, England*

G. F. J. GARLICK, *Physics Department, University of Birmingham, England*

P. GRIVET, *University of Paris, Paris*

GUNNAR HOK, *University of Michigan, Ann Arbor, Michigan*

G. LIEBMANN, *Research Laboratory, Associated Electrical Industries Limited, Aldermaston, Berkshire, England*

HILARY MOSS, *Division of Engineering and Research, Electronic Tubes Ltd., Reading, England*

GEORGE T. RADO, *Naval Research Laboratory, Washington, D. C.*

J. H. SIMPSON, *National Research Council of Canada, Ottawa, Canada*

PREFACE

The first volume of "Advances in Electronics" had so favorable a reception as to indicate that a yearbook of this kind has a definite place in the literature. We are, therefore, pleased to introduce this second volume. The editors and publishers wish also to acknowledge the suggestions and criticisms they have received.

The Editorial Board has undergone some changes. Before our first volume was completed, we suffered the inestimable loss of Harry Diamond. This is hardly the place to recount all the accomplishments of Harry Diamond in electronics, but it should be emphasized how much his enthusiasm and vast knowledge helped in starting off this new venture. In a desire to replace Mr. Diamond and at the same time to enlarge our Editorial Board, we obtained the collaboration of H. B. G. Casimir (N. V. Philips' Gloeilampenfabrieken, Eindhoven) and M. Ponte (Compagnie Generale de Telegraphie Sans Fil, Paris).

The contents of the present volume may perhaps show some predominance of physical electronics. This should not be taken as a trend. We intend to cover the field of engineering electronics as much as it is reasonably possible, but it turned out that the best available authors at the time of gathering the manuscripts were more in the field of physical electronics than in engineering electronics. The contents of Volume III, currently being assembled, will show a happier balance between the two fields.

Another fact in connection with the contents of the second volume should be emphasized. Whereas the first volume was written in its entirety by American authors, the second volume shows a large contribution from European ones. This, in addition to the change in the Editorial Board, is an indication of our efforts toward international cooperation. The editors and publishers of this volume feel they may contribute in this way to the free and international exchange of ideas among scientists.

L. MARTON

Washington, D. C.

CONTENTS

Intrinsic Dielectric Breakdown in Solids

By H. Fröhlich and J. H. Simpson, *University of Liverpool, Liverpool, England, and National Research Council of Canada, Ottawa, Canada*

The Microwave Magnetron

By Gunnar Hok, *University of Michigan, Ann Arbor, Michigan*

Ferromagnetic Phenomena at Microwave Frequencies

By George T. Rado, *Naval Research Laboratory, Washington, D. C.*

Microwave Spectroscopy

By Donald K. Coles, *Westinghouse Research Laboratories, East Pittsburgh, Pennsylvania*

Cathode Ray Tube Progress in the Past Decade with Special Reference to Manufacture and Design

HILARY MOSS

Division of Engineering and Research, Electronic Tubes Ltd., Reading, England

CONTENTS

Summary

The review is in three distinct parts. Part I discusses some of the changes in production methods. Part II gives a summary of the principal developments in the form of the finished product. Part III gives a brief critical survey of some of the more theoretical literature on cathode ray tubes, but is mainly an account of design methods which the author has found useful. This section is a subjective treatment addressed to the practising specialist designer. Parts I and II are intended for the more general reader. References are given to some twenty-nine papers covering most aspects of the review.

Introduction

During a large fraction of the period under review, major wars have been in progress among nations having a high degree of scientific competence. The cathode ray tube, along with many other thermionic devices, has found extensive use in military operations—notably in radar. In war it is an axiom that the best is the enemy of the good, so that research and development naturally tend to direct themselves towards the problems of better production of proved designs rather than towards more indefinite academic aims. This fact has had a marked influence on the nature of the progress which has been made. Increased knowledge of design methods and theory, improved production techniques, and a much larger industry, have all combined to yield a far greater range of cathode ray tubes, each with a more specialized field of application. In consequence it is certainly true to say that the average cathode ray tube "display" now represents a marked advance on that of ten years ago. But this has come about—in the author's opinion—mainly because of the closer matching of the cathode ray tube to its particular application. It is also the contention of the author, that although this state of affairs reflects the more wide-spread design knowledge, larger industry, etc., yet it does not necessarily imply fundamental progress, however practically useful the net result may have been.

In the last analysis this question as to whether or not fundamental progress has been made turns on the problem of defining a "figure of merit" for a cathode ray tube, which will be generally acceptable, and which can be applied as an acid test to compare tubes of today with those of ten years ago. This matter is discussed in Part III, and seems to lead to the conclusion that some "fundamental" progress has in fact been realized. For the moment however we must be content to note one fact—that the only two *quite* incontestible ways in which fundamental progress can be made are by reduction of lens aberrations and by improvements in

the fluorescent efficiency of screen powders, and it is equally i
that such progress in the past decade has not been startling.

It must also be remembered that much of the work dur
years was directed towards very specialized tubes for rada
afterglow screens) and that much of this effort has little ap
more normal cathode ray tubes.

I. Trends in Production Methods

1. General—Relative Progress between Nations—Political B

It is first necessary to remember that prior to about 193
no large scale production of cathode ray tubes anywhere in
Most of the large firms who were ultimately engaged in their r
during the war period were already making these devices, bu
small-scale hand-made basis. But around 1936 two factor
to give England an advantage for the next few years. The
establishment of a public television service in the followin
the second was the appreciation of the British Governme
with Germany was inevitable at no very distant date. T
gave Great Britain the lead in this field during the years
Doubtless the German government were equally aware th
war was coming, but their convenient philosophy of a short
offensive—did not require any particular efforts in the catho
since at that time the *offensive* possibilities of radar were
apparent. As for America, her geographical remoteness
prevent much pressure on her cathode ray tube industry a
European politics. Also her philosophy of "commercial"
sored by advertising retarded her television activities, and th
the corresponding cathode ray tube development.

Thus it came about that the gradual change from ha
semi-mass-produced cathode ray tubes began first in Engla
from timing, however, the same pattern of events was
America, although there were appreciable differences in
employed. An attempt is made to treat these differences as
proceeds.

2. Mechanical Features of Electron Gun Design

In spite of ten years activity there still seems no indicat
particular form of gun assembly has marked all-round
Broadly, all forms of mount may be divided according to v

* This statement excludes the U.S.S.R.—in fact the whole of this
reference to the state of the art in Soviet Russia, about which the au
unable to discover anything.

are self jigging or require external jigs, and according to whether they are radially or axially insulated. There is however a continuous spread of types between the completely self-jigging variety (somewhat unusual) and the forms requiring external jigs during the assembly process. The relative economics of the methods depend so much on the type of gun in question, and on quantity produced, that generalizations cannot be made. The nature of the labor available also has a marked bearing on the problem. The various electrodes, which are pressed into shape, are almost invariably made of nonmagnetic copper-nickel alloys, although stainless steel is sometimes used. Fabrication is done by electrical spot welding.

a. Radially Insulated Guns. Here the radial insulation is usually provided by means of thin mica discs, about 0.3-mm. thick. The metal electrodes are tight push fits into holes centrally bored in the micas, and are usually held in place by means of metal collars between which the micas are trapped. Holes punched near the rim of the micas are eyeletted, so that the micas can be threaded on metal rods disposed axially. Each metal electrode is mounted in its appropriate mica disc and the whole gun is assembled by stacking these discs on the axial metal support rods, spacing jigs being inserted between the electrodes while the eyelets are welded to the axial support rods. This is a typical example of a mounting system which is a half-way house between the completely self-jigging technique and the all-external jig type. Reliance is placed on the accuracy with which pressed metal parts can be made, and on the accuracy of punching of holes in mica, to ensure that the electrodes are aligned along a common axis.

This system has the advantage of producing a very rigid gun, and lends itself to quick fabrication on the bench with very simple jigs. The main drawback lies in the fact that since the insulation is radial, for a given electrode diameter the overall gun diameter is greater than would be the case if axial insulation were used. There are important advantages to be gained from increasing certain electrode diameters,[26] especially in all electrostatic tubes, so that radial insulation on high performance guns means rather large neck diameters. In addition large micas of the requisite high quality are very expensive. This latter objection can be avoided, although at the price of increased assembly difficulty, by using segmented micas for insulation as shown in Fig. 1.

This system is essentially English, and has not found much use in the United States—probably owing to the more critical mica supply position in America.

b. Axially Insulated Guns. In one form or another this general system is far more common, and is the one almost exclusively used in

America. The axial insulation is provided by rods, generally of ceramic material about 3–4 mm. in diameter. The rods are accurately ground straight. Many variations exist in the methods for securing the metal electrodes to the rods. Commonly this is effected by welding clips

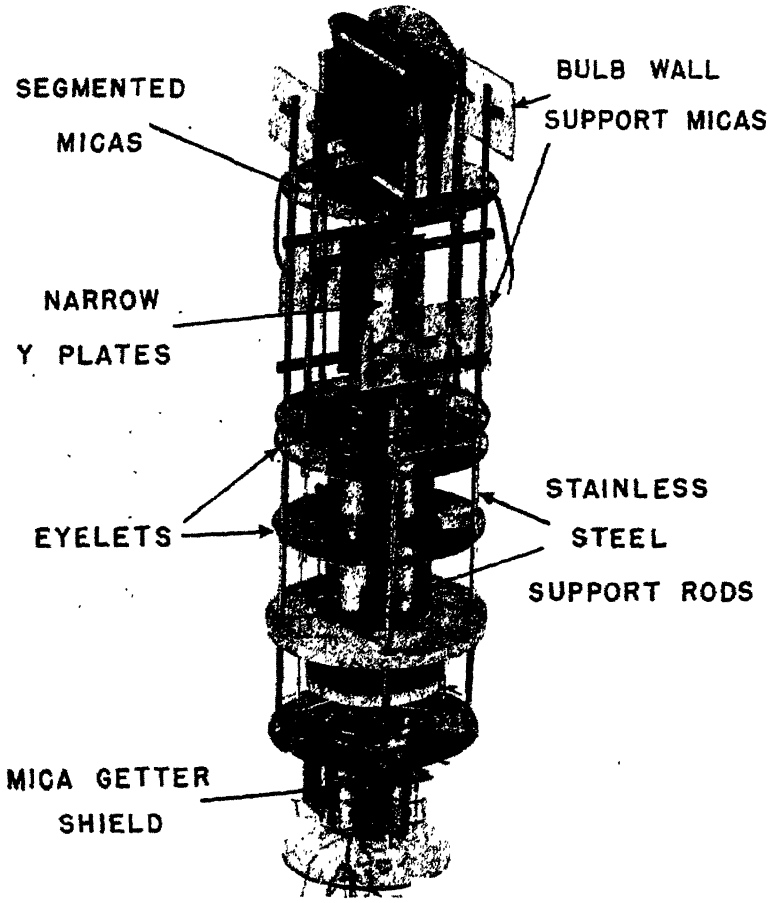


FIG. 1.—Radially insulated gun with segmented micas.

round the rods and relying on friction. A typical example of this type of construction is shown in Fig. 2. Another more ingenious method, developed in the United States, is to use a suitable paste. The whole assembly is held in a jig while the paste is dried by heating. The composition of the latter is based on aluminium oxide and sodium silicate.

Axial insulation combined with complete self jigging has been used by a prominent English firm. The support rods are of glass tubing, and the electrodes have holes punched in them so as to fit closely over the

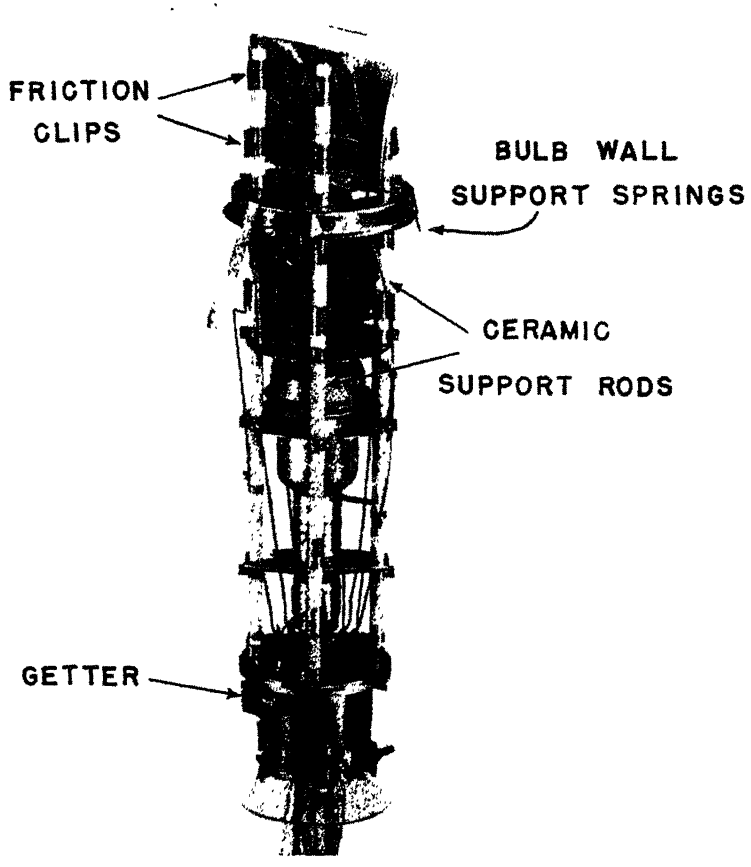


FIG. 2.—Axially insulated gun. (Courtesy of E.M.I. England.)

tubing. Their axial spacing is then controlled by insulating spacing washers cut to the appropriate length. This is illustrated in Fig. 3.

There is no doubt that mounting methods for cathode ray tubes, and indeed for all thermionic devices, give great scope for ingenuity. The simultaneous requirements of high dielectric strength, great mechanical

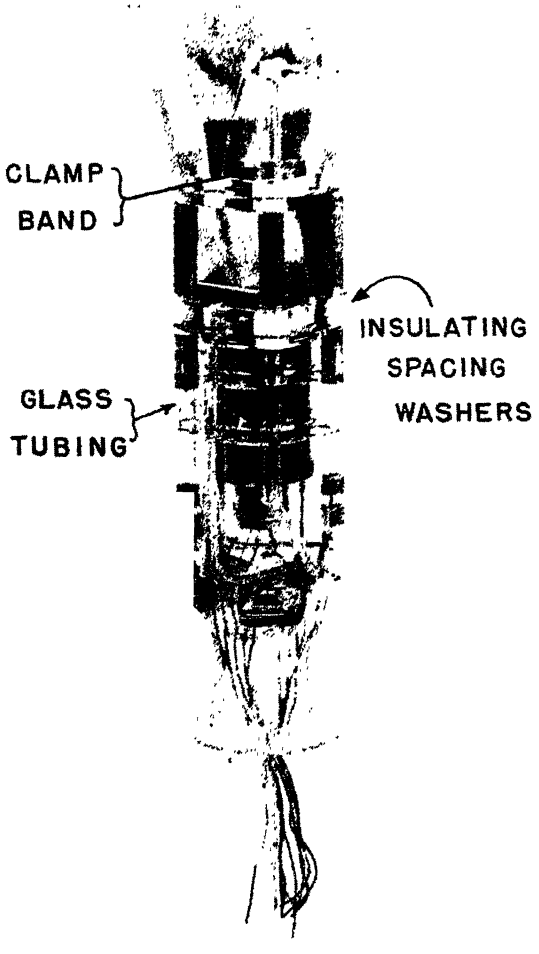


FIG. 3.—Self-jigging axially insulated gun. (Courtesy of G. E. C. Wembley, England.)

accuracy, cheapness, and ease of assembly, together with high temperature operation (at least in the cathode structure) present a combination of difficulties almost unparalleled in any other scientific instrument, in large scale production.

3. Production Techniques for Fluorescent Screens

There has been considerable progress in developing methods of applying fluorescent powder to the screens of cathode ray tubes. These

efforts have been directed to minimizing manual skill and to developing automatic or semi-automatic methods.

In consequence there has been a move away from the earlier technique of spraying the powder on to a sticky base of sodium or potassium silicate, which required considerable dexterity.

Three methods which were previously known but not so widely used have been increasingly employed during recent years. In the first, the screen is coated with a dilute solution of orthophosphoric acid in acetone, and the fluorescent powder is shaken on to this base when it is almost dry. With proper control the binder picks up the correct amount of powder and the excess is then shaken out of the bulb. In the second method the powder is ball milled into a viscous paste of trinitrocellulose dissolved in amyl acetate. The whole paste is then allowed to run over the area of the screen to be coated, the excess running down the walls. An air bake at about 450°C. then destroys the binder and leaves the powder adhering to the tube face. Powder on the tube walls is trimmed off mechanically. The third method consists of settling the powder through some suitable electrolyte such as dilute potassium silicate. The electrolyte is then decanted or siphoned off and the bulb dried by baking in air at about 200°C., preferably with an air stream in the bulb.

The first method has been mainly used in Germany and to a somewhat lesser extent in England. The second method has found main use in England, while America has made most use of settling. It is the opinion of the author, who has had experience in all methods, that settling is by far the best all-round technique.

Recently methods have been developed for "aluminising" cathode ray tube screens.[1] In these a thin layer of aluminium is flashed on to the back of the screen nearest the electron gun. On account of the roughness of the powder layer itself, the aluminum is not flashed directly on to the screen, but on to a very thin intermediate film usually of nitrocellulose. This latter forms a smooth surface on which the aluminium may be evenly deposited. The nitrocellulose is substantially destroyed by heating after the aluminium is deposited. Some comments on this process are given in Section II, 2*c*.

4. Tube Techniques in Relation to Glassworking—Basing and Capping

Glass working techniques have a first order influence on the design of nearly all thermionic vacuum devices, and the cathode ray tube is no exception to this rule. Progress in England has been slow by comparison with that in Germany and the United States. In England almost the only change in the last decade has been the partial introduction of a No. 4 alloy side contact seal for soft glass. This was used during the war to

make a convenient anode connection on certain post accelerator tubes and on one magnetic deflection tube (VCR 530). It is now finding its way into all magnetic television tubes but nothing like to the same extent as in America, where it is almost universal.

The situation is simpler in America since the use of soft glass (lead) has become the almost universal rule for all types of cathode ray tube bulbs. This is not so in Europe, however, where there is still a school of thought in favor of hard (high borosilicate) glasses. The argument is that such glass with its lower expansion coefficient permits hand working operations, and does not require such careful annealing. (It is of interest to note that this argument reflects a general difference between the two continents—namely, the ability of America to afford more expensive capital equipment than Europe.) Both hard and soft glass is therefore in large scale use in Europe, so that for this reason alone there is a greater diversity of glass-metal sealing techniques.

In England, hard glass tubes usually employ tungsten seals made through the usual flat glass pinch. The seal is then overcapped with the normal bakelite molding into which the contact pins are secured. In Germany however, the "all glass base" method has become common. The tungsten or molybdenum pins are sealed directly into a flat cup, and themselves form the contacts.

In England, soft glass tubes, which are in a considerable majority, generally use borated copper seals of the flat pinch or ring pinch type, and are overcapped. In America, the corresponding technique uses a more or less flat "wafer" through which the leads are sealed, and the latter are more usually of No. 4 alloy wire. Again overcapping is used to provide the actual contacts.

In the opinion of the author the general trend will be towards soft glass and No. 4 alloy seals. Probably this will be carried to the logical limit as on "all glass" valves, and the sealing wires will themselves form the pins.

5. Pumping Techniques

In America, and to a somewhat lesser extent in England, automatic pumping has become fairly common. This was not much in evidence in Germany up to the end of the war. Examination of records shows that, in general, pump schedules in Germany were very much more severe than in England and America. Pumping and baking times of 12 hours or more were in common use, whereas in England and America schedules in excess of about 2 hours were exceptional. Auto pumping methods are of course rather pointless with such long schedules, and this probably explains the German lack of interest in rotary or other automatic systems.

In England and America the pumps used bear a close resemblance to those employed in the earlier days of valve manufacture. The systems are usually of the "non indexing" variety, where the whole pump moves round steadily with the cathode ray tube, and the complete turntable consists simply of an assemblage of some twenty or thirty such units spaced uniformly round the periphery. The baking oven occupies perhaps 200° of the whole circle. Rotation times of the order of 90 minutes are representative. Fully automatic methods are still uncommon. Generally the eddy current heating and gettering is done by hand as the system rotates, although the lighting of the heater is usually automatically picked up from a track, as in the case of valves.*

Another more ambitious system in use in America employs a series of trolleys, each containing the complete pump system with its tube. The latter is sealed on to its trolley in a "siding" and the whole trolley is then connected up to a moving track which pulls it through the appropriate ovens and processes. This method permits some saving of space as compared with the circular rotary oven and gives greater flexibility.

There are appreciable differences of opinion as to whether diffusion pumps should or should not be used. If they are used, then mercury is the normal fluid employed as the working medium. This in turn necessitates liquid oxygen cooling traps, but this drawback is usually preferred to the use of oil diffusion pumps in which oil oxidation is the trouble. The more recent introduction of silicone-type fluids may change this position, but so far the development is in its infancy.

There is a definite tendency towards the use of backing rotary pumps *only* with the omission of diffusion pumps. Critics of this arrangement are often apt to forget that the apparent high speed of diffusion type pumps is frequently made useless by the throttling action of the connecting tubing. The pumping stem on a cathode ray tube is often not more then about 6-mm. internal diameter by some 60-mm. long, and at pressures below about 10^{-3} Tor the speed of such a pipe is only approximately 0.4 liters/second. Under such conditions and allowing for the extra throttling for the traps, etc. it is clear that multistage rotary pumps may be almost equally good, and certainly simpler. Very satisfactory† tubes have certainly been made on systems employing only rotary pumps.

The various organic materials used in the preparation of the colloidal graphite wall coating are an insidious source of trouble and may result in screen staining and/or cathode poisoning during the baking and pumping

* The author is not informed on developments in America after 1948.

† For example many hundreds of thousands of VCR139A's were made in Britain during the war on rotary oil pumps without diffusion pumps. This tube is a 2¾-inch diameter all electrostatic monitor. Thousands of 6 inch diameter all electrostatic tubes (VCR97 type) were also so manufactured.

operation. It has also been found that the decomposition products from these materials may contaminate the screen and cause it to "stick" (secondary emission coefficient less than unity) at quite low impact volt velocities (of the order of 6 kv.). These effects may be avoided in many cases by prevacuum baking the graphite wall coating before the fluorescent screen is applied. The temperature should be taken as high as the glass envelope will stand and in any case at least fifty degrees higher than the temperature it will attain during the final baking and pumping operation. After such treatment it will generally be found that the "sticking" potential of most sulfide materials is not lower than about 10 kv. This is an important point, for 10 kv. will permit a very satisfactory standard of brightness on a 10-inch diameter kinescope tube, so that aluminization on the score of screen potential maintenance is rendered unnecessary (see Section II, 2*c*). In addition, although the double pumping technique may seem extravagant it must be remembered that the *final* pumping time is reduced thereby, and that the maintenance of very good vacuum conditions during the first vacuum bake is unnecessary. Pressures of around 10^{-3} Tor are low enough.*

Experience has convinced the author that it pays to eddy-current heat the cathode and grid structure of the electron gun during the pumping process. This should be done immediately after the cathode is formed by the heating of the carbonates. There is evidence that during this period when the cathode is first lit, some emission poisoning substance is evaporated on to the grid surface from the cathode material. If it is allowed to remain there, it damages the emission to a greater or lesser extent. In particular it has a fairly low ionizing potential of around 10 volts, so that operation of the grid under positive conditions will severely damage the emission. This seems the origin of the so called 10 volt effect so well known in various forms to thermionic engineers. A theory has been evolved that HCl from the glass bulb wall is responsible, and this is supported by considerable evidence.[2]

It is certainly possible to make cathode ray tubes without eddy-current heating the mount, but in that case the emission tends to be more marginal, and it is necessary to clean out the pump system frequently. The only real drawback to eddy heating is the danger of deforming the carefully aligned electrodes in the triode. It is unwise therefore to allow the temperature during the eddy-current heating to exceed some 700°C.

6. Production Test Methods

Production of cathode ray tubes has not yet reached quantities at which automatic testing would be economic. In any case the develop-

* This paragraph is rather a statement of the author's personal opinion. Usual practice applies the fluorescent screen *before* the graphite wall coating.

ment of such test apparatus would present great difficulty since the estimation of some aspects of performance is still rather subjective.

The tests applied vary considerably but are usually on the following lines. The grid cut-off voltage is first measured, usually without deflecting signals applied since the cut-off voltage determines several important tube parameters (Section III, 4*a*). A close raster is then applied, so that the screen is excited over its whole operating area. With a standard "grid drive" (i.e., difference between grid voltage and cut-off voltage) the light output or screen brightness is measured. At the same time the test operator assesses the uniformity of screen brightness, generally subjectively. This is especially significant for television tubes, as is also the color. Measurement of the latter is very difficult, and it is usually left to the judgment of the operator. A useful refinement is to provide "limit" colors formed by the illumination of standard color transparencies at operational brightness. These should be made to give limit of blue, limit of green, etc., thus roughly defining the extremes beyond which the so called "white" television screen may not wander.

In producing the close raster it is in many ways preferable not to attempt to achieve uniformity of excitation. The latter demands crossed *linear* time bases which are not too easy to establish. It is far easier to form the raster by two sinusoidal sweeps of high nonsynchronous frequency ratio. If the brightness is measured over a rectangle at the center of the raster, such that the side of the rectangle does not exceed about one third of the side of the raster, then it is easily shown that the raster brightness which would be attained with linear scans is $\pi^2/4$ times as large.* In practice relative readings are in any case generally sufficient.

Focus testing presents many difficulties. Center focus may be checked by microscope and measuring scale, although there is some trouble owing to the gaussian nature of the line brightness. The edge of the trace is not sharp. Even more difficult is the measurement of deflection defocusing. Comparison of the trace with some standard light-illuminated photographic pattern provides a partial solution. Fortunately it is found (as is to be expected from theory) that if the gun centering alignment, etc. is in order, and if the cathode emission is normal, then the focus quality is almost certain to be satisfactory. It is therefore permissible to omit a focus test if the other factors are examined, and provided the resolution of a prototype tube is adequate.

The testing of radar tubes is more involved, since screen afterglow is a vital parameter. This subject has provoked intense controversy throughout the war years and is too specialized to be treated here.[3,4,5]

In making several types of test, e.g., centering, focusing of stationary

* Assuming that the screen shows no current saturation.

spot, alignment, etc. it is difficult to avoid screen burning, especially with high-beam power tubes, unless the grid is pulsed. A suitable blocking oscillator or multivibrator, with a large space-mark ratio is very helpful, therefore, in driving the grid so as to maintain a low *mean* beam current while giving peak currents and beam angles of operational level.

II. Developments in the Finished Product

1. General

The large growth in the size of the cathode ray tube industry has resulted in the appearance of a much greater range of tubes. Most of these are of no special interest since they merely represent small variations in local technique. In what follows the author attempts to indicate those developments which, in his opinion, are most worthy of note, but the choice is recognized as being subjective, and no attempt is made to include all the tube types which have appeared.

2. Classification of Principal Tube Types

Ten years ago there were approximately three basic types of cathode ray tube more or less available as "standard." Firstly there were the small 3–5 inch diameter gas focused electrostatically deflected tubes for low frequency laboratory work; then secondly a somewhat similar range of tubes which were electrostatically focused, and finally a range of narrow neck all-magnetic tubes with screen diameters from 10–15 inches. The last named were intended for television purposes. Other types of tube were available to those working in the industry, but were scarcely articles of general commerce.

Very broadly the modern position may be summarized thus:

(a) The gas focused tube is completely obsolete.
(b) The oscillograph all-electrostatic tube has considerably improved in all-round performance and numerous special forms have been developed such as
    (1) High voltage all-electrostatic tubes usually with multistage post acceleration giving very high writing speeds.
    (2) Double beam tubes.
    (3) Low transit time tubes intended for UHF use.
    (4) Radial deflection tubes.
(c) The hybrid electrostatically focused, magnetically deflected tube has appeared mainly on account of its suitability for centered P.P.I. displays.
(d) All-magnetic tubes intended mainly for television purposes have been improved mechanically. They have also been improved

electrically by the addition of such features as ion traps and/or aluminizing.

(e) Projection tubes and associated optical systems have begun to make their appearance.

(f) Several new types of afterglow screens have been evolved which have applications in radar and occasionally in oscillography.

We now proceed to examine these developments in more detail.

a. Developments in Oscillograph Tubes. In illustration of the "all-round" improvements which have occurred we shall consider two fairly representative general purpose oscillograph tubes made by A. C. Cossor. The type "26" was designed about eleven years ago, and the type "88" about two years ago. (Fig. 1 shows the gun of the latter.)

TABLE I

Parameter	Old type "26"	Modern type "88"
Over-all length (mm.)	375	400
Useful screen diameter (mm.)	95	88
Sensitivity "X"	480/v	750/v
Sensitivity "Y"	450/v	1100/v
Spot diameter at 1.5 kv. final anode potential	1.5 mm.	1.3 mm.
Maximum beam current at 1.5 kv. final anode potential (for satisfactory focus)	20 μa	25 μa
Maximum operating potential	2 kv	6 kv
Figure of merit (compared at 1.5 kv)	50	150

Table I shows this comparison. The spot diameter is defined as that diameter at which the spot density has fallen to one fifth of that on the center of the fluorescent spot. The figure of merit is based on the formula discussed in Section III, 6.

The all-round improvement is very largely due to better design of deflector plates so as to achieve maximum sensitivity, and to some reduction of lens aberrations.

TABLE II

Heater rating	6.3 v, 0.6 amp.
Grid voltage for cut-off	−115
1st and 3rd anode potential	3,000 volts
2nd anode (focus) potential	840 volts
Final intensifier potential	20 kv
Deflection sensitivity	
Front plates	0.22 mm./volt
Back plates	0.23 mm./volt
Spot diameter at 100 μa beam current	0.9 mm.

As an example of high speed post accelerator type tubes intended for single stroke transient analysis we may mention the Dumont 5RP11.[6]

A photograph of a very similar type of tube which also incorporates an ingenious coaxial deflector plate connection system for UHF working is given in Fig. 4. The main operating characteristics are given in Table II.

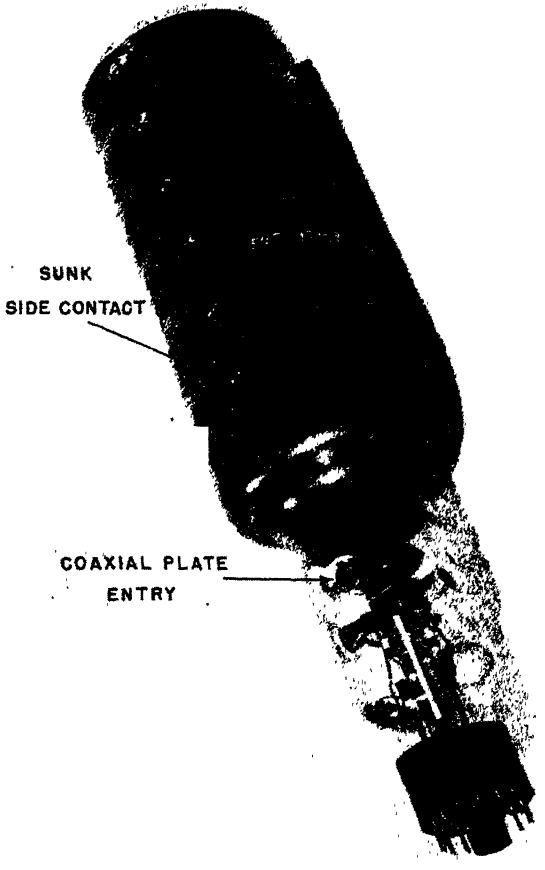


FIG. 4.—High writing speed post accelerator tube. (Courtesy of Dumont.)

The maximum writing speed for this tube approaches the velocity of light. (In passing we may note that there is nothing inconsistent with relativity theory if the writing speed exceeds the velocity of light. There is no reason from a relativity standpoint why the writing speed cannot be infinite.)

The introduction of such tubes will probably mean that the all metal continuously pumped type will gradually become obsolete.

It should be noted that the fact that this high performance tube is of the post accelerator type does not mean that post deflection acceleration enables an inherently better tube to be made.[7] It is shown in Section III, 6 that the figure of merit of a post accelerator tube is not higher than that for the more normal type, given a definition of a figure of merit which will probably prove generally acceptable. Post deflection acceleration is merely a convenient engineering way of achieving certain compromises in cathode ray tube design—as usual at the expense of certain other features. It is the author's opinion that exaggerated claims have at times been made for post accelerator tubes, which have given non-specialists the idea that this system has some fundamental all-round advantage.

Double beam tubes have appeared in several forms. In England the only form available uses a common electron gun, the beam being divided by means of a "splitter plate" placed immediately after the final anode limiting stop. This method has the important advantage of simplicity and avoids any increase in neck diameter. Its drawbacks are that the deflection system nearest the gun is necessarily asymmetric, and that the relative intensity of the two beams cannot readily be controlled without some complication.*

In America and the continent the method of employing two independent guns is used. Generally the final deflector plate system is common to both beams, since the whole essence of the double beam tube is to provide a common axis of deflection. For special purposes however, completely separate guns and deflectors have occasionally been assembled in one envelope. The objection to separate guns is the greater expense and the necessity for a much larger neck diameter.

It is the opinion of the author that the utility of double beam tubes has been over-emphasized. It will often be found that the more experienced oscillographer prefers to use a single beam tube and to dispense with the time base in many applications where time bases and double beam tubes may provide the most obvious solution.

Radial deflection tubes have appeared in small quantities, but there is as yet no *commercially* available tube of this type in England. The attractions of this system are the simplicity of the time base and its increased length for a given tube diameter.[8] The great drawback is the difficulty of modulating the circular trace with the radial deflection. If

* In point of fact this complication need not be more serious than the provision of a very small external permanent magnet which can be moved about in the region of the triode so as to alter the direction of the beam.

the tube is manufactured with a spike or cone supported only from the screen and projecting towards the electron gun, then the system tends to be mechanically fragile, and the center of the screen suffers from a disfiguring seal to which a connection has to be made *across* the face of the tube. On the other hand if the radial electrode is supported from the electron gun itself, then there is trouble from shadowing from the support webs. With either system it is impossible to make the deflection linear and deflection defocusing tends to be severe. Furthermore the apparent simplicity of producing the circular trace from double electrostatic deflection by means of simple phase splitting is illusory.[8] Such a circular trace is not generated by a conical path in the tube itself—the cone has two apices, one at the center of each deflector plate system. For this reason the radial deflection sensitivity is nonuniform around the periphery of the circular trace as well as being nonlinear. Uniformity of sensitivity can be attained by the use of a crossed field system of electrostatic and magnetic fields, which is somewhat less convenient than producing the circular trace from a phase splitting circuit.

The order of mean sensitivity obtainable from the use of a central spike or, preferably, cone, is 200/V. mm./volt. The associated capacity to ground of the spike is of the order of 7 $\mu\mu$ F.

b. "Hybrid" Electrostatically Focused Magnetically Deflected Tubes. The advantages of magnetic deflection over electrostatic deflection in respect to beam brightness are now well recognized.[9] This arises because it is possible to design deflector coils which permit the deflection of wider beams for a given amount of deflection defocusing and given scanning angle than is possible with known electrostatic deflection systems. For this reason it is fairly certain that the future of television lies with magnetically deflected tubes.

On the other hand the convenience of electrostatic focusing is also well recognized. The focusing electrode can easily be designed so as to consume no current, and the electrostatic system has the important advantage that no positional adjustment of a coil assembly is required.

These facts have brought about the use of tubes with electrostatic focusing and magnetic deflection. They usually assume the form shown in Fig. 5 where the main focusing lens is formed by the electrode A_2 and the wall coating of the tube neck. This method has the important advantage that the insulation between A_2 and the neck is good.

In this system the beam limiting aperture should be placed in A_1 and the A_1 potential made somewhat higher than A_2 to prevent secondary electrons from this limiting aperture reaching the screen.

The main drawback of these hybrid tubes is the incidence of ion burn in a concentrated spot at the center of the screen. It is a property of an

electrostatic field that it treats all charged particles alike, irrespective of their charge/mass ratio, so that the electrostatic field concentrates the ions into a focused beam just as it concentrates the electrons. On the other hand the magnetic deflection field has only a very small influence on the ions so that they remain more or less undeflected and strike the center of the screen.

For centered P.P.I. tubes this does not matter in the least, and the method has much to recommend it for that application. For television or off-centered P.P.I. systems however, it is necessary to consider carefully how ion burn is to be prevented if hybrid tubes are employed.

c. Improvements in Magnetic Tubes—Ion Traps—Aluminization. Apart from the already-mentioned development of neater glassworking techniques such as sunk side contacts and pressed glass bases, the

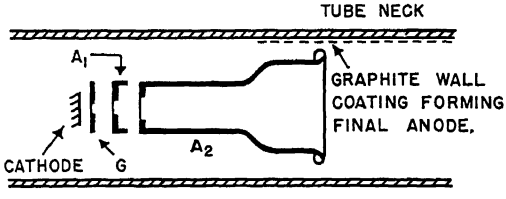


FIG. 5.

advances have been connected with ion traps and/or aluminization of the screen.

The basic idea of ion trapping goes back many years, but it is only more recently that traps of practical form have been developed. The simplest trap consists of bending the whole neck of the tube and correcting the electron path by a magnetic field too weak to have any appreciable effect on the ions. Such a method is not practical in manufacture since the neck departs from symmetry about a line and this raises great difficulties in processing.

Two methods of avoiding this departure from symmetry of the actual neck have been developed. The first system, American,* is diagrammatically illustrated in Fig. 6. The beam, a mixture of ions and electrons, is accelerated by the first anode A_1, and then to a much higher velocity by the second anode A_2. The gap between these two cylinders is cut skewed, so that the field directs the beam into the side of the A_2 tube. The component of electrons is then pulled straight by means of a magnetic field in the usual way, and passes through the hole H. Subsequent focusing is usually magnetic, although it will be appreciated that the same principle of ion trap is applicable to electrostatic focusing.

* American Patent Nos. 2,211,613 and 2,211,614.

The second method, developed in England,* makes use of the fact that if the first anode hole is unsymmetrically placed in respect to the grid hole then the beam is tilted. This is illustrated in Fig. 7. It will be noted that in this diagram not only the first anode but also the grid hole is off the axis of the gun. This bodily displacement is made so that

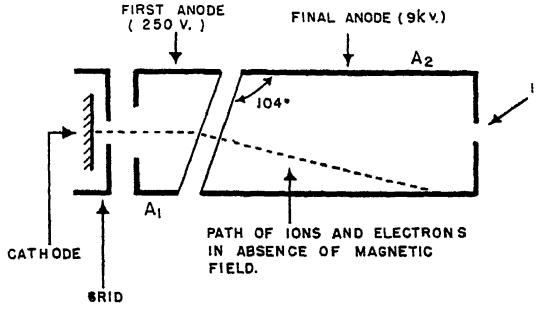


FIG. 6.—American type ion-trap.

when the trajectory is corrected by means of the usual magnetic field situated over the grid/anode gap, the resulting trajectory not only passes through the final limiting diaphragm H, but also is parallel to the neck of the tube.

Figure 8 shows curves relating the angular tilt of the beam with displacement of the triode components, for various anode to grid spacings f.

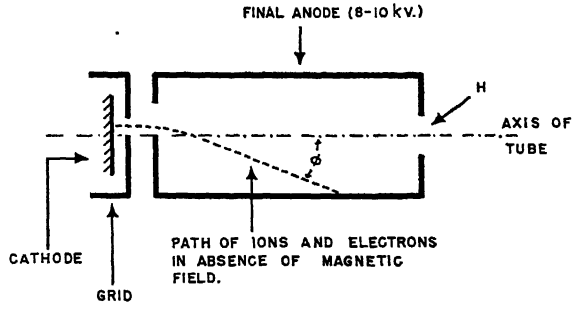


FIG. 7.—British type ion-trap.

The advantage of this method over the American lies in the fact that it avoids the necessity for supplying two anode potentials to the electron gun. On the other hand it is less flexible than the American system, in which the degree of tilt of the beam is controllable by the ratio of the potentials applied to the two anodes.

* British Provisional Patent No. 25964/44.

With both these systems of ion trapping it is important to confine the magnetic correcting field to a relatively short axial distance in order to bend the beam straight without excessive displacement from the axis. The magnetic field should be superimposed on the electrostatic skewing field. Pole pieces inside the tube assist in localizing the magnetic field.

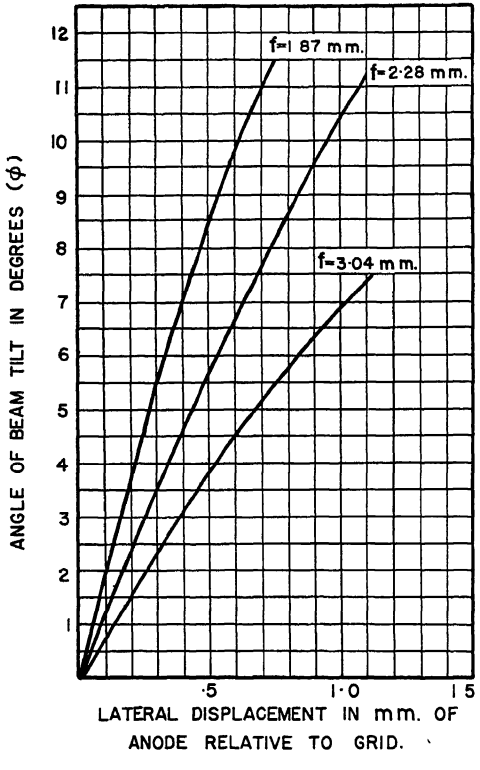


FIG. 8.

The technique of aluminization seems to have originated in America towards the end of the war.[1] The advantages claimed for it are:

(a) Increased light output by reflection of light from the side of the screen nearest the gun.
(b) Increased contrast by suppression of reflected light from the walls of the tube.
(c) Maintenance of screen potential by conduction.
(d) Elimination of ion burn.

Notwithstanding these claims the author has considerable doubts as to whether the process will find the wide use that is suggested. To begin

with, the technique is very difficult and singularly unsuited to mass production. As has already been stated the aluminium is evaporated on to the surface of a collodion film which is deposited on the surface of the fluorescent powder. The problem is to deposit this collodion film in such a way that surface blemishes and ridges are eliminated; such defects result in unequal thickness in the final evaporated aluminium film, and this in turn causes severe brilliancy variations over the screen. It seems fairly certain that even with good manufacturing control, the breakage will be serious, and this factor is vital in the highly competitive television tube market.

The claims themselves also require scrutiny. Claim (a) seems rather doubtful to any extent, since the powder grains themselves will attenuate the reflected light from the back surface of the screen. Claim (b) ignores the fact that the light scattering properties of the powders used are good, so that only a small amount of ambient light is sufficient to heavily reduce contrast. The practice of operating tubes in darkness (when the contrast increase due to aluminization is certainly apparent) is recognized as likely to cause eyestrain and is not acceptable. For these reasons claim (b) has only doubtful validity.

Claim (c) is perhaps the most important. The general characteristics of fluorescent screens in respect to light output against final anode potential at constant beam current are shown in Fig. 9, curve (a). At a potential around 6–7 kv the curve begins to flatten out, this being due to the failure of the screen to follow the potential increases on the final anode. (Secondary emission ratio falls below unity.) On the other hand the effect of aluminization of a similar screen is to produce a voltage/brightness curve of the form of curve (b) Fig. 9. The close physical contact between the aluminium and the screen surface enables the latter to follow the potential of the final anode to which the aluminium layer is connected. Thus the kinetic energy of impact of the electrons corresponds to the final anode potential, and not to the "sticking" potential of the powder itself. It will be noted from Fig. 9 that at low potentials the efficiency of the aluminized screen is lower than that of the non-aluminized type. This is explained by the energy loss in passing through the aluminium film, which is far outweighed at the higher velocities by the gain due to maintenance of screen potential. However, it is probably safe to say that up to the potential at which the screen begins to "stick" the process of aluminization will not increase the screen efficiency appreciably.* Any such increase must be due to the small effect of

* Experience seems to show the increase to be of the order of 30% for the same screen voltage and beam current once the beam voltage is sufficiently high and/or the aluminium layer sufficiently thin to avoid much energy loss in electron penetration.

increase of light output due to reflection from the back of the screen more than compensating for the loss in electron impact velocity.

It is therefore the opinion of the author that the following "theorem" is valid:

"Do not aluminize until the screen operating potential is required to be higher than that which can be achieved by secondary emission."

In Section I, 5 mention was made of a technique which assists in the maintenance of high "sticking" potentials. It is the view of the author that such methods give far more promise of meeting competitive television tube conditions, at least on the commoner sizes up to about 10-inch

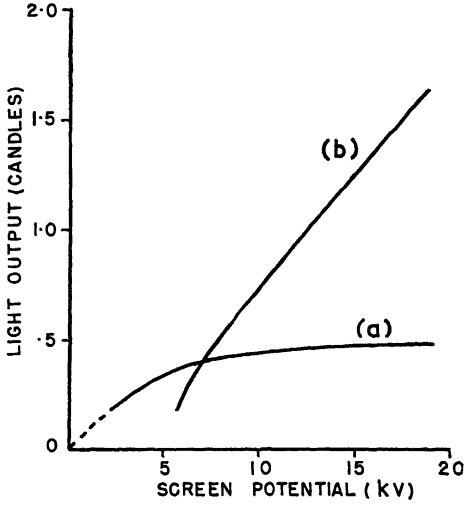


FIG. 9.

diameter at which operating potentials of not more than about 10 kv are adequate to attain satisfactory brightness.

Very recent evidence seems to show that claim (d) is not fully justified. "Postponed" rather than eliminated would be a better phrase. The main result of aluminization would seem to be the transfer of the breakage due to ion burn from the manufacturer's test bay to the consumer's premises.

d. Projection Tubes and Associated Optical Systems. The possibilities of projection tubes have long exercised the minds of cathode ray tube engineers, but various technical, and to a large extent commercial, difficulties have prevented large scale use of such methods. There is nothing impossible in an engineering sense in projection systems—the difficulty is mainly economic.

So far as the tubes themselves are concerned, Section III, 5*a* gives an outline of design methods whereby their geometry can be deduced from direct viewing prototypes. The design of the tubes cannot be dissociated from the optical system with which they are used, but if we assume that the light gathering power of the latter is some 30%, anode voltages of the order of 25 kv are required to produce television images with areas of the order of 2.5 square feet. This anode voltage might, however, be greatly reduced by the use of directive viewing screens, although this development is still in its infancy and it is difficult to predict how successful it will be.

However even assuming for the moment that directive screens are not employed, it is clear that projection systems are far from being impracticable. Aluminized transmission type screens, capable of operating at 30 kv, can be fairly readily produced and have useful lives of some hun-

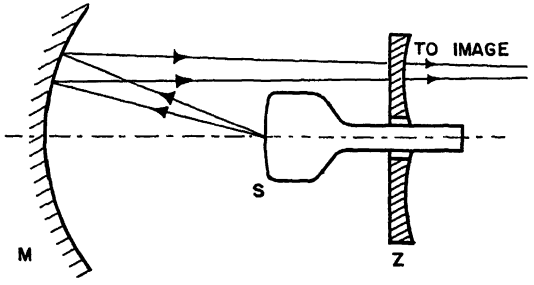


FIG. 10.—Principle of Schmidt mirror system.

dreds of hours. The problems of aluminizing are much simplified since the screen of the projection tube will normally have an area of only some 10 square inches, at the most.

By the use of RF or pulse driven power sources,[10] there is no great expense involved in generating the high tension supply for the tube. The increase in scanning power, proportional to anode voltage, is somewhat more serious but not prohibitively so.

On the optical side it is clear that the only hope lies in the use of mirror type systems of the Schmidt form. This will be appreciated from the fact that a normal transmission type lens of $f/2$ aperture has a light gathering power of only some 5%, and transmission lenses of larger aperture become quite impossibly expensive. Figure 10 shows a sketch of the normal type of Schmidt projection set up. The cathode ray tube is mounted facing a large spherical reflecting mirror—*M*. Rays from this pass through the correcting plate *Z* which has a nonspherical surface optically worked. With these systems it is necessary to employ a high grade convex surface *S* of accurately determined radius on the cathode

ray tube, in order that the image plane shall be flat. Alignment between the elements of the system is critical.

The main problem is one of cost. If the whole system were made of glass, then the optical working of the mirror M and the correcting plate Z would be prohibitively expensive. An ingenious attack on this difficulty has been made by manufacturing the system in optical plastics which are cast from master molds.[11] In particular this avoids the great difficulty of working a nonspherical surface on Z. Quantity production of such systems is starting in England with the additional feature that the correcting plate and mirror are mounted together in a preadjusted rigid framework into which the cathode ray tube can be inserted and

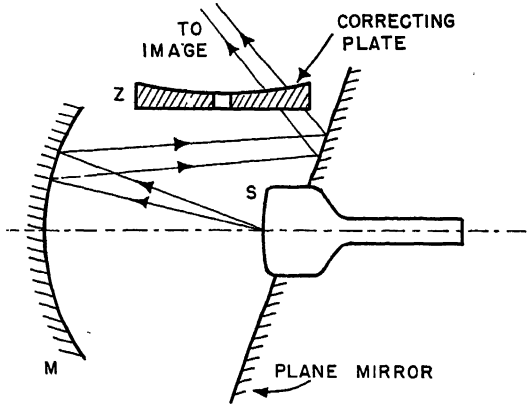


FIG. 11.—Principle of "folded" Schmidt mirror system.

located.[11] Another rather different optical system, termed a "folding" Schmidt system has also been described.[10] This latter has the advantage of lessening the length of the optical axis and so shortening the cabinet size. It also results in greater light gathering efficiency. The system is sketched diagrammatically in Fig. 11. It is the contention of the author that projection television will not come into large scale domestic use until the number of lines is raised above about 700. This opinion is based on considerations of visual acuity and viewing distances possible in the average home.

3. Fluorescent Screens

A large fraction of the more academic wartime work on cathode ray tubes was directed to this subject, and very especially to afterglow screens. Developments in nonafterglow screens have been mainly confined to improved application methods as discussed in Section I, 3.

The double layer, cascade screen (American terminology P7 type)

received much attention. This was introduced first by the British in 1938 in connection with their long range early warning radar system type "C.H.," but was later extensively employed in America.

This screen consists of two distinct layers. The one nearest the glass envelope is usually a copper activated zinc cadmium sulfide, giving a long green/yellow afterglow. Over this is deposited a second layer consisting of a silver activated zinc sulfide having a blue fluorescence and negligible afterglow. Ideally the yellow layer is stimulated only by near ultraviolet light emitted from the blue layer, although the characteristics of the combined screen may be modified to suit various operating conditions by variation of the blue layer thickness, thus permitting varying degrees of direct electron excitation of the yellow layer. Variation of the yellow layer thickness also alters the afterglow performance, so it is clear that a large range of characteristics can be obtained with the same powders.

When adequately excited this type of screen is capable of giving a longer afterglow than any other known system. It therefore found extensive use in low rotation speed P.P.I. sets when sufficient beam power was available. Originally however it was developed for the "C.H." application mentioned, as a means of discriminating between a repeating signal (from target) and a nonrepeating signal (noise). It owes this property to the slow build up characteristic of the yellow layer, aided by light discriminating filters.

As the war progressed, the use of short range radar greatly increased. This permitted higher rotation rates on P.P.I. systems, and the long afterglow times of double layer screens became unnecessary. Indeed they were undesirable on account of screen "clutter." Furthermore with increasing target speeds, successive echoes did not entirely coincide, so that the main quality required from the screen was no longer "build-up" but the ability to respond to relatively weak single signals against a comparable noise background.

Use of the yellow component only of the double layer screen, under direct electron excitation, was found to provide a reasonable solution. This class of screen was widely employed by the British in the H_2S blind bombing equipment.[12]

An important further advance was the introduction of "fluoride" screen materials. Zinc magnesium fluoride activated by manganese has proved very useful in many high rotation speed P.P.I. sets. The main characteristic of this class of material is the near exponential decay, in contrast to the inverse power law decay of sulfides. In consequence the fluorides give greater initial afterglow brightness and a shorter "tail," thus more closely approximating to the desired "square topped" afterglow characteristic.

It has been shown that an exponential decay can be explained on the assumption of only one electron trap depth in the phosphor, whereas the more common inverse power law is due to a distribution of such trap depths.[13]

III. Progress in Design Methods and Theory

1. List of Symbols

All potentials are measured with respect to the cathode.

E—generic symbol for potential difference
E_s—potential of fluorescent screen relative to cathode
E_1—potential of first anode
E_2—potential of second anode, etc.
I—generic term for current
I_a—total current from cathode
I_s—current reaching fluorescent screen
E_c—grid cut-off potential (negative)
E_d—grid drive, i.e., $|E_c| \sim |E_g|$ where $E_g < E_c$
E_g—grid potential
D—diameter of grid hole in mm.
t—thickness of grid material in mm.
b—cathode to grid spacing in mm.
f—anode to grid spacing in mm.
} See Fig. 12.
k—constant of scale
K_1, K_2—constants
ρ_c—cathode emission density
α—semi-vertical angle of cone of rays from triode measured at 250 mm. from grid plane
θ—semi-angle of convergence of rays in beam at any point
A—area of fluorescent spot
X—deflection at screen along X-axis for 1 volt potential difference between X deflector plates.
Y—deflection at screen along Y-axis for 1 volt potential difference between Y deflector plates
η—screen efficiency, i.e., light output/watt beam power
T—absolute temperature of cathode
y—"size" of electron concentration at crossover or screen
S—useful diameter of screen

Definitions

Cathode loading—the current emission density at the cathode surface.
Anodes are numbered outwards from the cathode, No. 1 being nearest the cathode. Electrodes which are concerned with beam generation

and focusing are termed anodes if they are normally positive. American nomenclature sometimes calls our anode No. 1, grid No. 2.

The final anode means the last anode on the *actual electron gun* through which the electrons pass. It may of course also be the first anode.

2. General—Analysis and Synthesis of Systems

It is characteristic of theory that analysis must precede synthesis but in addition analysis must satisfy certain conditions if synthesis is to be achieved.

Consider the design of any scientific instrument. It has certain *operational* characteristics $Z_1, Z_2, Z_3 \ldots Z_n$, and certain constructional parameters and operating conditions $a, b, c, d, e \ldots$. The operational characteristics are functions of the constructional parameters etc. so we may write briefly

$$\begin{aligned} Z_1 &= f_1(abcde \ldots) \\ Z_2 &= f_2(abcde \ldots) \\ Z_3 &= f_3(abcde \ldots) \\ &\ldots\ldots\ldots\ldots \\ Z_n &= f_n(abcde \ldots) \end{aligned} \tag{1}$$

Usually it will be found that n is less than the number of independent variables, indicating that there are some degrees of freedom in the design. On the other hand there will generally be some rather loose implicit relationships between $a, b, c, d, e \ldots$ based on such considerations as strength, durability, proportioning, fabrication, etc. These will limit the possible number of syntheses in design which achieve characteristics $Z_1, Z_2, Z_3 \ldots Z_n$.

However it still remains true that if analysis can establish a system of quantitative relations (1) then these may be used in synthesis. *We note that the designer is not interested in analysis, per se, but only in the inverse reasoning which will be possible if a complete set of analytic relations (1) can be established.*

Unluckily the task of establishing the rigorous complete system (1) usually proves impossible. It has not so far been achieved for cathode ray tubes. Some of the functions $f_1, f_2 \ldots$ are known approximately. By a series of bold approximations the author has set up a complete system (1) which has validity over restricted ranges.[14] This is discussed in Section III, 4. Otherwise the literature has been essentially confined to the analysis of parts of the tube, so attacking the problem of determining some of the individual functions $f_1, f_2 \ldots f_n$. The value of such analysis in guiding synthesis will be greatest when explicit formulae for the Z's can be found. Unfortunately this is often impossible since

analysis in electron optics often involves numerical computation of particular cases and cannot be more general. In that case the nature of the functions f_1, f_2, etc. may be expressed graphically. By constructing functions which agree approximately with these graphs we may then establish an approximate system (1) which can be used in synthesis. This technique has been employed in the author's treatment.[14]

3. *Summary of Literature*

A turning point is the classic paper by D. B. Langmuir[15] published in 1937. This established formally the fundamental principle that the brightness of any image in an electron optical system is limited by the same type of law that limits the brightness of an image in light optics. In the notation of this paper, Langmuir showed that the maximum peak current density ρ_0 in electron concentration was given by

$$\rho_o = \rho_c \left(\frac{Ee}{kT} + 1 \right) \sin^2 \theta \tag{2}$$

The derivation assumes no lens aberrations and freedom from space charge.

In 1939, J. R. Pierce,[16] in a monumental paper re-established this formula by the use of Liouville's theorem, and extended the theory to deal with the problem of average current densities which could be achieved in an image or crossover in terms of the fraction of the total emitted current which could be used. Again space charge was neglected. Both these papers are valuable in that they establish explicit formulae of the type (1) which are far reaching guides in system synthesis.

The first approach to the problem of space charge was also made in 1939 by Headrick and Thompson,[17] who extended earlier work of Watson. Their work investigated the problem of a circular beam of electrons homogeneous in axial velocity emerging from a hole and moving between there and the image in a field-free space. Focusing was allowed for by assuming that on emerging from the hole, each electron had an inward velocity component proportional to its distance from the beam axis, so that in the absence of space charge repulsions the beam would focus at an infinitesimal point. By graphical integration of the equations of motion, Headrick and Thompson established the shape of the resultant beam profile and calculated the minimum radius. This work was later extended by the present author[18] to include the case where the electrons were moving in a uniform externally impressed axial electric field.

These papers on space charge are less useful in synthesis because they present the analysis in the form of curves. The results are embodied in integral equations which do not yield explicit general formulae.

These papers established clearly the existence of two quite distinct limits on the minimum spot size which could be achieved for a given set of operating conditions, and showed furthermore how this minimum limit could be numerically determined. It will be noted however that their analysis treated the effects of initial emission velocities (for that is what is basically implied in eq. (2)) and space charge *separately*. This limitation must be carefully observed, for it is clear that an exact formulation must feed both sets of conditions *simultaneously* into the equations. A discussion of the relative bearing of the two restrictions on general tube design has been given by the present author.[19] Attempts to set up rigorous equations are probably futile in view of the complexity of the problem.

Considerable attention has been given to analysis of the final focusing lens and especially electrostatic lenses. Much work has been published on methods of determining cardinal points, principal foci, etc. of common two element electrostatic lenses.[20] The author feels, however, that although this work is excellent in itself and has required high analytic ability, yet its utility is not great not only because it *is* analysis and not synthesis, but also on account of a second factor. It is well known that the focal length of any common two element lens system is readily varied widely simply by variation of the voltage ratios applied to the lens. What this actual voltage ratio is, does not usually matter much since it merely involves appropriate setting of a voltage divider. It is thus sufficient to let the ratio take care of itself. If the theoretical analysis had concerned itself more with principles of lens design in regard to reduction of aberrations, etc. then the work would have been infinitely more useful.

Some work has been done on the analysis of electron paths in deflector plate systems, perhaps the most notable being that of Hutter.[21] But again we must comment that even this masterly work gives little in the way of clues to improved deflector plate design.

Very little has been published at all on the electron trajectory in the cathode region and even the analysis, let alone the synthesis, of this part of the tube is very sketchy.[21,22,23] This is not surprising in view of the necessity of simultaneously taking into account the effects of emission velocity, lens aberrations, and space charge.

4. A Design Theory of the Electron Gun

Recognizing that an attempt to establish a rigorous system of eq. (1), would result in expressions of impossible complexity, the author deliberately abandoned accuracy, and substituted functions which give approximate values over restricted ranges.[14]

This general technique has a close relation to the method of perturbations. This may be characterized thus; we set up a system S_1 which has characteristic Z_1, the system undergoes fairly small changes δS_1, and we seek to be able to determine the corresponding small change δZ_1. To permit this theory to cover the wide ranges in systems, we make up experimentally a series of systems S_2, S_3 . . . having *measured* characteristics Z_2, Z_3 . . . The experimental systems S_2, S_3 . . . are so spaced that wide extrapolations are not required, and thus the theory and experiment combine to cover the whole range of geometry and characteristic required. Experiment establishes fixed reference points—theory permits interpolation and small extrapolation.

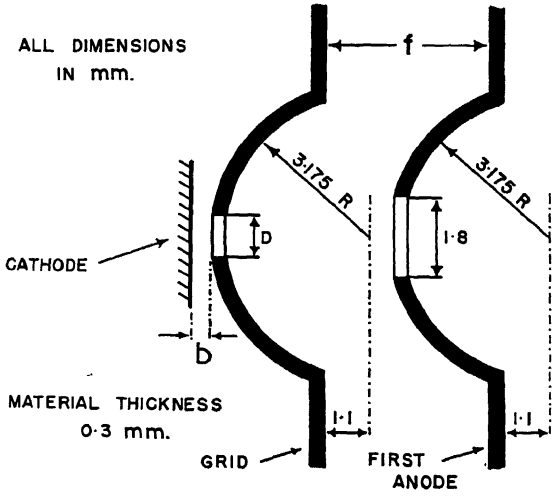


FIG. 12.—Triode geometry.

a. Conditions at the Triode. The following is the briefest summary of the results of the theory. It is given to indicate the nature of the work but there is no space here to discuss the many postulates and limitations which circumscribe the relations set out. The reader wishing to apply these methods *must* refer to the original paper.[14]

With reference to the triode geometry shown in Fig. 12, the following approximate relations hold

$$I_a = 3 \cdot E_d^{\frac{3}{2}} \cdot E_c^{-2} \tag{3}$$

$$E_c = \frac{0.0046 D^3 \cdot E}{t \cdot b \cdot f} \tag{4}$$

$$\rho_c = \frac{0.0145 \cdot E_d^{\frac{3}{2}}}{D^2} \tag{5}$$

$$\sin \alpha = \frac{0.27}{f^{\frac{1}{2}}} \cdot D \cdot \frac{|E_d|}{|E_c|} \qquad |E_d| \leqslant |E_c| \tag{6}$$

$$\frac{I_a}{A} = K_1 \cdot E \cdot \rho_c \cdot \sin^2 \theta \tag{7}$$

$$A = \frac{K_2 \cdot f^{\frac{1}{2}}}{E} \tag{8}$$

E is the potential of the first anode.

In this system of equations, all pressures are measured in volts, currents are in microamperes, current densities in milliamperes/sq. mm., angles in degrees, and all dimensions in millimeters. There is evidence that the same type of equations, but with modified constants, will be found to apply to other triode geometries.

In the paper from which these equations are derived, the results are presented in the form of curves. The constants given above have been derived from the curves as "best fit" values. In eqs. (7) and (8), constants K_1 and K_2 are not evaluated; such evaluation would involve careful definition of the concept "spot size" and has not been attempted.

b. Conditions at the Screen—Liebmann's Theory. Equations (3) to (8) refer to conditions at the triode. If, however, the final focusing lens is aberrationless and intercepts no current, then conditions at the screen will be identical, except for the consequences of a change in spot size due to the magnification of this focusing lens.

Usually the final lens incorporates a beam stop, so that only a fraction of the current given by (3) reaches the screen. The author, in a reply to the discussion on his paper,[14] has given a method of computing this fraction from first principles.

The beam stop in the final lens will affect the influence of aberrations in the *triode* immersion lens on ultimate spot size, since such a stop may prevent marginal rays in the triode from reaching the screen. Such marginal rays naturally contribute most to aberrations. In fact it is only the presence of such stops which justifies the practical application of eqs. (7) and (8). The latter are true for the triode only if the latter is aberrationless, but may be applied to the final focused spot (with a suitable magnification multiplying constant) in systems where the beam is severely trimmed in the final lens. The effect of such trimming is to make the triode *appear* substantially aberration free, through elimination of marginal rays.

Application of the Helmholtz-Lagrange law connects relations at the triode with those at the screen. If the subscript 1 denotes conditions at the triode, and subscript 2 conditions at the screen, then

$$y_2 = \frac{\sqrt{E_1}\, \alpha_1 y_1}{\sqrt{E_2}\, \alpha_2} \tag{9}$$

This relation holds regardless of the number of lenses interposed between the triode and the screen. It is based on the postulates of freedom from

aberrations and paraxial rays. Both these postulates are probably fairly well satisfied in practice. Normally, cathode ray tubes always work with narrow beams and small lens apertures, so that aberrations are greatly reduced.

Application of eq. (8) shows that multiplication of the voltage on the crossover forming electrode by k will multiply the crossover diameter by $1/\sqrt{k}$. Application of eq. (9) then shows that if the final anode potential is also multiplied by k, the final spot diameter y_2 is multiplied by $1/\sqrt{k}$. It is thus concluded that multiplication of all anode potentials by k multiplies the final spot diameter by $1/\sqrt{k}$.*

Now, can we make any deductions about what will happen if *only* the final anode voltage E_2 is multiplied by k, and if the voltage on the crossover forming electrode E_1 is held constant?

From eq. (9) we see that in this event we may deduce only

$$y_2\alpha_2 \cdot \sqrt{E_2} \text{ is constant}$$

The question as to what happens to the final spot size y_2 thus depends entirely on what happens to α_2. If α_2 is constant then clearly

$$y_2 \propto \frac{1}{\sqrt{E_2}} \tag{10}$$

On the other hand, if α_2 is inversely proportional to $\sqrt{E_2}$ the final spot size will be constant. This latter argument led Liebmann[24] to conclude that spot size was independent of final anode voltage. Unfortunately, his paper seemed to leave the impression that this was a general result which always applied. The present author,[25] however, has concluded that it is more usual for result (10) to hold since in his view the operating voltages for normal three anode guns are such that changes in the ratio of the first to second anode voltage consequent on changes in the third (final) anode voltage do not, in fact, cause much change in α_2. Of course the important fact is that $y_2\alpha_2 \cdot \sqrt{E_2}$ is constant and the exact effect of change in final anode potential will depend on the lens conditions.

c. Tabular Summary of Theory—Effect of Triode Perturbations. Tables III, IV, and V summarize the main results of the theory in convenient form. To make the results more general, we leave the values of the indices in eqs. (4) and (6) open by writing them as m and n.

* This result at first sight may appear inconsistent with the fundamental principle of electron optics, which states that in the absence of magnetic fields, multiplication of *all* electrode potentials by a constant factor does not alter the electron trajectory. But this principle assumes that the electrons are injected into the system with zero energies—a condition not satisfied in the triode where the crossover size is a function of the emission energies of the electrons from the cathode.

TABLE III

Operation		Consequences					
Independent variable and operation	*Sole* other variable and extent of adjustment to achieve new condition specified	Cut-off voltage	Crossover diameter	Beam angle (α)	Total current in beam	Current/unit solid angle	Cathode loading
Cathode/grid spacing multiplied by k	Grid material thickness, grid hole diameter and anode/grid spacing all multiplied by k. (Geometrical similitude)	$\times 1$	$\times k$	$\times 1$	$\times 1$	$\times 1$	$\times \frac{1}{k^2}$
	Grid drive multiplied by $\frac{1}{k}$ to maintain constant beam angle	$\times \frac{1}{k}$	$\times 1$	$\times 1$ (postulate)	$\times \frac{1}{k^{3/2}}$	$\times \frac{1}{k^{3/2}}$	$\times \frac{1}{k^{3/2}}$
	Anode/grid spacing multiplied by $\frac{1}{k}$ to maintain constant cut-off voltage	$\times 1$ (postulate)	$\times \frac{1}{k^m}$ $\left(\times \frac{1}{k^{3/4}}\right)$	$\times k^m$ $(\times k^{3/4})$	$\times 1$	$\times \frac{1}{k^{2m}}$ $\left(\times \frac{1}{k^{3/2}}\right)$	$\times 1$
	Grid hole diameter multiplied by $\sqrt[n]{k}$ to maintain constant cut-off voltage	$\times 1$ (postulate)	$\times 1$	$\times k^{1/n}$ $(\times k^{1/2})$	$\times 1$	$\times \frac{1}{k^{2/n}}$ $\left(\times \frac{1}{k^{3/2}}\right)$	$\times \frac{1}{k^{2/n}}$ $\left(\times \frac{1}{k^{3/2}}\right)$

TABLE IV

Operation		Consequences					
Independent variable and operation	*Sole* other variable and extent of adjustment to achieve new condition specified	Cut-off voltage	Crossover diameter	Beam angle (α)	Total current in beam	Current/unit solid angle	Cathode loading
Grid hole diameter multiplied by k	Anode/grid spacing, cathode/grid spacing and grid material thickness all multiplied by k to maintain constant cut-off voltage. (Geometrical similitude)	$\times 1$ (postulate)	$\times k$	$\times 1$	$\times 1$	$\times 1$	$\times \frac{1}{k^2}$
	Grid drive multiplied by k^{n-1} to maintain constant beam angle	$\times k^n$ ($\times k^2$)	$\times 1$	$\times 1$ (postulate)	$\times k^{\frac{3n-7}{2}}$ ($\times k$)	$\times k^{\frac{3n-7}{2}}$ ($\times k$)	$\times k^{\frac{3n-7}{2}}$ ($\times k$)
	Cathode/grid spacing multiplied by k^n to maintain constant cut-off voltage	$\times 1$ (postulate)	$\times 1$	$\times k$	$\times 1$	$\times \frac{1}{k^2}$	$\times \frac{1}{k^2}$
	Anode/grid spacing multiplied by k^n to maintain constant cut-off voltage	$\times 1$ (postulate)	$\times k^{m.n}$ ($\times k^{3/4}$)	$\times k^{1-m.n}$ $\left(\times \frac{1}{k^{3/4}}\right)$	$\times 1$	$\times \frac{1}{k^{2(1-m.n)}}$ ($\times k^{5/2}$)	$\times \frac{1}{k^2}$

TABLE V

Operation		Consequences					
Independent variable and operation	*Sole* other variable and extent of adjustment to achieve new condition specified	Cut-off voltage	Crossover diameter	Beam angle (α)	Total current in beam	Current/unit solid angle	Cathode loading
Anode/grid spacing multiplied by k	Grid hole diameter, cathode/grid spacing, grid material thickness all multiplied by k. (Geometrical similitude)	$\times 1$	$\times k$	$\times 1$	$\times 1$	$\times 1$	$\times \frac{1}{k^2}$
	Cathode/grid spacing, multiplied by $1/k$ to maintain constant cut-off voltage	$\times 1$ (postulate)	$\times k^m$ ($\times k^{3/4}$)	$\times \frac{1}{k^m}$ $\left(\times \frac{1}{k^{3/4}}\right)$	$\times 1$	$\times k^{2m}$ ($\times k^{3/2}$)	$\times 1$
	Grid hole diameter multiplied by $k^{1/n}$ to maintain constant cut-off voltage	$\times 1$ (postulate)	$\times k^m$ ($\times k^{3/4}$)	$\times k^{\frac{1}{n}-m}$ $\left(\times \frac{1}{k^{5/12}}\right)$	$\times 1$ •	$\times \frac{1}{k^{\frac{2}{n}-\frac{2}{m}}}$ ($\times k^{5/6}$)	$\left(\times \frac{1}{k^{2/n}}\right)$ $\times \frac{1}{k^{2/3}}$
	Grid drive multiplied by k^{m-1} to maintain constant beam angle	$\times \frac{1}{k}$	$\times k^m$ ($\times k^{3/4}$)	$\times 1$ (postulate)	$\times k^{\frac{7m-3}{2}}$ ($\times k^{9/8}$)	$\times k^{\frac{7m-3}{2}}$ ($\times k^{9/8}$)	$\times k^{\frac{3(m-1)}{2}}$ $\left(\times \frac{1}{k^{3/8}}\right)$

The results of substituting "best fit values," i.e., $n = 3$ and $m = \frac{3}{4}$, are inserted in brackets. It should be noted that most of these results are approximate only and accuracy on wide extrapolations must not be expected.

d. Examples of Use of Theory in Design Modifications. Problem No. 1. The gun of a television type cathode ray tube consists of a triode only, the focusing and deflection both being magnetic. No beam limiting stop is incorporated. When tested it is found that the central focus quality is satisfactory up to full working brilliance, but that the deflection defocusing is then excessive. Measurement shows that full brilliance requires 300 μa in the beam, and that the limit of tolerable deflection defocusing is reached when the beam current is only 200 μa. How may the triode geometry be altered to meet the requirements indicated for satisfactory operation?

The problem is to reduce the beam angle from the triode, so that the beam width when the beam current is 300 μa is the same as that obtaining in the prototype when the beam current is only 200 μa.

Table III shows that multiplication of the cathode-grid clearance by k, and the grid drive by $1/k$, multiplies the current/unit solid angle from the triode by $1/k^{\frac{3}{4}}$. The current into the maximum solid angle permissible on account of deflection defocusing has to be increased from 200 to 300 μa, so that $200/k^{\frac{3}{4}} = 300$, whence $k = 0.76$. Hence a solution is effected by multiplication of the cathode-grid clearance by 0.76, and the grid drive by $1/0.76 = 1.31$. The cut-off voltage is also multiplied by the latter figure.

Table III shows that the cathode loading is then multiplied by 1.5, as is otherwise apparent since the current/unit area/unit solid angle has been multiplied by 1.5.

Problem No. 2. Suppose that, for reasons of mechanical design, it is not practicable to use the solution to problem 1, and to reduce the cathode-grid clearance. Are any alternative triode changes possible?

Looking through Table IV we notice that an increase in grid hole diameter does not affect the spot size, and that multiplication of the grid hole diameter by k, and the grid drive by k^2 also multiplies the current/unit solid angle by k. The current has to be increased from 200 to 300 μa into the same solid angle, whence $k = \frac{3}{2}$. The grid drive, therefore, has to be multiplied by approximately $\frac{9}{4}$ and the cathode loading is multiplied by $\frac{3}{2}$.

It will be noted that this solution is electrically less satisfactory than the previous one, since the grid drive has to be more than doubled, against a previous increase of only about 30%. Note that the cut-off voltage has been more than trebled.

5. *Use of Scaling and Similitude Theory in Design*

A rigorous system of equations (1) will permit designing *ab initio*. The eqs. (3) to (8) are sufficiently accurate to go a long way towards such "designing on the drawing board," although it is recommended that they are used rather as "correction" equations, as indicated in Tables III, IV, and V.

If, however, one is prepared to abandon any idea of designing without reference to established models, and is prepared to work by modification of existing types, then much simpler methods are possible which do not depend on first setting up the analytic expressions (3) to (8). This is the approach most used in practice, since prototype tubes are always available. It enables theory to restrict itself to evaluation of *relative* rather than *absolute* characteristics.

These methods depend on the application of the fundamental principles of scaling, dimensional homogeneity, and energy conservation, together with a few very simple electron optical laws. Space does not permit discussion of the development of these ideas, given in the original paper,[26] but we will illustrate their application by a very topical example.

a. Application to Geometry of Projection Tube Design. The problem is to investigate the design of a projection tube or tubes which will, after optical projection, yield the same final picture quality (size, definition, brightness, etc.) as the picture on a 15-inch direct viewing prototype tube.

The characteristics of this direct viewing tube are set out in insert 1, Table VI, which also gives the principal *relative* characteristics of two derived tubes.

The first obvious derived tube to investigate is a proportionally scaled-down replica. Insert 2 shows this with a three-to-one linear reduction, which seems reasonable. For practical reasons the base only is not scaled down.

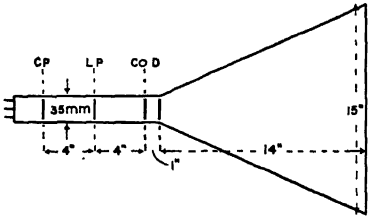
The first step is to decide on the postulates. It will be assumed that the cathode loading remains constant, which is quite reasonable. From this it immediately follows that the beam current in the scaled-down tube can only be $\frac{1}{9}$ of that of the prototype.

Next, in order to insure equal final projected picture brightness, some assumption as to the efficiency of the projection lens system must be made. A figure of 20% is assumed, which can be readily achieved with a Schmidt mirror system. Then, if E_p, I_p are the beam voltage and current respectively in the projection tube, and E and I those of the direct viewing tube, the requirement of equal final picture brightnesses gives

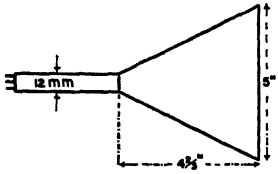
$$0.2E_p^2I_p = E^2I \tag{11}$$

The screen brightness is here assumed to be proportional to the square of the beam voltage, which is a good working approximation. In eq. (11) only E_p is unknown—its value works out at 54 kv.

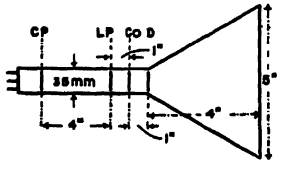
TABLE VI



1. Prototype 15 in. direct viewing tube.	
Final picture illumination	10 e.f.c.
Efficiency of optical system	100 per cent.
Anode voltage	8 kV.
Full white—	
Beam current	150 μA.
Grid drive	E_d
Cathode loading	.35 a/cm^2
Beam power	1.2 watts
Screen loading	1.6 mW/cm^2
Cut-off voltage	E_c
Scanning power	18 watts
Final picture size	12×9.6 ins.



2. One-third scale projection tube.	
Final picture illumination	10 e.f.c.
Efficiency of optical system	20 per cent.
Anode voltage	28.4 kV.
Full white—	
Beam current	60 μA.
Grid drive	$\left(\frac{60}{150}\right)^{2\cdot3} \times E_d$
Cathode loading	.35 a/cm^2
Beam power	1.7 watts
Screen loading	20.6 mW/cm^2
Cut-off voltage	$\left(\frac{60}{150}\right)^{2\cdot1} \times E_c$
Scanning power	6.4 watts
Final picture size	12/9.6 ins.
Grid drive	$\left(\frac{60}{150}\right)^{2\cdot1} \times E_d$



3. Practical projection tube.	
Final picture illumination	10 e.f.c.
Efficiency of optical system	20 per cent.
Anode voltage	13.6 kV.
Full white—	
Beam current	250 μA.
Grid drive	$\left(\frac{250}{150}\right)^{2\cdot3} \times E_d$
Cathode loading	.35 a/cm^2
Beam power	3.4 watts
Screen loading	41.5 mW/cm^2
Cut-off voltage	$\left(\frac{250}{150}\right)^{2\cdot1} \times E_c$
Scanning power	30.5 watts
Final picture size	12×9.6 ins.

C.P. = Crossover plane.
L.P. = Lens plane.
C. of D. = Centre of deflection

Next we investigate the picture definition. By geometrical similitude it immediately follows that if the anode voltage were unchanged, then the definition must also be unchanged. However, the anode voltage has been changed, and from Section III, 4*b* it follows that the new spot size is multiplied by $\sqrt{\frac{8}{54}}$. Hence the scaled-down replica will not give the constant picture quality required.

A solution involves multiplying the scale of the triode by $\sqrt{\frac{5}{8}}$. This then means that cathode loading is only $\frac{8}{5}$ of that on the prototype. This suggests the current could be increased for the same cathode loading, so that the extreme voltage increase might be avoided.

To solve this problem, let k be the linear multiplying factor for the triode portion *only*. The remainder of the tube size is, as before, multiplied by $\frac{1}{3}$.

Inserting the values shown in insert 1 of Table VI, the following equations will be found to obtain.

(1) Energy balance:

$$0.2E_p^2I_p = 8^2 \times 150 \tag{11}$$

(2) Constant cathode loading (postulate):

$$I_p = 150k^2 \tag{12}$$

(3) Constant final spot size:

$$3k = \sqrt{\frac{E_p}{8}} \tag{13}$$

The solution of this system gives $E_p = 28.4$ kv, $I_p = 60$ μa, and $k = 0.63$. Thus, although the tube is reduced to one-third size, the triode portion is reduced to approximately two-thirds of its previous value.

It is now necessary to consider how the solution of eqs. (11), (12), and (13) can be realized in practice. It must be appreciated that the operations performed on the prototype tube described by the solutions given do not automatically enable the derived tube to give the desired performance. This can be seen by the following reasoning.

Suppose that the prototype 15-inch tube operates at a grid bias of $-E_g$ when giving the performance set out in Table VI. The derived projection tube operates at an anode voltage 28.4/8 times as high. In order that the same fraction of the cathode surface shall be employed, it follows that the derived tube must operate at a grid bias of $-\frac{28.4}{8}E_g$. This fact follows from the principle of voltage similitude. It is essential that the same fraction of the cathode surface be employed in order to satisfy the equality of cathode loading condition as determined by eq. (12).

But by the principle of voltage similitude[26] the current in the beam would now be $\left(\frac{28.4}{8}\right)^{\frac{3}{2}} \times 150$ μa, whereas we require the current to be 60 μa in order to satisfy eqs. (11) and (12).

The solution of this problem involves one fact which does not appear to belong to the province of similitude. If any two triodes, identical in all respects except for their cathode to grid spacings, are operated at different first anode potentials, so that their cut-off voltages are identical, then to a very close approximation their modulation characteristics, both in current delivered and in beam angle, are identical. As a corollary, it follows that so far as the current delivered and beam angle are concerned, the effect of multiplying the cut-off voltage by k, by alteration of cathode to grid spacing, is the same as the effect of multiplying the cut-off voltage by k by alteration of first anode potential. This property of the triode is implied in eqs. (3) and (6).

Applying this fact to the problem, the cathode to grid spacing of the derived tube is adjusted so that the cut-off voltage at 28.4 kv becomes $(\frac{60}{150})^{\frac{3}{2}}$ times that of the prototype at 8 kv. The grid drive from cut-off is also multiplied by the same quantity, so that the new beam current is $\frac{60}{150}$ of that in the prototype for the same emitting fraction of the cathode surface. The specification set out in insert 2, Table VI, has now been met, and in such a way as to avoid invalidating eq. (12).

Another solution for a derived tube, due to Mr. J. Sharpe,* is illustrated in insert 3 of Table VI. The solution begins by slicing off the bulb of the direct viewing tube. This is followed by a reduction in separation of the lens plane and the center of deflection, so as to maintain a constant mean beam width in the deflectors. This operation, in view of the constant scanning angle, ensures an unchanged relative picture quality at the edges of the raster.

Following the notation used in the last treatment, we have the following equations to define the operating conditions of the derived tube.

(1) Energy balance:

$$0.2E_p^2 I_p = 8^2 \times 150 \tag{14}$$

(2) Constant cathode loading (postulate):

$$I_p = 150k^2 \tag{15}$$

(3) Constant definition:

$$\frac{4.5}{3 \times \sqrt{8}} = \frac{1.5k}{\sqrt{E_p}} \tag{16}$$

The solution of this system gives $E_p = 13.6$ kv, $k = 1.3$ and $I_p = 250$ μa, as shown in Table VI, insert 3. Again we note that eq. (15) requires that the same fraction of the cathode surface is emitting. This condition

* Late of G. E. C. Research Laboratories, Wembley, England, now at Atomic Energy Research Establishment, Harwell.

is achieved by the method treated in the case of the first derived tube. The cut-off voltage and grid drive are both multiplied by $(\frac{250}{150})^{\frac{3}{4}}$ by alteration of the cathode to grid spacing.

It will be noted that the second derived tube requires less anode voltage to achieve the same performance, and for the same cathode loading. This fact is due to the use of a larger deflector system.

Further examples of these design methods are given in references 27, 28 and 29.

6. Figure of Merit for a Cathode Ray Tube

Although it may be very misleading to quote a figure of merit for any multiapplication device, yet it is perhaps worth while to consider whether some expression can be advanced which will give an over-all picture of the "goodness" of a cathode ray tube for general application. Such considerations may lead to valuable conclusions as to whether all round progress has been effected.

For all electrostatic oscillograph tubes, the following formula is suggested

$$\text{Figure of Merit (F.M.)} = \frac{E_s \times I_s \times \sqrt{XY} \times S \times \eta}{A} \qquad (17)$$

Examination of this formula shows that it makes the F.M. independent of the overall size of the tube. This is reasonable. The precise dependence of the F.M. on screen voltage depends on the way in which the screen efficiency varies with this factor, but it is clear that this formula shows a rapid increase in F.M. with voltage. Again this is reasonable. In comparing the effectiveness of general design of two tubes one would of course assume the same screen potential.

Application of eq. (17) to some modern oscillograph tubes does lead one to conclude that some all-round progress has been made (Section II, 2*a*). However there are many oscillograph tubes on the market especially in America which do not show much improvement over those of ten years ago if the result (17) is accepted as a criterion.

a. Application to Post Accelerator Tubes. It is of interest to examine the application of (17) to post accelerator construction.

A sketch of such a tube is shown in Fig. 13. In this the post accelerator principle has been "idealised" so that such after acceleration takes place without *any* deflection. The screen consists of an electron permeable conducting plate T separated from a second permeable layer L by the dielectric Z (also permeable). The fluorescent screen is in contact

with L, and is assumed to have the same potential. Acceleration takes place across the dielectric Z.

We shall now compare the F.M.'s of this tube when operating as a post accelerator and as a non-post accelerator type. This is done in Table VII. The same screen potential has been assumed so that we avoid discussion of the response of the screen to different voltages. In order to make the conditions of operation of the triode portions of the

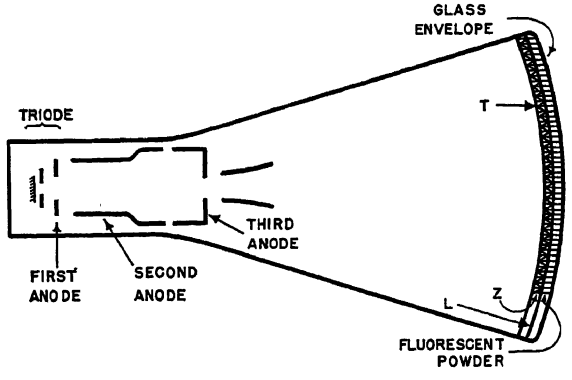


FIG. 13.

gun strictly comparable, the cathode-grid spacing of the tube when operation in the non-post accelerator condition, has been adjusted to make the cut-off voltage constant (see Section III, 4a). This is sufficient to insure constancy of cathode loading, beam current, and beam angle.

TABLE VII

Method of operation	Screen potential	Potential on final anode	Beam current*	Grid drive*	Spot area	Geometric mean of deflections/volt	Figure of merit
Non post accelerator..	E_s	E_s	I_s	E_d	A	$\sqrt{XY}$	$\frac{E_s I_s \sqrt{XY}\, S}{A}$
Post accelerator.....	E_s	E_s/k	I_s	Ed	kA	$k\sqrt{XY}$	$\frac{E_s I_s k \sqrt{XY}\, S}{kA}$

* Kept constant by multiplication of cathode-grid spacing by $1/k$.

Table VII shows the comparison, and indicates that on the definition (17) there is no improvement in employing post acceleration. The

reason is that in this "perfect" post accelerator system, the post accelerating voltage does not contribute to reduction of the spot size, but merely serves to increase the electron speed.

In practical post acceleration tubes the after acceleration is gradual, with a result that there is some decrease in spot size but also some loss in deflection sensitivity. In addition there will be some reduction in effective screen area. It is believed therefore that the analysis based on an "idealized" system does not prejudice the case for post accelerator tubes. Furthermore, in practice, the effects of space charge here neglected must to some extent adversely affect the post accelerator system.[18]

Of course this does not mean that post acceleration construction is not useful in the appropriate circumstances. When high beam voltages and high sensitivities are required, it forms a very convenient engineering method of achieving these features. In particular the problem of voltage insulation is greatly eased since the high potentials are introduced through side contacts which can be widely spaced in the glass envelope. The operation does show, however, that on a reasonable definition of a figure of merit, post acceleration on electrostatically deflected tubes gives no over-all electron optical advantage.

The result of applying definition (17) to magnetically deflected tubes is of some importance. In this case we define X and Y as the deflections along "X" and "Y" axes for a unit *current* in the deflector coils. Following the method of Table VII we readily prove that operation of a magnetically deflected tube with screen potential E_s and final anode potential of $\frac{E_s}{k}$ results in the figure of merit being multiplied by $\frac{1}{\sqrt{k}}$. *Hence on a definition analogous to eq. (17) post acceleration results in a definite degradation in the performance of a magnetically deflected tube.* This arises because with magnetic deflection the deflection/ampere is inversely proportional to the *square root* of the beam potential, whereas with electrostatic deflection the deflection/volt is inversely proportional to the beam potential.

Pierce,[7] reaches the same conclusion by quite different reasoning.

ACKNOWLEDGEMENTS

Individual acknowledgements are almost impossible in a wide review of this type. The author can only therefore express his thanks to many friends and colleagues on both sides of the Atlantic for opportunities afforded and information provided.

REFERENCES

1. Epstein, D. W., and Pensak, L. Improved cathode ray tubes with metal backed luminescent screens. *R.C.A. Rev.*, **7**, 5–10 (1946).

2. Hamaker, H. C., Bruining, H., and Aten, A. H. W. On the activation of oxide coated cathodes. *Philips Research Repts.*, **2**, 171–176 (1947).
3. Johnson, R. E., and Hardy, A. E. Performance characteristics of long persistence cathode ray tube screens: their measurement and control. *R.C.A. Rev.*, **8**, 660–681 (1947).
4. Jesty, L. C., Moss, H., and Puleston, R. Wartime developments in cathode ray tubes for radar. *J. Instn. Elect. Engrs.*, **93**, IIIA, 149–166 (1946).
5. Bradfield, G., Bartlett, J. G., and Watson, D. Stewart. A survey of cathode ray tube problems in service applications with special reference to radar. *J. Instn. Elect. Engrs.*, **93**, IIIA, 128–148 (1946).
6. Lempert, I. E., and Feldt, R. The 5RP multiband tube: an intensifier-type cathode ray tube for high voltage operation. *Proc. Inst. Radio Engrs. N.Y. Wav. Electrons*, **34**, 432–440 (1946).
7. Pierce, J. R. After acceleration and deflection. *Proc. Inst. Radio Engrs. N.Y. Wav. Electrons*, **29**, 28–31 (1941).
8. Moss, H. Cathode ray tube traces. Part 3—Circular time bases. *Electronic Engng.* **17**, 723–729 (1945).
9. Bull, E. W., and Stanley, V. A. Comparison of electrostatic and electromagnetic deflection in cathode ray tubes. *J. Inst. Elect. Engrs.*, **93**, III, 364 (1946).
10. Rinia, H., de Gier, J., van Alphen, P. M. Home projection television. Parts 1, 2, and 3. *Proc. Inst. Radio Engrs. N.Y. Wav. Electrons*, **36**, 395–411 (1946).
11. Starkie, D. Plastic lenses in television projection. *J. Televis. Soc.*, **5**, 86–92 (1947).
12. Garlick, G. F. J., Henderson, S. T., and Puleston, R. Cathode ray tube screens for radar: their development and measurement for intensity modulated displays, especially H_2S. *J. Instn. Elect. Engrs.*, **93**, IIIA, 167–168 (1946).
13. Randall, J. T., Wilkins, M. H. F., and Garlick, G. F. J. Phosphorescence and electron traps. *Proc. Roy. Soc. Lond.*, **184**, 366–433 (1945).
14. Moss, H. The electron gun of the cathode ray tube—Part 2. *J. Brit. Instn. Radio Engrs.*, **6**, 99–129 (1946).
15. Langmuir, D. B. Theoretical limitations in cathode ray tubes. *Proc. Inst. Radio Engrs. N.Y. Wav. Electrons*, **25**, 977–991 (1947).
16. Pierce, J. R. Limiting current densities in electron beams. *J. Appl. Phys.*, **10**, 715–724 (1939).
17. Thompson, B. J., and Headrick, L. B. Space charge limitations on the focus of electron beams. *Proc. Inst. Radio Engrs. N.Y. Wav. Electrons*, **28**, 318–324 (1940).
18. Moss, H. A space charge problem. *Wireless Engr.* **22**, 316–321 (July 1945).
19. Moss, H. The electron gun of the cathode ray tube—part 1. *J. Brit. Instn. Radio Engrs.*, **5**, 10–25 (1945).
20. Spangenberg, Karl R. Vacuum Tubes. McGraw-Hill, New York, 1948, 369–386.
21. Hutter, R. G. E. Electron beam deflection. Parts 1 and 2. *J. Appl. Phys.*, **18**, 740–758 (1947).
22. Klemperer, O., and Mayo, B. J. Electron optics and space charge in simple emission systems with circular symmetry. *J. Instn. Elect. Engrs.*, **95**, III, 131–141 (1948).
23. Jacob, L. A new type of electron optical voltmeter. *J. Instn. Elect. Engrs.*, **91**, III, 512–515 (1944).
24. Liebmann, G. Image formation in cathode ray tubes and the relation of fluorescent spot size and final anode voltage. *Proc. Inst. Radio Engrs. N.Y. Wav. Electrons*, **33**, 381–389 (1945).

25. Discussion on paper No. 24—Liebmann vs. Moss. *Proc. Inst. Radio Engrs. N.Y. Wav. Electrons*, **34**, 580–586 (1946).
26. Moss, H. Simplification of cathode ray tube design by application the theory of similitude. *J. Televis. Soc.*, **4**, 206–219 (1946).
27. Moss, H. Television Cathode Ray Tubes. *Wireless World*, **55**, No. 6, 202–205; No. 7, 261–263 (1949).
28. Moss, H. Letter to the Editor. *Wireless World*, **55**, No. 10, 408–409 (1949).
29. Moss, H. High Power Cathode Ray Tubes. *Wireless Engr.*, **26**, No. 312, 293–296 (1949).

Electron Lenses*

P. GRIVET

University of Paris, Paris

CONTENTS

* Rationalized practical units are employed throughout.

I. Electrostatic Lenses

1. Image Formation in a Thin Lens

a. An Experimental Scheme. First we study a simple experiment, which may be carried out with the apparatus shown in Fig. 1. An electron gun, consisting of a thermionic cathode and a perforated accelerating anode at a potential ϕ_1, sprays electrons onto a thin object (a few millimicrons thick) in the plane of the aperture D_1. The result of this complex action may be explained by assuming that each point of the object emits a narrow pencil of particles of equal velocity, the intensity of these elementary bundles varying from point to point in accordance with the

specimen's fictitious electron transparency (Marton,[1] Gabor,[2] Boersch[3]). The beam of electrons crosses the lens and finally forms an image on the fluorescent screen D_2. This screen remains at a constant potential under electron bombardment because of its secondary emission and the secondary current to the nearby conducting electrode which forms a closed circuit for the beam current. The electrons therefore flow from D_1 to D_2 in a field-free space, with the exception of the short distance between the planes H and H' where they come under the influence of the lens. Suppose in theory that these planes are thin conductive grids made up

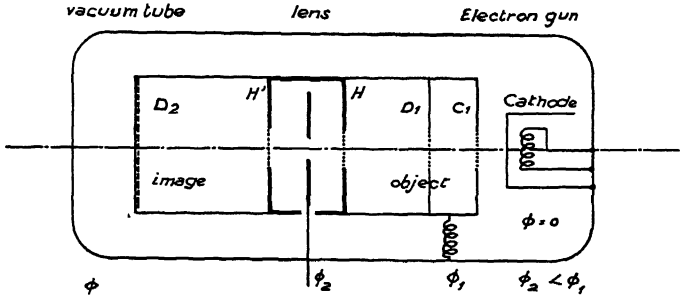


FIG. 1.—Experimental scheme.

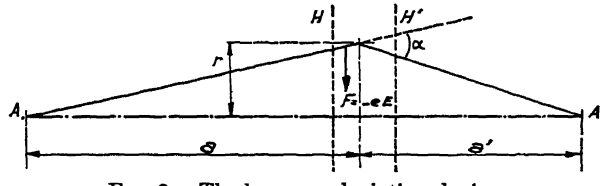


FIG. 2.—The lens as a deviating device.

of extremely fine and closely spaced wires, and that between the two planes H and H' is a disc-shaped electrode at a negative potential with respect to the grids.

b. The Lens Field. We shall now find what essential field properties are necessary to obtain a lens effect. If an image of D_1 is formed on D_2, then all the rays coming out of A, Fig. 2, on the lens axis, converge on A'. Therefore each ray must be deflected toward the axis through an angle α, and the lens must produce an outward-directed radial field, since the force on each electron is $\mathbf{F} = -e\mathbf{E}$. If the ray falls on the lens at a distance r from its center, the angular deflection α will be

$$\alpha = r\left(\frac{1}{a} + \frac{1}{a'}\right) \tag{1}$$

Because the deflection is proportional to the field strength, the radial field intensity must increase with the distance r from the axis according to the law $E_r = K(z)r$, where $K(z)$ is constant or varies with z only. Starting with this relation one may easily prove that α/r has a constant value $1/f$ for any ray, so that

$$\frac{1}{a} + \frac{1}{a'} = \frac{1}{f}.$$

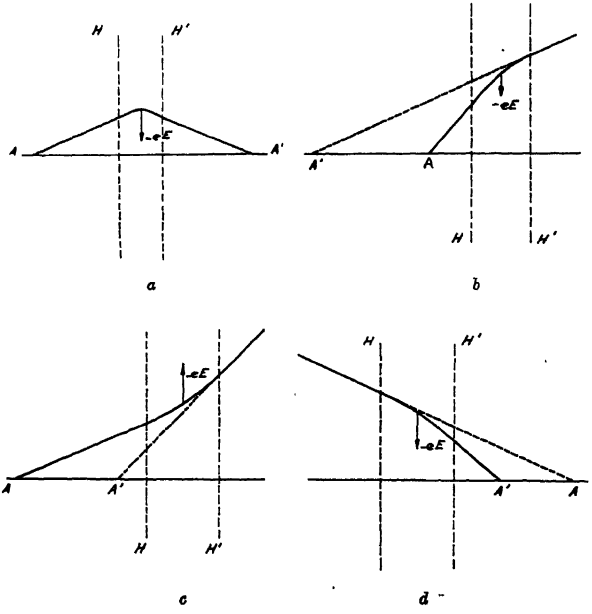


FIG. 3.—*a*. Convergent lens, real image. *b*. Convergent lens, virtual image. *c*. Divergent lens, virtual image. *d*. Convergent lens, virtual object.

A lens effect will be found if the particles are influenced by a radial field whose intensity increases proportionally with the distance from the axis. The rays will converge on a point such as A', if:

1. The force is directed toward the axis.
2. The lens action is sufficiently strong.

When these two conditions are realized simultaneously, we have a convergent lens producing a real image as shown in Fig. 3a. In a lens of this type we can reduce the field to such a low value that it becomes incapable of forcing the rays that issue from a nearby point, A, to converge at a point A', but even in this case, if we extend the emerging rays by drawing straight dotted lines, these "virtual rays" will converge in a point A'. The real rays would then seem to come from A', which would be called the "virtual" image of A (Fig. 3b). The same device permits

us to define an image when the radial force is directed away from the axis. In this case, the lens is said to be divergent and the image of a real object is virtual in every case (Fig. 3c). A further extension of the same idea gives us also "virtual" objects, as can be seen from Fig. 3d. Such a virtual object is produced if a second lens is placed between the position of the first lens and the position of the (real) image projected by the first lens alone.

Thus an image exists for every lens field strength and electron velocity (within certain limits beyond which the lens behaves like a mirror); these two factors govern the lens effect or, expressed in optical terms, the lens convergence. The slower the electrons and the stronger the radial field, the greater will be the electron deflection.

c. The Inseparability of the Radial and Longitudinal Field Components. Electrodes such as those in Fig. 1 give a satisfactory field shape, but the field is not a simple one since each radial component is combined with a longitudinal one directed parallel to the axis.

From the field map of Fig. 4 we see that the radial component will not be important except in regions where the curvature of the lines of force is great and where, therefore, the amplitude of the longitudinal component varies considerably over a short distance. By means of electric field laws we may establish a relation for near-the-axis points: The flux of electrical force across the surface of a small cylinder centered on the axis is zero, and omitting the third and higher order terms we find for a field with rotational symmetry

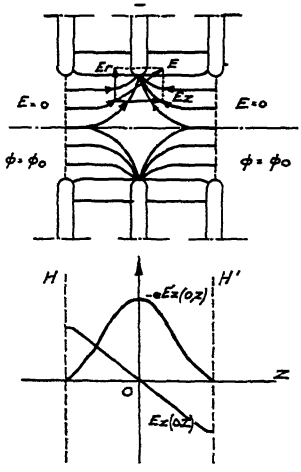


FIG. 4.—Field map in a grid lens.

$$E_r = -\frac{1}{2} r \frac{\partial E_z}{\partial z} \tag{2}$$

which gives the radial field component in terms of the field gradient along the axis. The intensity of E_r varies with r, as required, and the previously introduced function $K(z)$ appears to be proportional to the gradient of the field intensity along the axis. The lens thickness is then established as that thickness that causes the E_z derivative to have an appreciable value.

The longitudinal component slows down or accelerates the electrons and, by influencing the deflection of the electron beam by the radial field, changes somewhat the magnitude of the convergence.

d. Ordinary Gridless Lenses. The effect of the longitudinal component is much more important in the case of a gridless lens—the most widely used type.

In addition to being difficult to construct with satisfactory transparency, a grid is undesirable because:

1. It is a source of secondary electrons and therefore a source of haze in the image.

2. It is the cause of a weak but definite image blurring. The present theory supposes that H and H' are plane conductors, i.e., the lines of force terminate perpendicularly at all points in the plane. In reality, when H and H' are grids, each line of force ends at the grid in a pronounced bend (Fig. 5). The influence of the bend may be evaluated by comparing the variation of potential along it with the potential difference between the grids and the center electrode. Although this effect is not very large, of the same importance as the pitch of the grid, it would be objectionable in most applications.

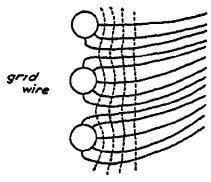


FIG. 5.—Distortion of the field in the vicinity of a grid.

In this section the previously described grid lenses are transformed into a three-electrode single lens, with two circular holes in the two plane electrodes taking the place of the grids. In this arrangement the longitudinal component becomes the most important. This will be more clearly understood if we consider first the case of the symmetrical two-electrode lens, shown in Fig. 6.

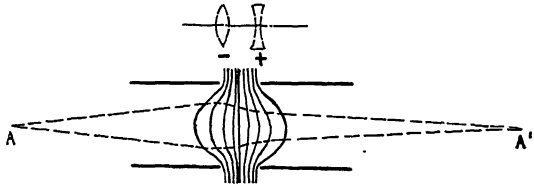


FIG. 6.—The symmetrical two cylinder, two potential lens (+ and − affect the charges on the electrodes).

The radial field reverses its direction in crossing from the first lens half to the second half, and, if the radial field were acting alone, the total effect would be zero. But the longitudinal component accelerates the electrons as they cross the lens and they thus move more slowly in the first half of the lens than in the second. Therefore, the convergence is increased in the first lens half. At the same time, the divergence decreases in the second lens half, and, as a whole, the lens is convergent.

An analogous effect also occurs in a three-element lens, as can be seen from Fig. 7 which shows a convergent element between two divergent elements.

The influence of the longitudinal component explains many peculiarities of ordinary electron lenses:

1. They are always convergent. The lines of force begin positively on one electrode, terminate negatively on the other, and are convex to

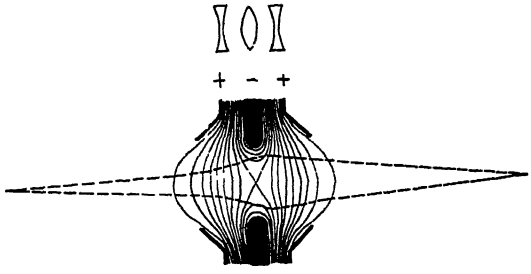


FIG. 7.—The independent three diaphragm lens.

the axis because of their mutual repulsion. Thus a convergent region is always accompanied by a divergent region, with the electrons traveling more slowly in the convergent region than in the divergent region, so that the converging action outweighs the diverging action. In the special case of a plane cathode adjacent to an accelerating hole, this rule breaks down. This is the only known gridless divergent lens. A three-electrode gridless lens is convergent even if the central aperture is at a positive potential.

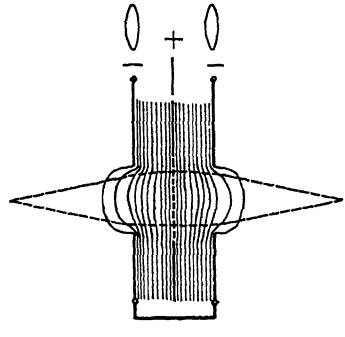


FIG. 8.—The independent three diaphragm grid lens.

2. In an ordinary lens it is always necessary to use a potential comparable to the beam accelerating voltage in order to obtain a nonnegligible convergence by the differential mechanism. In a grid lens the action of the field is uniform throughout and we could obtain the same convergence by using only one tenth of the potential required when no grids are used (Fig. 8 shows a simple model design for the one-grid lens which possesses the same optical properties as the two-grid model previously considered).

e. Electron Mirror. The Minimum Focal Length. Let us consider a lens consisting of rotationally symmetrical electrodes and having the

field distribution (in a meridional plane) represented by the equipotential lines $\Phi(r,z) = \text{constant}$. By using the potential $\Phi(r,z)$ we obtain for the electron velocity

$$v = \sqrt{\frac{2e}{m}}\sqrt{\Phi} \tag{3}$$

where e is the electronic charge and m is the electronic mass, assuming the zero potential to be at the cathode where the electrons are released at negligible speed.

These fundamental relations immediately demonstrate the possibility of transforming an electron lens into an electron mirror. Let us apply to the central electrode of the single lens (Fig. 7) a potential that is negative not only with respect to the nearby electrode but also with respect to the cathode. The field thus produced will still act as a lens as long as the point at the lens center ($r = 0$, $z = 0$) is maintained at a positive potential ($\Phi(0,0) > 0$).

When the center potential becomes negative the electrons are slowed down and turned back before reaching center of the lens, and we have a mirror working on the same principle as an optical mirror. In transitional cases where $\Phi(0,0)$ is only slightly negative the behavior is more complicated and requires special treatment (Recknagel[4]).

It should be noted that mirrors of this type are convergent only when the repelling electrode is not too strongly negative, permitting the electrons to enter the lens deeply and be influenced by the converging fields before they are reflected. When the electrode is very strongly negative the electrons are repulsed before they can be influenced by the convergent fields and the mirror is divergent. The problems arising when $\Phi(0,0)$ is small make it difficult to obtain a focal length smaller than a certain limit which is of the same order of magnitude as the radius of the electrode apertures in the most favorable case when the electrodes are separated by that same distance. This minimum focal length is approximately the same for the mirror and for the lens.

f. The Lens as a Set of Prisms. Similitude. An attack based on the separation of an electron lens into a set of thin prisms is more artificial than our preceding one, but the analogy leads to suggestive results. The elementary electronic prism is the classical condenser as it is used to deflect the electron beam, in cathode ray oscillographs and as such it gives us an explanation of the action of the radial component E_r. Each elementary volume of the field may be represented by an elementary condenser (distance d between the plates, distance l for the length of the plates parallel to the axis, acting difference of potential $E_r d = u$, and mean value of the potential within the element Φ) (Fig. 9), deflecting the

ray through the angle

$$\alpha = \frac{1}{2}\frac{u}{\Phi}\frac{l}{d} \tag{4}$$

This shows clearly that in either a thick or a thin lens:

1. The path of the rays is independent of the particle constants e and m since neither the charge nor the mass appears in the equation. Thus the focusing property of the lens would be unchanged if the source were to emit ions instead of electrons. This property opens up the possibility of ion focusing, which is of wide use in particle accelerators and which has been brilliantly introduced into the domain of mass spectrography by Cartan[5] in France, and into the domain of microscopy by Boersch[6] in Germany, but which we will not take up here. However, for tubes possessing oxide-coated emitting cathodes or luminescent layers there is always, with electrostatic focusing, the danger of excessive ion bombardment which might injure or destroy these delicate coatings (Broadway and Pearce[7]; Schaefer and Walcher[8]).

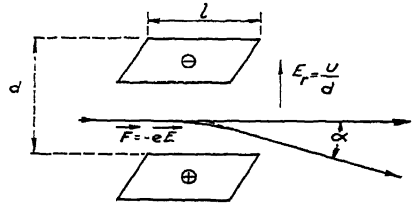


FIG. 9.—The electrostatic prism.

2. The path is independent of the absolute dimensions of the lens.

3. The path is influenced only by the ratios Φ_1/Φ_0, Φ_2/Φ_0, etc. of the electrode voltages to the accelerating voltage Φ_0. The operation of an electrostatic lens will not be altered by small variations in these applied voltages, which may even alternate provided that their mutual ratio remains the same. It appears possible to connect the electrodes to a potentiometer fed from a single voltage source, but time lag effects resulting from a rapidly fluctuating voltage source makes this method difficult to achieve. This possibility has been experimentally checked and has been utilized in some cruder electron optical applications such as high-voltage X-ray tubes. However, in more highly refined applications, such as the electron microscope, it can not be used due to the relativity correction necessary at high electron speeds. Equation (4) thus becomes

$$\alpha = \frac{1}{2}\frac{u}{\Phi}\frac{l}{d}\left(1 + \frac{e\Phi}{2mc^2}\right) \tag{5}$$

where c is the velocity of light, and the linearity in eq. (4) is lost, since the effect of $e\Phi/2mc^2$, which equals $0.977 \times 10^{-3}\Phi$ (where Φ is in kilovolts), can not be neglected (Ramberg[9]).

2. Gaussian Approximation

a. Gaussian Conditions. The foregoing considerations show that the necessary conditions for good image formation exist only in a tubular-shaped space around the axis. In this case, when describing the ray paths as given by the electric field laws, we can neglect the terms in higher powers of r and dr/dz. This is a first order approximation of the ray characteristics and is known as "Gaussian Optics."

The first order terms of the power series for E_r and E_z are given by eq. (2) which we write in a more convenient notation

$$E_z = -\varphi'(z) \qquad E_r = \tfrac{1}{2} r\varphi''(z) \tag{6}$$

denoting by $\varphi(z)$, $\varphi'(z)$, $\varphi''(z)$ the axial potential value $\Phi(0,z)$ and its derivatives.

b. Equations of Motion. We shall now attempt the precise calculation of the electron trajectories within the limits of the Gaussian approximation. Applying Newton's law in the r direction, we get (since the electron charge is $-e$)

$$m\frac{d^2r}{dt^2} = -\frac{1}{2} er\varphi'' \tag{7}$$

For the z-axis we may write down at once, using the kinetic energy law, a first order differential equation

$$\frac{1}{2} m\left(\frac{dz}{dt}\right)^2 = e\varphi \tag{8}$$

The two equations (7 and 8) allow us to determine the paths of those rays which remain in a meridional plane: this is the case for all rays when the object point is on the axis, for some of them when it lies off the axis. Eliminating the time parameter from the two equations, we have

$$\sqrt{\varphi}\frac{d}{dz}\left(\sqrt{\varphi}\frac{dr}{dz}\right) = -\frac{1}{4} r\varphi'' \tag{9}$$

or

$$\frac{d^2r}{dz^2} + \frac{\varphi'}{2\varphi}\frac{dr}{dz} + \frac{\varphi''}{4\varphi} r = 0 \tag{10}$$

We need not worry about the lost skew trajectories as a simple mathematical trick brings them under the control of eq. (9) or (10). Having chosen two axes $0x$ and $0y$ in a plane perpendicular to $0z$, we can describe the quantities depending linearly on x and y in condensed form by using the imaginary symbol j as follows: $u = x + jy$

$$\frac{du}{dt} = \frac{dx}{dt} + j\frac{dy}{dt} \qquad \frac{d^2u}{dt^2} = \frac{d^2x}{dt^2} + j\frac{d^2y}{dt^2} \tag{11}$$

$$E_u = E_x + jE_y = -\frac{1}{2}u\frac{\partial E_z}{\partial z}$$

The equation which defines u and z in space, defines r and z in the meridional plane. Replacing r by u and interpreting the results in terms of eq. (11) we find that the skew trajectories exhibit the same optical properties as the meridional trajectories.

c. Similitude. Homogeneity. The similitude properties may be seen at once by observing the homogeneous equations of zero degree as regards φ and in which e and m do not appear.

The loss of homogeneity, caused by relativity at high electron speeds will be dealt with later.

d. Cathode Immersion Lenses. When the object is itself a thermionic, photoelectric, or secondary source of electrons a new difficulty arises; the conditions required for a Gaussian approximation are no longer fulfilled since the electrons from the cathode escape in all directions. This, does no harm however, because initial velocities are small, corresponding to a few tenths of an electron volt. On the other hand the accelerating voltages are of the order of one to several kilovolts, and the field is initially normal to the cathode (that is, parallel to the axis), nearly uniform, and very strong. Under these conditions the start of the trajectory is an arc of a parabola, and as we proceed along the curve the velocity very soon changes its direction so that it is nearly parallel to the axis when we come to some plane, $z = z_1$, a few tenths of a millimeter away from the cathode. In this plane, rays originating at A in the cathode (Fig. 10) now seem to travel along straight paths diverging from a small spot around a point A_1 located at $z = -z_1$; the radius of the spot is $\rho = 2\varphi_c/E_c$, where φ_c is the voltage equivalent to the most probable initial velocity of the electrons ($\varphi_c = 11600/T$ where T is the absolute temperature of the cathode) and E_c the field at the cathode (Recknagel[10]). The radius of the spot is a negligible quantity ranging ordinarily between 2 and 0.02 μ; the angular apertures of the beam corresponding to each point of the cathode are correspondingly small ($2\alpha_1 = 2\sqrt{\varphi_c/E_c z_1}$). What remains of the lens effect farther than $z = z_1$ from the cathode needs no special attention since the conditions there are truly Gaussian.

Experiments confirm this theory, and emission electron microscopes working at 30 kv give images surpassing the best light microscopes in sharpness (Mecklenburg[11]; Kinder[12]).

e. Electron Mirrors. The standard reasoning seems also to fail in the case of electron mirrors.

Indeed, in the regions where the electrons are reflected, the motion is toward the axis, and the quantity dr/dz takes on a finite value first, then an infinite value at the vertex S of the trajectory. Since we treated dr/dz as a small quantity when establishing the ray equation (10), it is no longer evident that the curve C, which is defined by this equation and the initial conditions, is the real trajectory (T) to a first order approximation. That this is indeed still true may be shown by discussion of the terms in r^3 in the neighborhood of the return point. The use of mirrors in the electron microscope (Mahl[13]) has shown that the quality of the images is comparable to that obtained with lenses.

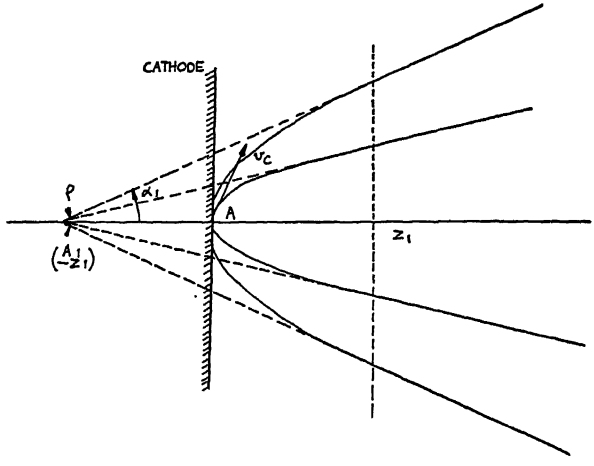


FIG. 10.—Rays in a cathode lens.

A more difficult analysis is required in the case where the return point is at the center of a symmetrical lens (Recknagel[4]); but this case is of small interest in optics since the resulting images are quite poor.

3. *Optical Formalism*

a. Image Formation. In the general case, for any lens thickness and convergence, the ability of the lens to form an image can be deduced from the linearity of eq. (10). If we know the paths of two independent rays, e.g., $r_1 = X(z)$ and $r_2 = Y(z)$ (whose equations are not connected by any linear relation) we can deduce the paths of all other rays, because, for a linear grouping of these two, we have

$$r = \lambda X(z) + \mu Y(z) \tag{12}$$

where λ and μ are arbitrary constants.

The following results may be deduced from relation (12) by simple calculations.

1. All the rays originating at a given object point A generally converge at a "real image point A'" after crossing the lens. If they fail to converge we can extend the real rays emerging from the lens backward by straight lines (dotted lines in the figures to conform to the ordinary convention of glass optics); these rays obey a linear relation similar to (12) and converge in a "virtual image" point. Object point and image point are called "conjugate points." The conjugate points of an infinitely distant point on the axis, left or right, are said to be the two "foci" of the lens.

2. A figure located in a plane perpendicular to the axis is imaged in a conjugate plane perpendicular to the axis (frontal plane), and the image is similar to the object.

3. The linear magnification of the image is a fixed quantity for any pair of conjugate planes. The angular magnification is defined as the ratio of the angular aperture of an elementary pencil at the image to the angular aperture of the corresponding pencil at the object; the angular magnification is also fixed for a pair of conjugate planes.

b. The Classical Cardinal Points. In classical electron optics we assume that the object and the image are in a field-free region where the rays are straight lnes. This is a very valuable physical approximation because at a distance of a few lens diameters from the center the field has disappeared physically although it exists mathematically. In general in electron optics, lenses have a well defined finite thickness within which E (and later H^2) have an appreciable effect on the electrons.

Mathematically speaking, classical electron optics does not deal with the trajectories themselves, but only with the straight "asymptotes" of the rays. The classical method leaves us, therefore, without any information on what happens to the curved portions of the trajectories within the lens. It does not deal with any real image or object located within the lens, since in this region we consider only virtual objects and images given by straight dotted-line extensions of the real rays.

Except for this gap, the classical cardinal points offer a highly efficient method of representing the optical properties of a system of lenses; moreover, it should be noted that virtual images are of interest in themselves. For instance, the electron microscope objective forms an intermediate image which is ordinarily located inside the projector lens (see Fig. 13) and plays the role of a virtual object.

Now let us consider a ray $Y(z)$ parallel to the axis in the object space. In the image space it intersects the axis at the image focus. The equation of another ray, $U(z)$, parallel to the axis in the object space in $r = U(z) = \mu Y(z)$.

It follows that we have $U(z_0)/U'(z_i) =$ constant, whatever may be

the choice of μ or U. In optical notation this is represented as $h_0/\alpha_i = f_i$ which defines the focal length f_i on the image side (Fig. 11). The geometrical interpretation of these equations is that if we extend the initial and final parts of the U-rays they intersect in a fixed front plane, which is called the principal plane on the image side. The rays issuing from the object focus and, emerging parallel to the axis ($V(z)$), allow us to define the principal plane and the focal length on the object side in a similiar manner.

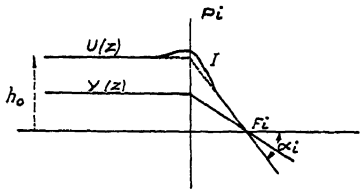


FIG. 11.—Principal plane and focus.

The two principal plane positions and the two focal lengths being known, the optical properties are completely defined. Given an incident ray we can obtain any emerging ray quite simply and we can easily determine the location and magnitude of the image for any object, using simple geometrical considerations and formulae, which may be found in elementary books on optics.

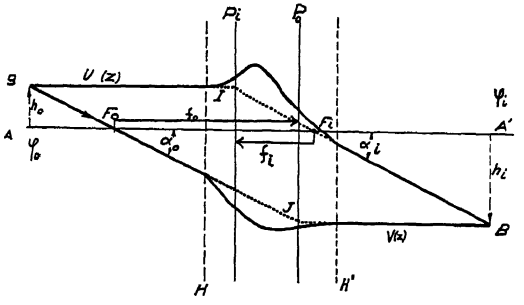


FIG. 12.—Cardinal elements.

c. Lagrange-Helmholtz Law. This law establishes a relation between the two focal lengths:

$$\frac{f_i}{f_0} = -\frac{\sqrt{\varphi_i}}{\sqrt{\varphi_0}} \tag{13}$$

To derive it we establish a relation between the two formulae of rays $U(z)$ and $V(z)$. Combining their differential equations symmetrically to eliminate the r term and integrating, we get

$$U\frac{dV}{dz}\sqrt{\varphi} - V\frac{dU}{dz}\sqrt{\varphi} = \text{a constant}$$

In optical notations (Fig. 12)

$$\sqrt{\varphi_0}\, h_0 \tan \alpha_0 = \sqrt{\varphi_i}\, h_i \tan \alpha_i \tag{14}$$

This may be read: the product of the angular magnification (α_i/α_0) and the linear magnification h_i/h_0 is constant and equal to $\sqrt{\varphi_0}/\sqrt{\varphi_i}$. In terms of cardinal points we finally obtain (13), requiring only two values of φ to characterize the object space and the image space.

d. Objects or Images Immersed in the Lens Field. In glass optics the curved parts of the rays are of small interest since they reduce to sharp breaks at the boundary between the two media; the classical cardinal points alone are sufficient. In electron optics, on the other hand, it often happens that the lens is thick and convergent enough to form an image inside the lens, which still has the properties described in Section I, 3, *a*. For example, the symmetrical single lens may be strong enough to refract an incident ray parallel to the axis and make it cross the axis within the lens, at a point G_i which may be considered as a real focus (Fig. 13). This class of objects and images is of no practical interest in the case of electrostatic lenses where the field would be badly distorted by the presence of the specimen or of the screen or plate so that an image like A_1B_1 (Fig. 14) would not be observable. In a magnetic objective the magnetic field is left unchanged by the presence of nonmagnetic

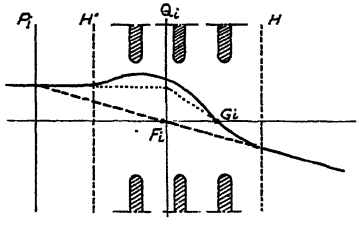


FIG. 13.—Object immersed in an electrostatic lens.

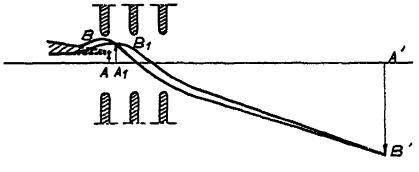


FIG. 14.—Object immersed in a magnetic objective.

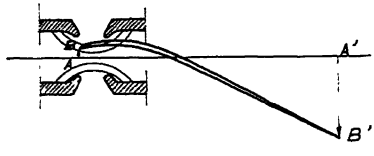


FIG. 15.—Glaser's[14] cardinal elements compared to the classical ones.

materials and an object position such as AB (Fig. 15) would be perfectly usable.

As long as we considered only the straight parts of the rays the procedure was simple because we had to deal only with a linear association of straight lines which cut the axis at mathematically defined points. But, in general, no method of this kind applies when we are dealing with the curved parts themselves, and the real and immersed images (or objects) of the previous paragraph are not joined with their corresponding objects (or images) by any simple relation. Consequently, if we deal with the images really formed within the lens we must resort to direct

calculation of two rays from the differential equation. The Lagrange-Helmholz law (14) remains valid, but φ_0 and φ_i are now characteristic of the particular pair of conjugate points considered.

Glaser,[14] Glaser and Lammel,[15] and Hutter[16] discovered a class of fields for which relations of the classical form hold, even for immersed objects or images. Such a field can be used as a good approximation to the ordinary magnetic lens. The relation is valid, whether the points considered are inside or outside the lens, and the magnification is given by Newton's laws as a function of the image position. These relations can be represented geometrically by an ordinary construction if we define G_i and G_0, the two real points where a ray, initially parallel to the axis, first effectively crosses the axis as foci, and as principal planes,

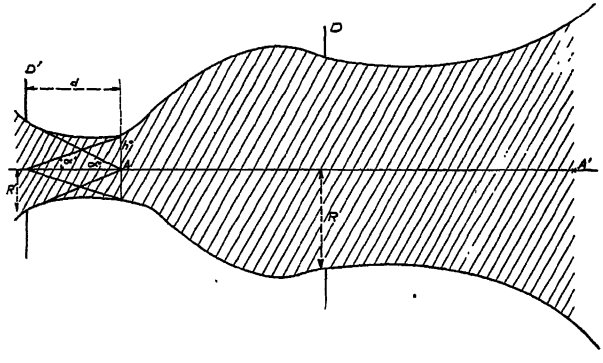


FIG. 16.—Importance of aperture diaphragm D and of the entrance pupil D′ in photometry.

Q_i and Q_0, which are the planes defined by $G_0Q_0 = g_0$ and $G_iQ_i = g_i$, with the focal lengths $g_0 = r(+\infty)/r'(G_0)$ and $g_i = r(-\infty)/r'(G_i)$ (Fig. 13).

But it should be pointed out that the straight lines on the drawing are no longer related to the real rays or to the classical points F_i, F_0 and P_i, P_0. The Lagrange-Helmholtz law (eq. 14) can no longer be converted to the form of (eq. 13).

e. Photometry. In developing electron photometry we can start from the Lagrange-Helmholtz law in the form of eq. (14) (the constant angular magnification applies to skew rays as well as to the meridional rays). A more suggestive relation may be found by taking the area covered in a frontal plane by the entire beam. This surface may be considered as an aperture (of radius R) in a plane. Introducing the image D' (of radius R') of D through the part of the system which precedes D, the rays between D' and the object are straight lines and the value of α_0 at the object is $\alpha_0 = R'/d$, d being the distance from the object to D' (Fig. 16). The first term in the L-H law is h_0R'/d and it

may be read as the product of R' and the aperture $\alpha' = h_0/d$ of an elementary pencil in the plane of D'. Thus: $h_0\alpha_0\sqrt{\varphi_0} = R'\alpha'\sqrt{\varphi_0}$. Applying the L-H law to planes D and D' we get $R'\alpha'\sqrt{\varphi_0} = R\alpha_D\sqrt{\varphi_D}$, since D' is in the object space. Therefore,

$$h_0\alpha_0\sqrt{\varphi_0} = R\alpha_D\sqrt{\varphi_D} \tag{15}$$

and the product $R\alpha_D\sqrt{\varphi_D}$ (R radius of the beam's section, α_D angular aperture of the elementary pencil in plane D) is a constant whatever may be the choice of the plane D. The most interesting application of formula (15) is to electron sources. Calling the effective radius of the cathode r_c, and taking into account the peculiarities of cathode lenses as previously explained ($h_1 = r_c$; $\alpha_1 = \sqrt{\varphi_c/E_c z_1}$; $\varphi_1 = E_c z_1$) and taking $\varphi_c = T/11.600$, one obtains

$$r_c\sqrt{T/11.600} = R_s\alpha_s\sqrt{\varphi_s} \tag{16}$$

which gives the minimum possible radius R_s of the spot in an oscillograph tube working with a final voltage of φ_s and a beam of final aperture α_s. In terms of the specific emission i_c and current density i_s on the screen ($I_s = \pi r_c^2 i_c = \pi R_s^2 i_s$; I_s is the total current) we have

$$i_s/i_c = \left(\frac{11.600\ \varphi_s}{T}\right)\alpha_s^2 \tag{17}$$

a famous formula due to Langmuir[17] and perfected by Pierce.[18,19]

Since r_c and α_s are related, formula (17) may still be simplified. First considering the image focal plane of the cathode lens, where the first cross-over lies, we have

$$\alpha_f = r_c/f \qquad r_f = f\sqrt{T/11.600\ \varphi_f} \tag{18}$$

f is the image focal distance, r_f is the cross-over radius, α_f is the pencil aperture at the focus, φ_f is the potential.

If the second lens splits the distance between the cross-over and the screen into two parts whose quotient is m, we have

$$R_s = mf\sqrt{T/11.600\ \varphi_s} \tag{19}$$

This expression is valid as long as r_c and α_s have such a small value that the spherical aberration is negligible in the cathode lens and in the second lens (if this is not the case the L-H law would lose its simple form). These conditions may be partially satisfied by suitable cathode construction or by limiting the beam with a stop in the second lens.

Going further into the details, as did Law,[20] first we take into account the Maxwellian distribution of the particle speed, and get the current

distribution in the spot. In the image focal plane of the cathode lens (cross-over) we have

$$i_s = (I_s/\pi m^2 f^2) \cdot (11.600\ \varphi_s/T) \cdot \exp\left[(11.600\ \varphi_s/T)(-r^2/mf^2)\right] \quad (20)$$

This result has been thoroughly checked by experiment, with the one exception of the value of the absolute magnitude of f which depends on the space charge correction (Dosse[21] and Jacob[22]). The current density i_c is found to reach saturation value in some electron guns employing a tungsten hair-pin filament; the influence of space charge on oxide cathodes is so strong that emission depends upon the degree of vacuum in the sealed tubes (Klemperer and Mayo[23]).

Spot size reduction by "post acceleration" is another example of the importance of geometrical photometry; Pierce[24] has given this problem an exhaustive treatment.

4. Properties and Determination of the Cardinal Points

a. Reduced Equations. In attacking the integration of eq. (10) to find the two fundamental rays it seems necessary at first to measure the $\varphi(z)$ values with considerable precision, if we are to deduce from them the φ'' values appearing in eq. (10) without undue errors. However, this is generally unnecessary. Introducing a new function R, where

$$r = R\varphi^{-\frac{1}{4}} \quad (21)$$

the equation becomes

$$\frac{d^2R}{dz^2} + \frac{3}{16}\left(\frac{\varphi'}{\varphi}\right)^2 R = 0 \quad (22)$$

Where φ'' has disappeared and we need only φ' and φ. The value of φ' may be deduced from φ by an interpolation formula. φ'' remains useful only where the transformation (21) is unworkable, near the cathode or near a reflection point where φ approaches zero. In these cases, however, the real field can be efficiently approached by a simple uniform field where $\varphi'' = 0$, so that the trajectory is a parabola, and no real difficulty appears.

The curve $R(z)$ can be used directly to fix the position of the focus, since $R(z)$ and $r(z)$ have the same roots. For the principal planes we use the formula

$$f_i = r(-\infty)/r'(+\infty) = [R(-\infty)/R'(+\infty)]\varphi_0^{-\frac{1}{4}} \cdot \varphi_i^{\frac{1}{4}} \quad (23)$$

which often reduces to: $f_i = R(-\infty)/R'(+\infty)$.

b. Peculiarities of Electron Optical Systems. The positive sign of the coefficients in (22) makes it easy to prove 1) that no divergent lens

exists between two regions at the same potential, and 2) that we cross the image principal plane before the object principal plane when proceeding from the object space to the image space. This property holds for both electrostatic and magnetic lenses.

c. Calculation of the Two Fundamental Rays. Adams-Störmer's Method. In most cases, to find the path of the electron we must resort to some method of direct numerical computation of the integral. The most efficient method was originally devised by the British astronomer Adams in his research on the theory of capillary action (1883). Störmer in 1907, adapted it to the computation of electron trajectories between the sun and our magnetically acting earth.

In ordinary applications the only difficulty presented by the Adams-Störmer method is that it requires a series of points (3 or more) to begin with in order to calculate other trajectory points, and these starting points are sometimes difficult to calculate from the ordinary initial conditions (one point and the initial speed). But in our case this difficulty disappears since we know that the ray is a straight line before entering the lens. The reader can refer to Störmer[25] himself for a clear presentation of the method and its advantages in the case of electron trajectories, and also to Kryloff[26] and von Mises.[27] Automatic ray tracing, using an electrolytic tank, was unsuccessful in the field of Gaussian optics because the tank must work with a model geometircally similar to the original, and unfortunately this preserves the angle values. The machine must then deal with angles as small as the actual electron trajectory angles, and the cardinal points are thus inaccurately determined by the intersections of nearly parallel lines.

5. Optical Analogy

a. Descartes-Snells' Law. Equation (14) becomes identified with the optical Lagrange-Helmholtz law if we write $\sqrt{\varphi_0} = n_0$; $\sqrt{\varphi_i} = n_i$. In this particular case we recognize the fundamental analogy between mechanical particles and dioptrics, which can be explained as follows: the electron path inside an electrostatic lens is identified with that of a light ray starting from the same source point and proceeding through a transparent space of variable index where

$$n = \sqrt{\Phi(r,z)} \tag{24}$$

and the cathode potential is taken as zero. ($n' = Kn$ would do as well and often one chooses $K^2 = e/mc^2$ to get a dimensionless index $n' = v/c$, where c is the velocity of light in vacuo.)

We can put this analogy on a firmer basis by explaining the elementary deviation along an element of path as a refraction at the equipo-

tential curves, with a constant index between them. Geometrical electron optics may be developed from this, as is done in most elementary expositions on glass optics. The base relation (10) appears in the guise of the light ray equation in inhomogeneous media, applied in the theory of the mirage and in the diffraction of light by ultrasonic waves. In electron optics it is adapted to the new case where the repartition of n presents axial symmetry.

b. Fermat's Principle. The optical analogy can be explained more concisely by introducing the principle of least action from mechanics and comparing it to Fermat's principle in optics. With the help of eq. (24) they may be merged in a unique formulation that may be accepted in the most general way as governing all electrostatic geometrical optics. However it should be remarked that some electronic peculiarities are still noticeable.

1. The values for n_i/n_0 vary over a much greater range than in glass optics.

2. The convergence of the elementary diopters depends on their radius R and on n. However, in the case of electrons these two factors are no longer independent, and are, in fact, related through $R = 2\varphi'/\varphi''$. It is this limitation which makes it impossible to correlate glass and electron optics completely (for example, we do not usually find divergent lenses, or any means to correct spherical aberration in electron optics).

3. The absence of reflected electrons accompanying the elementary reflections is related to the impossibility of making the electrical field strong enough to act perceptibly along a length as short as an electron wavelength (a few thousandths of a millimicron).

4. In electron optics n is proportional to the particle speed v. In glass optics, on the contrary, we have $n = c/v'$, where v' is the phase velocity of the light in glass. The discrepancy disappears if we introduce the phase velocity of the electron waves, w, which is related to v by $vw = c^2$.

Curiously enough, we can point out that a similar difficulty seriously delayed Fermat's publication of his principle (from 1657 to 1661, cf. Caratheodory,[28] pp. 1–10).

II. Magnetic Lenses

1. Magnetic Force and Equation of Motion

The force induced by a magnetic field **B** on an electron moving with velocity **v** is

$$\mathbf{F} = -e\mathbf{v} \times \mathbf{B} \tag{25}$$

The position of the vector is illustrated in Fig. 17 (right hand screw convention and current $-ev$). In all lenses the magnetic field B has an axis of symmetry $0z$ in the direction of motion of the electron. It can be defined by two components only, one radial, $B_r(r,z)$, the other axial,

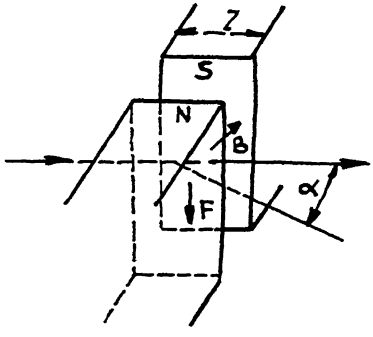


FIG. 17.—Magnetic prism.

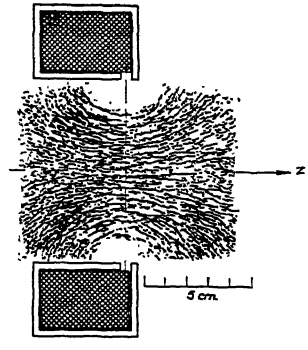


FIG. 18.—Shrouded magnetic lens.

$B_z(r,z)$. The tangential component $B_\theta = 0$. Thus the components of the force are

$$f_r = -ev_\theta B_z \qquad f_\theta = e(v_z B_r - B_z v_r) \qquad f_z = +ev_\theta B_r \tag{26}$$

Magnetic lenses consist, in most cases, of short coils, and often an iron magnetic circuit to concentrate the field within a small area. A current model is shown in Fig. 18. The spectrum of the lines of force illustrates that the electron will be under the simultaneous influence of a longitudinal component B_z and a radial component B_r, whose actions are equally important. The path of an electron emitted from A follows a straight line until it reaches the plane H where the action of the lens begins (Fig. 19). At that point the field is strongly radial (this is the α, β zone on the frontal projection of the trajectory and has a large curvature) and the force causes the trajectory to curve laterally, repelling it in front of the plane of the diagram. When this motion starts, a transverse component of the velocity v_t appears perpendicular to the axis. The longitudinal field then starts acting again through a radial force which sends the electron back toward the axis. Then, for a certain distance, the field is nearly longitudinal and uniform—this is the β,γ

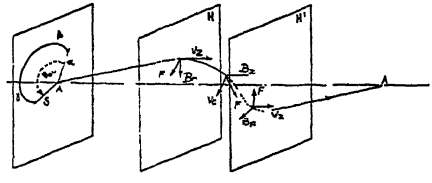


FIG. 19.—Ray path in a magnetic lens.

zone where the frontal projection is approximately a circle. The electron finally reaches the exit leakage field where the radial component B_r becomes noticeable again. The curvature again increases but B_r has changed its direction and the resulting force on v_z repels the trajectory toward the axis. The respective strengths of these actions is such that the emerging ray cuts the axis. The radial component of the exit leakage field forces the trajectory back to a meridional plane, which makes an angle θ_{0i} with the mean incident plane, thus causing a rotation of the image for an extended object.

It seems difficult to make a simple representation of the electron's motion. However, the problem may be simplified by the introduction of a new integral of the first order.

1. In fact, the old kinetic energy integral still holds. The force is perpendicular to the velocity and there is no tangential acceleration, so the velocity and the energy of the electron remain constant. Eq. (3) still holds even if the lens is not purely magnetic.

2. There is another integral of the first order, discovered by Störmer[29] at the beginning of the century. It was forgotten and then rediscovered by Busch in 1926. This integral of Störmer's is quite general, but we will use it first for the Gaussian approximation where the calculations are simpler.

2. Gaussian Approximation

a. Equation of Motion. 1. *Gaussian field.* The magnetic field, like the electrostatic field, satisfies the equation

$$B_r(r,z) = -\frac{1}{2} r \frac{\partial B_z(0,z)}{\partial z} \tag{27}$$

so that the optical properties in the Gaussian approximation depend only on the function $B_z(0,z)$, which we will denote simply by $B(z)$.

2. *Störmer's integral.* Equating the rate of change of the angular momentum to the torque around the axis, we get

$$m \frac{d}{dt}\left(r^2 \frac{d\theta}{dt}\right) = r \cdot e \left(B \frac{dr}{dt} + \frac{1}{2} r B' \frac{dz}{dt}\right) \tag{28}$$

Since the second member of the equation is the derivative of $\frac{e}{2} r^2 B$, upon integrating one finds

$$\frac{d\theta}{dt} = \frac{e}{2m} B + \frac{C}{r^2} \tag{29}$$

The constant of integration C is zero when the ray crosses the axis. This appears to be the case (1) for any ray direction when the object point

lies on the axis ($r_0 = 0$) or outside the lens $\left(r_0 \neq 0, \left(\frac{d\theta}{dt}\right)_0 = 0\right)$; (2) for meridional rays, when the object has any position outside the lens or is a virtual object inside; and (3) when a real object is immersed in the lens, for rays which cross the axis while still in the illuminating system or when emerging into the field-free space. Such a case exists in the microscope objective for rays which are meridional in the electron gun and in the image region. Störmer's equation then takes the simple form

$$\frac{d\theta}{dt} = \frac{e}{2m} B \qquad (e/2m = 8.797.10^{10} Cb/Kg) \tag{30}$$

This shows that a ray of the previous category upon leaving the lens as a straight line will cross the axis again ($B = 0$, for which $d\theta/dt = 0$).

3. *Larmor's moving frame.* Following Larmor's theorem, we further simplify the description of the motion, using a moving frame rotating about the axis, with the Larmor angular speed

$$\omega = \frac{e}{2m} B \tag{31}$$

as measured on the axis.

The simplification is evident for the family of rays with $C = 0$ since the relative trajectories will appear as plane and meridional as $d\theta/dt = \omega$. The transformation is still more valuable for the other rays, as will now be shown.

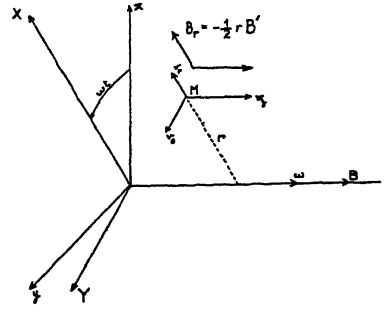


FIG. 20.—Larmor's rotating frame.

Applying Newton's equations in the moving frame,* to the true force, f, we must add two fictitious forces: the Coriolis force, $-2m\boldsymbol{\omega} \times \mathbf{v}'$, and the "moving space force" (corresponding to the transfer acceleration), which has two components, the centrifugal force $m\omega^2$ and the transverse force $-mr(d\omega/dt)$. The choice of ω enables us to cancel all but one component of the applied force. This can be seen from the following table where the speed and its transverse component are designated by v and v_θ in the true movement, by v' and v_θ' in the relative movement; ω is a vector of algebraic value ω along $0z$ (Fig. 20). The longitudinal

*See, for example, L. Brand, Vector and Tensor Analysis. Wiley, New York, 1947, pp. 123–124.

component of the force need not be considered since it is of a negligible order in the Gaussian approximation.

	Applied force in the true movement	Coriolis force $(-2m\omega \times v')$	Moving space force
Transverse component	ev_rB	$(-2m\omega \cdot v_r) \rightarrow -ev_rB$	
	$(-ev_zB_r) \rightarrow \frac{e}{2}\frac{dz}{dt}rB'$		$\left(-mr\frac{d\omega}{dt}\right) \rightarrow -\frac{e}{2}rB'\cdot\frac{dz}{dt}$
Radial component	$-ev_\theta B$	$(2mv_\theta' \cdot \omega) \rightarrow ev_\theta' B$	$(m\omega^2 r) \rightarrow \frac{e^2}{4m}rB^2$

The mutual cancellation of the transverse components is obvious; taking into account the relation $v_\theta = v_\theta' + \omega r$, the composition of the radial components leaves the net force f_r':

$$f'r = -e(v_\theta' + \omega r)B + ev_\theta' B + (e^2/4m)rB^2 = -e\omega rB + (e^2/4m)rB = -(e^2/4m)rB \quad (32)$$

Thus the equations of the relative movement in a frontal projection reduce to

$$\frac{d^2r}{dt^2} + \frac{e^2}{4m^2}B^2r = 0 \quad (33)$$

and on the $0z$ axis:

$$dz/dt = \sqrt{2e/m}\,\sqrt{\varphi_0} \quad (34)$$

φ_θ being the constant value of the electric potential throughout the lens.

1. For the class of rays for which $C = 0$ the transverse components of the initial velocity disappear, the trajectory is plane, and eliminating the time, in its meridional plane we have,

$$r'' + (e/8m\varphi_0)B^2r = 0 \quad (35)$$

It is useful to introduce the radius of the circle described by the electrons when they are fired at right angles to a uniform field of intensity B with a velocity corresponding to the accelerating voltage expressed φ_0 in volts.

$$(1/\rho)^2 = eB^2/2m\varphi_0 \quad (36)$$

(e.g. $\rho = 3.371\sqrt{\varphi_0}/B$; (centimeter; volt; gauss) or $\rho = 3.371.10^6 \sqrt{\varphi_0}/B$ (meter; volt; weber/m.^{2}). Equation (35) then reads

$$\frac{d^2r}{dz^2} + \frac{r}{4\rho^2} = 0. \tag{37}$$

2. The rays with $C \neq 0$ remain skew in the rotating frame. Using the complex representation as in the electrostatic case, we resolve this complication and prove that the skew rays coincide with the meridional rays at the image point.

3. The rotation of the meridional plane between object and image is given by

$$\theta_{0i} = \theta_i - \theta_0 = \sqrt{e/8m\varphi_0} \cdot \int_{z_0}^{z_i} B(z)dz = \int_{z_0}^{z_i} (\tfrac{1}{2}\rho)dr \tag{38}$$

4. One may easily translate the terms of our previous geometrical reasoning into a simple calculation, using the complex representation for the forces, as in Section I, 2, *b*.

$$f_u = je[Bu' + (1/2)B'u(dz/dt)]$$

Using the transformation

$$W = ue^{j\Omega} \qquad \text{with} \qquad \Omega = \int_{t_0}^{t} \omega dt$$

we are led immediately to the trajectory.

b. Optical Properties, Cardinal Elements. The rotation θ_{i0} of the meridional plane of the particle is a lens constant when the object and image remain outside the field. We are able to determine the image in two steps:

1. We first calculate the function of z which yields r. The corresponding plane curve represents the apparent trajectory for an observer located in the rotating frame. The differential equation allows us to determine the Gaussian cardinal points and to construct the image of a plane front.

2. Then a rotation of an angle θ_{0i} around the z-axis gives the actual image. When a variable electrostatic field is superimposed on the magnetic field, the ray equation takes the form

$$\frac{d}{dz}\left(\sqrt{\varphi}\,\frac{dr}{dz}\right) + \frac{1}{4\sqrt{\varphi}}\left(\varphi'' + \frac{e}{2m}B^2\right)r = 0 \tag{39}$$

The differential equation for the rays shows the same linear behavior as in the electrostatic case and leads to the same optical formalism.

c. Physical Properties. The physical properties are definitely modified, as one can guess from an inspection of the deflection formula (Fig. 17)

$$\alpha = l/\rho = \sqrt{e/2m\varphi_0} \cdot Bl \tag{40}$$

1. The coefficient e/m now appears in the equation and the form of the trajectory depends on the nature of the particle. For ions, which are heavier, the focal lengths are greater than for electrons. Magnetic lenses are of little use with heavy, high-speed ions.

2. The coefficients are no longer homogeneous with respect to the electric potential $\varphi(z)$. The values of the cardinal constants vary with the voltages applied to the instrument, and the currents feeding the lens. In particular, accidental fluctuations in these voltages will disturb the formation of images, and consequently the input voltage must be highly stabilized. The same precaution is necessary for the coil current.

3. An observer looking at the receding particles will see them turn clockwise where the field points in the direction of the flight and counterclockwise where the field points away. Thus, the shape of the trajectory depends on the direction of travel with respect to the field. Two trajectories with the same tangent at the same point are not the same if their speeds are different in sign (Fig. 21). The optical principle of the reversibility of light paths thus loses its validity and thereby bars any close analogy with the optics of crystalline media.

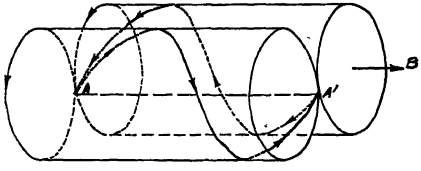


FIG. 21.—Irreversibility of trajectories.

In practice, however, this peculiarity does no harm since we are interested in the properties of the image itself rather than in the rays that produce it.

The rotation changes its sign with the field although the convergence does not vary with that change because it is dependent on B^2. It is therefore possible to suppress the rotation without destroying the focusing action. All that is necessary is to use two lenses back to back with reversed current directions. Actually this method is seldom employed because, for a given focal length, it leads to increased power consumption in the magnetizing coils and also emphasizes some aberrations of the system (Becker and Wallraff[30]).

3. Simple Models

Consideration of the simple case of the long coil (uniform field) and of the discontinuous model of the short coil is very instructive. We refer

the reader to Cosslett's textbook[1'] for a treatment of these cases and for an extensive bibliography.

4. *Tracing the Trajectories—Glaser's Field*

The step by step methods are applicable here as well as in the electrostatic case, but a simple model, first proposed by Glaser[14] offers new possibilities.

The $B(z)$ curve for shielded lenses has a characteristic bell-shaped form that can be approximated closely by the function

$$B(z) = B_c/[1 + (z/a)^2] \tag{41}$$

where B_c is the field at the lens' center, and a is the distance in which the longitudinal field drops to one half of B_c. Putting

$$x = z/a \qquad y = r/a \qquad k = a/2\rho_c \qquad p = \sqrt{1 + k^2} \tag{42}$$

where ρ_c is the value of ρ corresponding to the lens' center, the trajectory may be described in terms of elementary functions, and for a ray parallel to the axis we have

$$x = ctnu \qquad y = \frac{1}{p}\frac{\sin pu}{\sin u} \tag{43}$$

These ray equations enable us to determine the cardinal constants easily. Glaser's field is the first known type to provide two definitions for the cardinal points.

1. *Classical cardinal points.* The asymptote of the ray (eq. 42) is given by

$$y = -\cos p\pi + x(\sin p\pi)/p \tag{44}$$

and this leads to the classical constants of this symmetrical lens which, for the image side, are

$$f/a = p/\sin(p\pi) \qquad z_F/a = pctn(p\pi) \qquad z_P = pctn(p\pi/2) \qquad \theta_{0i} = \pi k \tag{45}$$

2. *Glaser's cardinal points.* Dealing with the trajectory (eq. 43) directly, for immersed objects (the image side) we get

$$g/a = 1/\sin(\pi/p) \qquad z_G/a = -ctn(\pi/p) \qquad z_Q/a = -ctn(\pi/2p)$$
$$\theta_{0i} = \pi k/p \tag{46}$$

It is noted that the rotation is still a lens constant.

3. *Relations between the two sets of cardinal points.* When the object and image are outside the lens, the two sets of cardinal points give the same result. This suggests the approximate relation $g - f = z_F - z_G$, which has been verified in practice.

The minimum practical focal length is $f = 1.5a$ for $k = 1$ in the first

case, and $f = a$ for $k = \sqrt{3}$ in the second case. If the same lens type could be used in the two stages of a microscope under the condition of minimum focal length, the objective would be more convergent than the projector, but these conditions are not feasible, and actually the opposite proportions are observed.

We will not go further in our study of this interesting model since it has been extensively treated in textbooks and in Marton's report.[1]

4. *Correction method.* Glaser's model duplicates an optical system with an accuracy of a few per cent for various pole-piece shapes and magnetic conditions. But it also happens that the experimental curve for $B(z)$ differs from Glaser's bell-shaped one, particularly when the iron is far from saturation.

In these cases an improvement devised by Siegbahn[31] is found useful. It brings out the solution for the field in the form

$$B = B_c e^{-(z/b)^2} \tag{47}$$

Developing $B(z)$ in a power series, we find a differential equation which reduces to Glaser's form, except for the small term in $(z/b)^4$ and for k which takes the form $K' = b/2\rho_c$.

The preceding simple formulae for f, z_f, g and z_g remain valid under the substitution of a for b, and k for k'; the half-height width of the field is then $a' = 0.83b$. This has been checked rather closely (to within a few parts in a thousand) by Siegbahn with his measurements on the β-ray spectrometer.

More generally we may measure the experimental field which approaches Glaser's equation and consider the small difference as a disturbance of the field. This can easily be done by Scherzer's[1'] method of calculating the aberrations. It gives a clear picture of the general optical features as well as a precise determination of the cardinal points.

5. *Optical Analogy*

a. Index for the Fictitious Movement. 1. *Gaussian approximation.* If we now examine the structure of the Lagrange-Helmholtz law in the pure magnetic case, we come to the surprising result

$$f_i/f_0 = -1$$

If we wish to resort to the use of an index we must accept the value $\sqrt{\varphi_0}$ on the axis or, in other words, we must recognize that the presence of a magnetic field only changes the off-axis index.

In the Gaussian tubular space, the value of the electrostatic index is

$$n^2 = \varphi(z) - (r^2/4)\varphi''(z) \tag{48}$$

On the other hand, examining the path equation, we see that we only need to substitute φ'' for $\varphi'' + \frac{e}{2m} B$ to account for the presence of a magnetic field in the mixed focalization. This suggests that we make the same substitution for the electrostatic index. Doing this we shall get the general solution

$$n^2 = \varphi(z) - (r^2/4)[\varphi'' + (e/2m)B^2] \tag{49}$$

which correctly describes the fictitious trajectory in the meridional plane. This is apparent from the observation that the electrostatic equations applied to the fictitious potential $\varphi_1(z) = n^2(r,z)$ give the correct laws of motion in the rotating frame.

2. *General case—Störmer's function.* To obtain the expression for the index in the general non-Gaussian case, we need the general expression of Störmer's integral. He remarked that, since the symmetry of revolution is $B_\theta = 0$, the condition div $\mathbf{B} = 0$, fulfilled by the field, reduces to

$$\frac{\partial(rB_r)}{\partial r} + \frac{\partial(rB_z)}{\partial z} = 0 \tag{50}$$

But under this condition the expression $-rB_r dz + rB_z dr$ is a total differential of a certain function $Q(r,z)$, and we can write

$$dQ = -rB_r dz + rB_z dr$$

The equation for the angular momentum may be integrated, giving

$$r^2 \frac{d\theta}{dt} = \frac{e}{m}[Q(r,z) + C] \tag{51}$$

$Q(r,z)$ is the magnetic flux through a circle of radius r, centered on the axis and located in the frontal plane z; thus $Q = rA_\theta(r,z)$, where $\mathbf{A}$ is the ordinary vector potential. It is then simply a matter of calculations to show that the two equations of motion for the coordinates r and z may be written as for an electron moving in an electrostatic field of potential $\Psi(r,z)$ given by

$$\Psi(r,z) = \varphi_0 - (e/2m)[A_\theta + C/r]^2 \tag{52}$$

Thus the trajectory, as seen by an observer located in the rotating meridional plane, is the same as the real trajectory in the corresponding hypothetical electrostatic case. (It must be remarked that the speed of rotation differs from the ω of eq. (31) even in the Gaussian approximation, if the constant C of eq. (29) is different from 0.)

We can then follow the refraction of the ray step by step, using the new index which, in the most general case, takes the form

$$n^2 = \Phi(r,z) - (e/2m)[A_\theta + C/r]^2 \tag{53}$$

Figure 22 gives the general shape of the surfaces n = constant for a pure magnetic lens. Proposals for a practical use of Ψ (Sandor[32]) have brought no response so far. The usefulness of Störmer's theory is that it clarifies the relations between the electrostatic and magnetic systems. This will be apparent in the theory of aberrations where it is possible to correlate quantitatively the magnetic and electrostatic aberrations.

b. The Index for a Stationary Observer. It is also possible to use a refractive index for the direct determination of the real motion. This

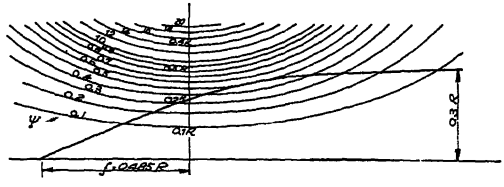


FIG. 22.—Surfaces n = constant, in the rotating frame, for a circular current I = 1000 Amps., ϕ = 5000 volts.

technique was introduced by Schwartzschild in 1903, who extended the validity of the variational principle to the magnetic case, taking the magnetic field into account through a new term in the kinetic energy. This procedure (see Gabor[2]) leads to the following definition of the index

$$n^2 = \Phi(r,z) + \mathbf{v} \cdot \mathbf{A} \tag{54}$$

The equations of motion would then appear as Euler's equation attached to the variational equation

$$\delta \int_A^B n ds = 0 \tag{55}$$

The magnetic field behaves like an optically anisotropic medium but of a rather strange kind. The index of refraction n has two different values for the two opposite directions of motion of the particles and the principle of the reversibility of the light path is no longer valid. On the other hand, n can be arbitrarily modified by adding to **A** any function of the form grad $f(x,y,z)$ without changing the path of the particles, since the grad function disappears in the Euler equations. The index (eq. 54) manifests only one simple property, it is constant for the rays contained in a plane perpendicular to the vector **A**, and since **A** moves in closed

circles perpendicular to the meridional planes the index is constant for meridional rays in systems with axial symmetry. It is in this peculiarity, as Gabor,[2] has mentioned, that the success of the rotating plane theory and of the fictitious index lies.

Because of these difficulties the general index is seldom used. Nevertheless the general principle (eq. 55) is useful as a starting point. It can be employed to cover the relativistic case also and its great generality enables us to treat any kind of focalization along the same deductive lines. In addition it gives us an easy way to eliminate the time parameter from the trajectory equations. There, we recognize the well known automatism brought about in mechanics by the use of Lagrangian theory.

III. Relativistic Lenses

At high speeds relativistic mechanics governs the electrons' behavior. The equations are more complex, but the general line of reasoning and the results are similar. Starting with Einstein's equation instead of Newton's, one can establish the existence of a first order imagery; whatever the speed of the particles may be the differential equations of the rays are still linear and thus we again have all the Gaussian properties.

1. Electrostatic Lenses

Let us define $\epsilon(z) = (e/2m_0c^2) \cdot \varphi(z)$ (with $e/2m_0c^2 = 0.977.10^{-3} \cdot (kv)^{-1}$). The relativistic Gaussian expression reads

$$\frac{1+\epsilon}{1+2\epsilon}\varphi r'' + \frac{1}{2}\varphi' r' + \frac{1}{4}\varphi'' r = 0 \tag{56}$$

and the loss of homogeneity appears in the first coefficient. Putting

$$R = r\varphi^{\frac{1}{4}}[1+\epsilon]^{\frac{1}{2}} \tag{57}$$

we get a reduced equation similar to (22) with the exception of the coefficient $(\frac{3}{16})(\varphi'/\varphi^2)$ which now appears multiplied by the factor $K(\epsilon)$ defined as

$$K(\epsilon) = [1 + (4\epsilon/3) + (4\epsilon^2/3)]/(1+\epsilon)^2 \tag{58}$$

As a rough approximation we may neglect the variation of ϵ with z and take a constant mean value $\bar{\epsilon}$. Equation (58) then shows that as this mean value increases the correction factor K decreases from a value 1, at low voltages, to a minimum value of 0.88 at 600 kv, again reaching the value of 1 at 2 Mv, and $\frac{4}{3}$ at still higher voltages. It follows that if we multiply the voltages by the same factor in the electron gun and in

the lens, we should find the focal length first increasing, reaching a maximum, and then decreasing to an asymptotic value. An accurate evaluation confirms this qualitative prevision, as can be seen from Fig. 23 which shows the behavior of a three-electrodes unipotential lens (Laplume[33]).

The danger of a breakdown prohibits the use of potential differences much larger than 100 kv between adjacent electrodes. Thus in the megavolt region grid lenses with their low voltage requirement become interesting (Gabor[34]).

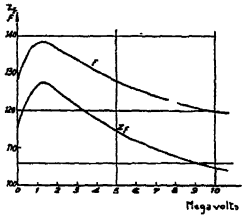


FIG. 23.—Focal length at relativistic speed, electrostatic case (lens of Fig. 30).

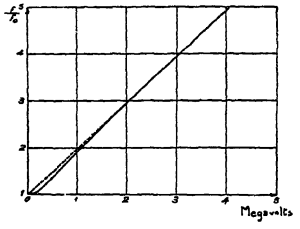


FIG. 24.—Focal length at relativistic speed, magnetic case.

2. *Magnetic Lenses*

Relativity manifests its influence by a constant factor applied to φ_0 or to ρ^2, which becomes $\varphi_0(1 + \epsilon_0)$ or $\rho^2(1 + \epsilon_0)$ in the ray equations (35) and (37) as well as in Glaser's model (eq. 42). The focal length increases directly with φ_0 (Fig. 24). At megaelectron volt speeds magnetic lenses become ineffective because the upper value of B cannot exceed 26,000 gauss.

IV. Image Defects-Limits of the Gaussian Approximation

In practice, the Gaussian theory has had considerable success. A great variety of electron optical instruments obey its laws and furnish images which appear sharp and faithful. Nevertheless, a searching study always reveals "defects" in the image: instead of an image "point" we have a spot which may be displaced from its Gaussian position at the border of the field.

Is the origin of these defects to be found in the terms we deliberately neglected when establishing the Gaussian theory or are they due to other causes we have not yet mentioned? An answer to this question can only be searched for in a careful discussion of each particular instrument and is not known in every case.

We will not discuss space charge defects, a phenomenon without analogy in glass optics and one which does not affect the truly imaging devices (microscope, image tube) where a sufficient image brillancy is obtained with minute current densities and high voltages. We also neglect chromatic aberration, which is adequately treated in its instrumental aspect in textbooks, but whose origin in the interaction between the beam and the specimen is still largely mysterious (Gabor,[2] Hillier,[35] Boersch[36]).

However, we will, treat in greater detail the geometrical aberration of the third order (from the term in r^3 which were neglected in the Gaussian theory) and the mechanical aberrations resulting from the imperfections of the mechanical design, which determine the upper limit of the Gaussian tubular space materialized by a set of diaphragms along the axis.

It is impossible to improve the quality of the images indefinitely by reducing the diameter of the diaphragms. One is soon stopped by the increasing lack of intensity (for example, one must choose between intensity and sharpness in the cathode ray tube), or by diffraction: this is the dominant effect in the electron microscope when the aperture α of the elementary pencils from a point in the specimen is less than 5×10^{-4} radians. Then the image of a point appears as a small Airy spot, the diameter of which, when divided by the magnification, is

$$d = 24/\alpha \sqrt{\varphi} \tag{59}$$

(d = millimicrons, α = milliradians, φ = kv).
(The working voltage φ, varies over too small a range to be a significant factor in the phenomenon.) Thus, in the microscope, diffraction is balanced against mechanical or geometrical defects to give maximum sharpness (Gabor[2]). We will come back to this problem later after we have become better acquainted with the defects themselves.

V. Third Order Aberrations

1. Definition of the Aberration Coefficients

x_0 and y_0 are the coordinates of A in the object plane $z = z_0$; x_i and y_i are the coordinates of A', the Gaussian image of A in the plane $z = z_i$ (Fig. 25). A ray coming from A crosses at point $N(x_a, y_a)$, in the plane of the aperture that limits the beam diameter. When the real aperture is inside the lens one should locate this point N in an imaginary plane, which is in the object space outside the lens field influence. It is convenient to choose for the position of this plane the image of the aperture plane through the first part of the lens. (This imaginary plane is known

as the "entrance pupil," it is plane D' in Fig. 25.) The ray proceeding through A and N emerges from the lens and intersects the image plane $z = z_i$ at $A''(x_i', y_i')$ and not at the Gaussian image $A'(x_i, y_i)$; the length $A'A''$ depicts the aberration

$$A'A'' \begin{cases} \xi = x_i' - x_i = f(x_0, y_0, x_d, y_d) \\ \eta = y_i' - y_i = g(x_0, y_0, x_d, y_d) \end{cases} \tag{60}$$

We want to compute the serial development for the f and g functions just to the third order. The symmetry of revolution causes the second order terms to disappear since the quantities x_0, y_0, x_d, y_d, x_i, y_i change their signs simultaneously in a 180° rotation. This also results in

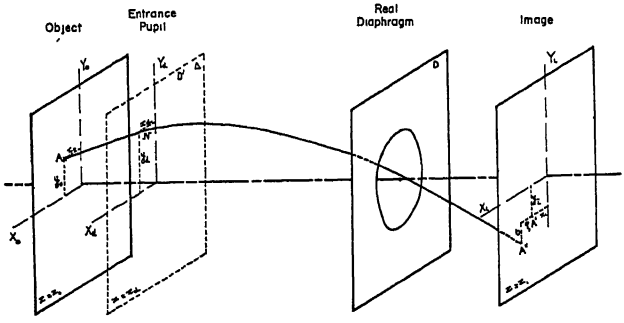


FIG. 25.—Definition of the transversal aberration.

a simple structure for the third order terms (as is explained in detail by Ramberg[37]). If we write

$$u = x_0 + jy_0 \qquad v = x_d + jy_d \qquad w = \xi + j\eta \tag{61}$$

w becomes a function of u,v, and of the complex conjugates

$$\bar{u} = x_0 - jy_0 \qquad \bar{v} = x_d - jyd \tag{62}$$

(because the correspondence between w and x_0,y_0, x_1,y_1 is of a more general character than that which would be established by a function of u,v, alone). An arbitrary θ rotation of the coordinates around the two axis transforms u,v, into $u' = ue^{j\theta}$, $v' = ve^{j\theta}$ whose conjugates are now $\overline{u'} = \bar{u}e^{-j\theta}$, $\overline{v'} = \bar{v}e^{-j\theta}$, and changes w to $w' = we^{j\theta}$. The new relation between w' and u',v', $\overline{u'},\overline{v'}$ must therefore remain the same as the old one between w and u,v, $\bar{u},\bar{v}$. This proves, if we write the third order terms in the form $(au + bv) \cdot I_2(u,\bar{u},v,\bar{v})$, that I_2 is a second degree polynomial which is invariant under the same transformation, e.g., of the type $cu\bar{u} + du\bar{v} + e\bar{u}v + f\bar{v}v$.

For w the result is

$$w = (A + jA')v^2\bar{v} + (B + jB')u^2\bar{v} + (C + jC')u\bar{u}v + (D + jD')\bar{u}v^2 + (E + jE')uv\bar{v} + (F + jF')u^2\bar{u}$$

A ray starting from an object point on the axis intersects the axis again in the image space: this condition gives $A' = 0$.

A further simplification is brought about by Malus' law (we refer the reader to standard optics texts for its development). This law states that the rays are normal to a family of surfaces, the wave surfaces. This property is expressed by the condition that the

$$\text{imaginary part of } \left(\frac{dw}{dv}\right) = 0 \tag{63}$$

and one gets $C' = 0$, $2D = E$, $2D' = -E'$. The aberration is now determined by eight coefficients

$$W = Av^2\bar{v} + (B + jB')u^2\bar{v} + Cu\bar{u}v + (D + jD')\bar{u}v^2 + 2(D - jD')uv\bar{v} + (F + jF)u^2\bar{u} \tag{64}$$

This formula applies to magnetic lenses. In the electrostatic case the non-meridional components of the force disappear and every meridional plane becomes a plane of symmetry. Replacing $\bar{u},u,\bar{v},v$ with $u,\bar{u},v,\bar{v}$ does not change w, and consequently we have

$$B' = 0 \qquad D' = 0 \qquad F' = 0 \tag{65}$$

The number of coefficients is reduced to five as in glass optics.

The preceding formulae are sufficient to determine the form and the position of the small spot that plays the role of the image point. When all the terms are of the same importance, the resulting observed figure is very complicated (Bricka and Bruck[38]). But it is possible to choose the experimental conditions (Cosslett[1']) so that all but one term is negligible. Now that the defects have been separated we may give a name to each in accordance with our glass optics experience.

The reader is referred to textbooks[1'-2'] for descriptions of the individual aberrations.

2. *Further Comparison between Electrostatic and Magnetic Lenses*

Störmer's potential Ψ will allow us to distinguish further between the electrostatic and the magnetic case. The distinctions arise from differences in the power series developments of the two functions Φ, the electrostatic potential, and Ψ given by eq. (53). There exist three cases:

1. The constant C plays no part. This condition exists a) for spherical aberration since the object point is on the axis, $C = 0$; b) for

distortion because this defect does not change the sharpness, and rays with $C \neq 0$ must give the same result as rays with $C = 0$, so that C disappears from the final formulae; c) for coma, where two skew rays starting in two directions symmetrical with respect to the initial meridional plane are reunited in the image plane. Since they correspond to different values of the constant C, this C must also disappear from the coma expression.

In these cases one may pass from electrostatic to magnetic expressions by simply substituting φ for φ, φ'' for $\varphi'' + \frac{1}{2}\frac{e}{m}B^2$, and φ^{IV} for $\varphi^{IV} + \frac{2e}{m}BB''$. (These are corresponding terms in the power development of Φ and Ψ.)

2. Two other defects are common to magnetic and electrostatic focalization: curvature of the field and astigmatism. In the magnetic expression, C must be taken into account by adding two new terms in the coefficient expression.

3. Three aberration coefficients apply only to the magnetic case. These so called "anisotropic" defects correspond to the differences between the true rotation of the meridional planes of the particles and the Larmor rotation of the Gaussian image given by the field's value on the axis. Their position depends on the sign of $B(z)$ and changes to one, symmetric to the first one, with respect to the meridional plane by reversal of the magnetizing current.

3. Calculation of the Aberration Coefficient

The detailed evaluation of the aberration coefficients may be obtained by the trajectories method, which we followed in the first order approximation. It is only necessary to reintroduce the neglected r^3 terms in the equations. The equations are now much more complicated but may still be solved by the perturbation method as first proposed by Scherzer. We refer the reader to the basic and introductory treatment of Cosslett.[1']

By partial integration, aided by the first order equation and the Lagrange-Helmholz law, Scherzer was able to eliminate all the higher derivatives of the field. Thus an evaluation is possible knowing only φ, φ', φ'', B and B'. When we use the reduced variable R in the determination of the Gaussian trajectory, it is convenient to conserve the variable R in the calculation of aberrations; this is possible at least in the case of spherical aberration (Recknagel[39] and Bruck[40]). The values of the coefficients may be also determined experimentally, and the results agree with the theoretical findings. The methods used are transformations of standard techniques in glass optics and offer no special difficulty

except in the case of microscope lenses where the dimensions involved are very small. For spherical aberrations an indirect method is successful, using the close relationship between spherical aberration and distortion (von Ardenne,[41] Dosse,[42] and Mahl[43]). The magnitude of this latter defect is easy to measure, when the object is formed of concentric circles[43] or parallel lines.[42]

4. *Value of the Aberrations*

The literature gives much information about aberrations, but this information is difficult to use and correlate because of the great diversity of the parameters chosen by different authors. The only generally accepted convention is to divide the aberration $|w| = |A'A''|$ by the magnification M to obtain a length $|w|/M$ which is significant in the object space. It would be advisable to free ourselves from any scale factor, however, and it seems appropriate to divide the aberration by the focal length of the lens, using $|w|/Mf$.

The most difficult problem is the choice of variables. x_d, y_d are convenient theoretically, but angular variables give a more concrete presentation of the results. In any case the choice of a coordinate in the plane of the aperture itself instead of at the entrance pupil plane is unfortunate since it introduces a parasitic variable in the final results (the aberration coefficients then include the effect of the magnification between entrance pupil and aperture) when the aperture is immersed in the field. For example, this would lead us to believe falsely that spherical aberration depends on the position of the aperture.[44]

5. *Correction of the Geometrical Aberrations*

a. Influence of the Diaphragm's Position. In electron optics there are fewer opportunities to correct for aberrations than in glass optics.

The sole parameter that we can adjust as freely as in glass optics, is the position of the aperture. In this way we may act upon some of the aberrations that affect the image of a point away from the axis: coma, astigmatism, curvature of field, and distortion. In this case it is clear that when the aperture is displaced along the axis a different portion of the imaging field is put into action (Fig. 26). For reasons of symmetry this remedy is not effective with those defects originating in magnetic rotation, or in spherical aberration.

The possibility of this correction has been investigated in detail for Glaser's magnetic model (Glaser and Lammel[44]). It seems probable that it exists for the electrostatic lens too. In particular the suppression of distortion has been investigated in detail and studied experimentally by

Hillier[45] for the magnetic projector lens of the microscope and by Le Rutte[46] for its electrostatic counterpart and for the electrostatic mirror.

b. Spherical Aberration. There is no possibility of a complete correction for electron lenses.

This impossibility becomes apparent when one considers the ideal distribution of the index, such as when all rays emanating from a particular point intersect again at a single point. Gabor[47] resolved this problem and found a distribution for n identical with the one Maxwell attributed to fish eyes and which seems impossible to approach even roughly by any kind of electrode or pole piece (Glaser[48]).

A more direct demonstration of the same impossibility, and also the earliest, is due to Scherzer.[49] He showed that the aberration coefficient

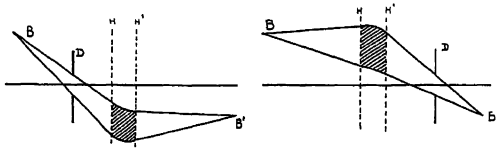


FIG. 26.—Importance of the diaphragm position.

may be expressed as a sum of perfect squares which in the case of purely magnetic focusing reduces to (with $r(z_0) = 0$, $r'(z_0) = 1$):

$$\frac{|w|}{M} = \frac{1}{16}\frac{e}{m\varphi_0}\alpha^3 \int_{z_0}^{z_i} \left\{ \left(B' + B\frac{r'}{r} \right)^2 + B^2\frac{r'^2}{r^2} + \frac{e}{4m}\frac{B^4}{\varphi_0} \right\} r^4 dz \qquad (66)$$

It can only be nullified by equating the individual terms to zero, but then the focalization also disappears. The situation is very different in glass optics: there, divergent lenses afford a negative contribution, convergent lenses a positive contribution, to the defect and a cancellation is possible by an association of the two, while still retaining the convergence. This is impossible with electrons since the divergent as well as the convergent portions of the lenses have a positive aberration.

Glaser[50] thought at one time that he had nullified this conclusion, and he has indeed transformed relation (66) to

$$\frac{|w|}{M} = \frac{e}{96m\varphi_0}\alpha^3 \int_{z_0}^{z_i} \left(\frac{2e}{m\varphi_0} B^4 + 5B'^2 - BB'' \right) r^4 dz \qquad (67)$$

where the terms are of mixed signs and where he found a field satisfying the equation

$$(2e/m\varphi_0)B^4 + 5B'^2 - BB'' = 0 \qquad (68)$$

This field would be aberration-free if formula (67) were valid for it. However this is not the case. Equation (67) is equivalent to eq. (66) only for strong fields and Glaser's field, on the contrary, always gives a

negligible convergence (Recknagel[51] and Rebsch[52]) because of its small longitudinal extension. Thus the proposition is contradictory and of no value. Formula (67) nevertheless remains useful for ordinary strong lenses.

c. Partial Correction. It is still possible to choose the fields so as to reduce the defect. Physically, we can characterize the defect by the fact that the lens is more convergent for marginal than for paraxial rays, and we find two causes for this effect:

1. The longitudinal component of the speed is smaller for the marginal rays and we can do nothing to change this.

2. The longitudinal component of the field may be stronger for fringing rays in regions where $B''(z) < 0$, since $B_z(r,z) = B(z) - (r^2/4)B''(z)$, contributing positively to the aberration; or the longitudinal component may be weaker for fringing rays in regions where $B''(z) > 0$, this reducing the aberration.

Siegbahn[53] found a means of realizing magnetic fields of the second kind (the intensity of B is largest in the object and image region) and obtained by this means a substantial improvement of brightness in his beta-ray spectrograph.

In such an apparatus however, the magnification is nearly unity. Marton[54] has successfully extended the solution to highly magnifying lenses but in the electron microscope the resulting progress is negligible because the sharpness of the image depends on the fourth root of the aberration, a very unfavorable law.

d. Lenses and Mirrors. Scherzer's formula does not hold for mirrors, as has been shown by Recknagel.[55] The aberration may be negative for a converging mirror and associating it with a lens may produce a total correction. In the microscope, however, it is not practical to locate the mirror close to the lens and no position has yet been found near enough to permit any useful cancellation. When the mirror is far from the lens its correcting contribution to the aberration is much too weak because of the small inclination of the ray to the axis at this point. Ramberg[56] has recently reviewed this question.

e. Correction of Spherical Aberration by Means of Cylindrical Lenses. We shall postpone the description of this topic until Section VI, 3 in order to utilize the introductory treatment of asymmetrical fields in Sections VI, 1 and 2.

VI. Mechanical Defects

Up to now we have supposed that the electrode systems were geometrically perfect, that is, that the electrodes actually possessed the ideal geometric form which we gave them in theory. But in practice

these systems are manufactured on lathes which have a limited precision, and the assembly of these pieces may leave much to be desired. We may have planes that are inclined rather than parallel and axes that are displaced rather than coincident.

These imperfections appear in the image to an extent that depends on the degree of the mechanical precision. But these defects are still noticeable when we employ all the known refinements of the mechanical art. This is especially true of the microscope objective whose electron optical qualities give an accurate test of the precision of the mechanical construction and assembly. An ellipticity of a fraction of a micron in the bore of modern electron microscope objectives (whose diameter is a few millimeters) is observable; a still greater accuracy would be desirable.

Conditions are different in glass optics and defects of this type are, for the most part, negligible. This advantage is principally due to the extreme machinability of the material and to the geometric simplicity of the desired forms, planes and spheres. The error may be reduced to less than one tenth of a micron if necessary.

1. Classification of the Mechanical Defects

We are able to classify these defects according to their origin: a) alignment defects, decentralization and loss of parallelism and b) asymmetrical defects of the pieces themselves.

An elaborate theory of the first group exists (Cotte,[57] Wendt,[58] and Hachenberg[59]) but seems superfluous in most practical cases. It is sufficient to break up the lens system into decentralized elementary lenses and to follow the formation of the image as it passes from one to the other (Bruck[60]). The general conclusion is that below one one-hundredth of a millimeter decentralization is insignificant and that it is always possible to make the electrodes big enough to render any small cause of lack of parallelism imperceptible.

Asymmetrical defects are more serious and also more difficult to study. The easiest way to start is to consider a very simple model. Let us suppose that the central electrode has a hole which is slightly elliptical (Fig. 27). Consequently the potential Φ no longer possesses a symmetry of revolution but instead has two planes of symmetry, that is, those which contain the axes of the ellipse. We are able to represent them by a development of the form

$$\Phi(r,\theta,z) = \varphi(z) - (\tfrac{1}{4}) \cdot \varphi''(z) \cdot [1 - \epsilon(z) \cos 2\theta] r^2 + \cdots \qquad (69)$$

Therefore, the study of the paraxial trajectories, when the object point is on the axis, enables us to define two only slightly different Gaussian systems for two families of rays in the planes of symmetry $x0z$ and $y0z$.

These two families produce a clear and sharp image at two points along the z-axis, A_1 and A_2. For other rays, their projections on the planes of symmetry will cross at A_1 and A_2, respectively. The result is that the rays themselves cross two small straight lines each situated in one of the planes of symmetry. As a result, a system of object points in the plane front of A gives rise, in the plane front A_1 and A_2, to two systems of short straight lines called "focals" (Glaser[61]). We observe that the approximate image is given by a small "confusion circle" in an intermediate

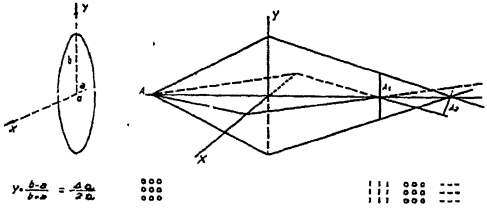


FIG. 27.—Elliptic aperture and its optical effect.

plane and that the radius d of this circle, which is taken as the measure of the defect (called astigmatism of ellipticity), is given by the formula

$$d = C_e\alpha \tag{70}$$

where α is the half aperture of the pencil and C_e is related to the coefficient ϵ, which defines the field dissymmetry by the formula

$$C_e = \frac{1}{4r_0'^2\sqrt{\varphi_0}} \int_{z_0}^{z_1} \frac{\varphi''\epsilon r^2}{\sqrt{\varphi}}\, dz \tag{71}$$

When the object is near the lens focus point a first approximation to C_e may be given by the equation

$$C_e = 1.2f\bar{\epsilon} \tag{72}$$

(assuming that ϵ has an average value $\bar{\epsilon}$ independent of z).

Finally, we must relate the coefficient of asymmetry of the field to the ellipticity of the lens' opening, which is rather difficult and which has been calculated by F. Bertein.[62–65] The results have been checked experimentally by Bruck[66] by delicate measurements in the electrolytic tank and by Cotte[67] using an independent theory.

The equation relating ϵ to η, the ellipticity of the *aperture* ($\eta = \delta a/2a$, where a is the radius of the hole and δa the error in the radius) is

$$\epsilon\varphi'' = \frac{0.88E_r\eta}{a} \frac{1}{(1 + 0.7z^2/a^2)^{3/2}} \approx \eta\varphi_e'' \tag{73}$$

(E_r is the radial field on the inner surface of the hole and φ_c'' is the value of $\varphi''(z)$ at the center of the lens). Finally the defect may be linked directly to its cause δa by the relation

$$C_e \approx f\eta \approx \delta a \tag{74}$$

This theory explains the operation of the microscope in accordance with the physical facts. Careful observation of the image will show that it suffers from an astigmatism characteristic of mechanical asymmetry (Hillier[68]; Bertein and Regenstreif[69]). Combining this defect with that of diffraction, the following equation is obtained (Bruck and Grivet[70]; Bruck[71]) for the resolving power d and the optimum half aperture α of the illuminating system.

$$\begin{aligned} d &= 0.85\lambda \sqrt{C_e/\lambda} \\ \alpha &= 0.71 \sqrt{\lambda/C_e} \end{aligned} \tag{75}$$

where λ is the electronic wavelength (mμ) and $\lambda^2 = 1.24/\varphi$ (volts) (which gives correctly and to scale the values observed when δa is approximated in accordance with the measured errors in cross section of the central electrode). One important result in regard to the machining of the pieces is that the resolving power depends on the absolute value of δa, that is, on the precision of the lathe operation. It is independent of the bore diameter, which we can make small enough to give us the minimum focal length necessary for large magnification.

This theoretical treatment also gives the optimum objective aperture, in reasonable agreement with our experimental observations.

A similar theory for magnetic lenses has been developed by Hillier and Ramberg[68] and applied with success to the magnetic microscope. In this case the cause of the defects resides perhaps more in the magnetic inhomogeneity of the pole pieces than in their mechanical precision.

This theory gives an order of magnitude only, since the methods employed to calculate the combined mechanical and diffraction defects are still very rough. The difficulties of the problem may be appreciated by considering the merging of diffraction and spherical aberration, which is better, but still insufficiently, understood. The theory only gives the result for the Gaussian image plane, which is not the most favorable one (Glaser,[72] Bruck,[73] and Marechal[74]).

2. Correction of Mechanical Defects

The choice of the elliptic model was completely arbitrary and not at first justified, but then Bertein[64,75] was able to show its importance by giving a general theoretical treatment of mechanical defects, exposing the causes in detail and giving much new information on the subject.

The principal value of this theory is that it shows us how we may correct mechanical aberrations. This possibility was also independently discovered by Hillier and Ramberg[68] and Scherzer.[76]

The method suggested is the use of weak cylindrical lenses with a very marked mechanical asymmetry. A weak voltage is applied to this electrode system in the case of an electrostatic lens and a weak field is used in the magnetic lens case. The corrector lens is placed immediately in front or behind the lens to be corrected to introduce a defect of the same nature but of opposite sign. The difficulty is to determine the

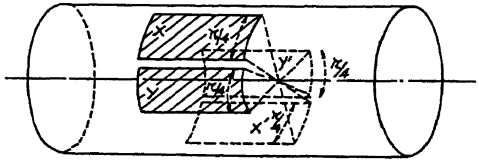


FIG. 28.—Correction of the ellipticity effect.

type of corrector necessary to nullify any particular defect without creating new defects. Care is necessary, for example, in correcting astigmatism, in order not to introduce a distortion defect. Some very simple correctors, as can be seen in Fig. 28, possess this important property.

3. Correction of Spherical Aberration by Means of Cylindrical Lenses

Compensation for astigmatism due to ellipticity is possible even if the defect is large. This is also true for glass optics; we know that two cylindrical glass lenses produce the same effect as a single spherical lens. It is necessary to employ three cylindrical glass lenses if we wish to eliminate distortion.

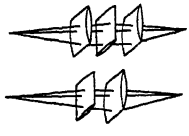


FIG. 29.—Two crossed cylindrical lenses are equivalent to a spherical lens.

Scherzer[76] has noticed that such an assembly of strong correctors in electron optics possess some remarkable advantages over ordinary round lenses. In fact, we are practically able to nullify spherical aberration along with the chromatic defect (Fig. 29).

If we do not have symmetry of revolution, the annoying expression $R = 2\varphi'/\varphi''$, which links the index $n = \sqrt{\varphi}$ and the curvature $1/R$ of the elementary diopters, disappears. According to the meridional plane chosen, the diopter meridians have a curvature that varies between two limits. The difference between these limits can be chosen at will, since it depends on the asymmetry of the cylindrical lens and is independent of the index n of the diopter. If we could obtain a refraction which was an

average between the two curvatures considered, nullification of the spherical aberration would then be possible using this new parameter. The average curvature would then not be related to the index of the diopter. Scherzer's plan realizes these conditions.

The idea is very engaging, but the practical realization presents many difficulties. First of all, the complexity of the correctors is annoying, involving mechanical alignment and regulation. Still other schemes have been proposed for the same purpose, the most recent ones by Gabor[77] and Hubert,[78] but they are all difficult to achieve. It would be very worth while to overcome these difficulties; it is possible that the first microscope to give us an image of the atom will utilize such complicated structures.

VII. Some Typical Lenses

1. Two Cylinder Lenses

This widespread type is adequately discussed in standard texts and we shall only describe an important advance which has recently been made by Hutter.[79] Transposing Glaser's model to electrostatics, he can give a complete description of the optical properties of the two cylinder lenses. The proposed formula for $\varphi(z)$ is

$$\varphi(z) = \varphi_0 e^{K \arctan(z/a)}$$

which fits the experimental curve reasonably well. The results are expressed by Hutter in terms of Glaser's cardinal point set G and Q, which, as previously explained, does not cover the case of virtual immersed objects. The classical cardinal elements covering this case could easily be found by a slight manipulation of the formula (eq. 45). The model is very flexible and the treatment is quite complete, giving both chromatic and spherical aberration.

Gundert[80] gave a thorough discussion of the adaptation of this type of lens to the function of a second lens in the oscillograph tube. The same arrangement of electrodes, but used as a mirror, has been described by Nicoll.[81]

2. Three Electrode "Independent" Lens

This type is extensively used in the electron microscope and as a second lens for cathode ray tubes, where it presents certain constructional advantages.

Microscopists prefer to keep the central electrode at the same potential as the cathode (unipotential lens), which simplifies the electrical connections. The Gaussian elements are given in Figs. 30, 31 and 32, after Bruck,[40] who also computed the spherical and chromatic aberration.

(Typical values of $|w|/Mf$ are 8 in one case, 4 in the other.) The results are in agreement with Mahl[43,82] and Dosse,[42] who gave a complete table of all the aberrations.

The use of a "unipotential" lens efficiently protects the system against quick voltage fluctuations. A good resolving power (better than 2 mμ and depending on the machining accuracy of the central aperture Grivet

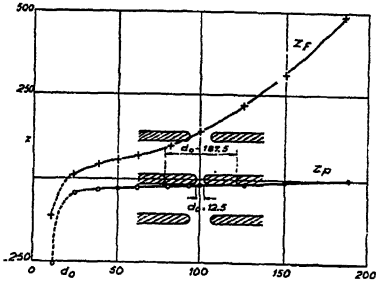


FIG. 30.—Three diaphragm independent lens: focus and principal planes position (Bruck[40]).

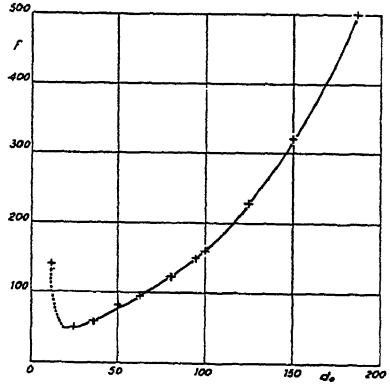


FIG. 31.—The three-diaphragm independent lens: focal length (Bruck[40]).

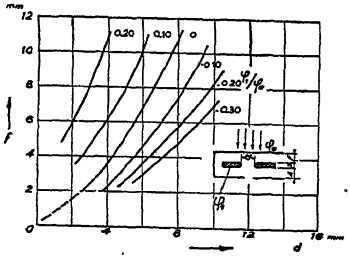


FIG. 32.—The three-electrode independent lens: focal length vs. central electrode voltage (Mahl[82]).

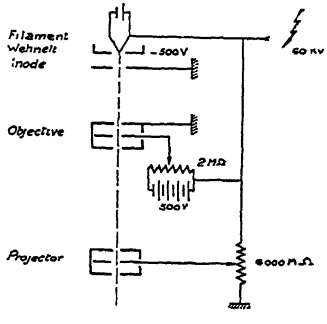


FIG. 33.—Use of nearly unipotential lens in the electrostatic microscope (Bruck[84]).

and Regenstreif[83]) may be obtained at 65 kv with interelectrode clearances of 4 mm. But a truly "unipotential" lens would have a fixed focal length and focusing would then have to be achieved by a very smooth displacement of the specimen because of the high magnification (from 10,000 to 50,000×).

This difficulty may be overcome (Bruck[84]) by the system represented in Fig. 33, which provides a useful variation of the potential of the central

electrode without spoiling the objective's precious insensitivity to sudden voltage fluctuations. Figure 32 shows how f depends on the two parameters, aperture diameter and potential of the central electrode (Mahl[82]). Other information may be found in Chamson,[85] Plass[86] and Ramberg.[9]

The same electrode arrangement may be used as a mirror, giving sharp images with a very small distortion (Zworykin,[2] Regenstreif[87] and Le Rutte[46]). Mirrors have been used successfully in the second stage of electrostatic microscopes (Mahl and Pendzich[13]).

3. *Immersion Lenses*

a. Cathode Microscope. The work of Johannson[88] may still be considered as fundamental since he thoroughly describes the main parameters (see other information by Einstein and Jacob[89]).

	Johannson (22)	Boersch (22)	Mahl (22)	Mecklenburg (22)
C, mm	Variable	1	0.2	0.2
First aperture, mm	1	1.3	1.8	2
Second aperture, mm	1.2	1.4	5	2.4
Voltage (kv)	1	30	20	30

Figure 34 shows the more recent cathode microscope electrode assembly, which finally gave a resolving power of 20 millimicrons (Brüche[90]).

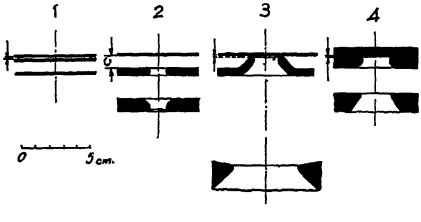


FIG. 34.—Cathode microscopes (90).

b. The Electron Gun in Cathode Ray Tubes. In the preceding application the cathode is heated just enough to give a current of about 10 μa. In oscillograph work, on the other hand, the current density is high and the space charge is of paramount importance. Reviews of the cathode-ray tube field have recently appeared in America (Morton[91]), in England (Moss[92]) and in Germany (Knoll[93]).

c. The Electron Gun in Microscopes. In electron microscopes, the oxide cathode has been discarded because of the poor vacuum (about 10^{-5} mm.) and a tungsten hair pin filament is used instead. The problem, then, is to apply a high field at the tip of the filament so as to obtain full saturated emission and to bunch the electrons in a narrow beam of 10^{-3} or 5×10^{-4} radians. Many proposals have been studied using two lenses (Hillier,[94] Ellis[95] and von Borries[96]). A comparable efficiency

has been obtained by Bruck and Bricka[38] with the immersion lens alone. Figure 35 shows the shape of the electrodes (the filament protrudes slightly beyond the negative electrode) and the efficiency of the gun. The current (i.e., the brightness of the images) may be regulated by adjusting the potential of the Wehnelt cylinder.

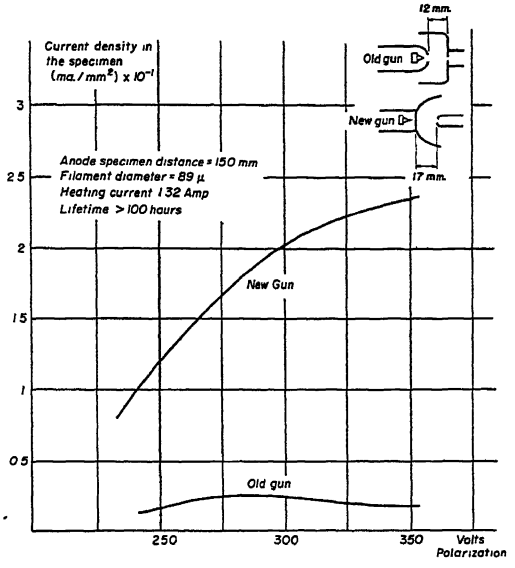


FIG. 35.—Bruck's[38] electron gun for microscopy.

4. *Magnetic Lenses*

a. Iron-Free Short Coils. These coils are used as a second lens in cathode-ray tubes and generally where a moderate convergence is sufficient. An old approximate formula (Ruska[99]) for the focal length is still useful

$$\text{A.T.} = 220\sqrt{\varphi_0 \frac{d}{f}} \tag{76}$$

(where A.T. stands for ampere turns, φ_0 is the accelerating voltage in kilovolts, f the focal length, and d the mean diameter of the coil).

Various experimental details and characteristics may be found in the references (97 and 98).

b. Iron-Clad Coils. Iron-clad coils give a more pronounced concentration of the field and a higher maximum value of the field for a given current. They are more economical. The same formula still applies for the focal length and d is now taken to be approximately the inner

diameter of the shield in the region of the gap. Glaser's model fits the experimental data rather well. References 97 and 98 are still valuable in this connection.

c. Microscope Lenses. The objectives particularly have been studied in great detail. The objective differs from the projector because of the conditions imposed by the specimen holder: the objective focus cannot be located as far into the bore of the lens as the projector focus and the focal length is thus longer. On the other hand, the important defects in the objective are the mechanical and aperture defects, whereas in the projector the pencils are very narrow and we have only the distortion

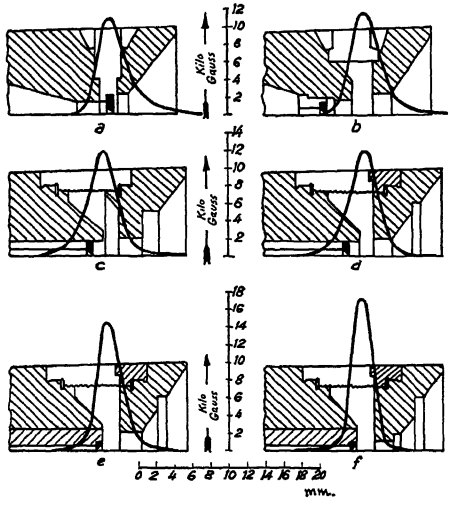


FIG. 36.—Ruska's[99] magnetic objective.

defect. Pole pieces are employed to give a strong concentration of the field in a small space.

Glaser's model permits a thorough calculation of the aberrations (Dosse[42]; Glaser and Lammel[44]); these results have been checked experimentally in the case of the aperture defect and distortion. This last defect may be corrected as indicated by Hillier.[45]

The same author has proposed an efficient corrector for the objective's mechanical and inhomogeneity defects and has succeeded in attaining a resolving power of 1 mμ by its use (Hillier and Ramberg[68]).

Microscopists endeavor to obtain as small a focal length as possible, a condition which keeps the effect of spherical aberration imperceptible, or at least minimizes it, and so gives the best resolving power. The maximum usable field intensity depends on the degree of saturation of the steel, and with cobalt steel we can reach 26,000 gauss. Saturation

broadens the field, and this effect is more pronounced with a small bore in the pole pieces.

This last effect may be important for high-magnetizing current, but does not materially change the general behavior of the lens, as may be inferred from the consideration of Glaser's model. In particular, when plotting the variation of focal length f or g against the parameter

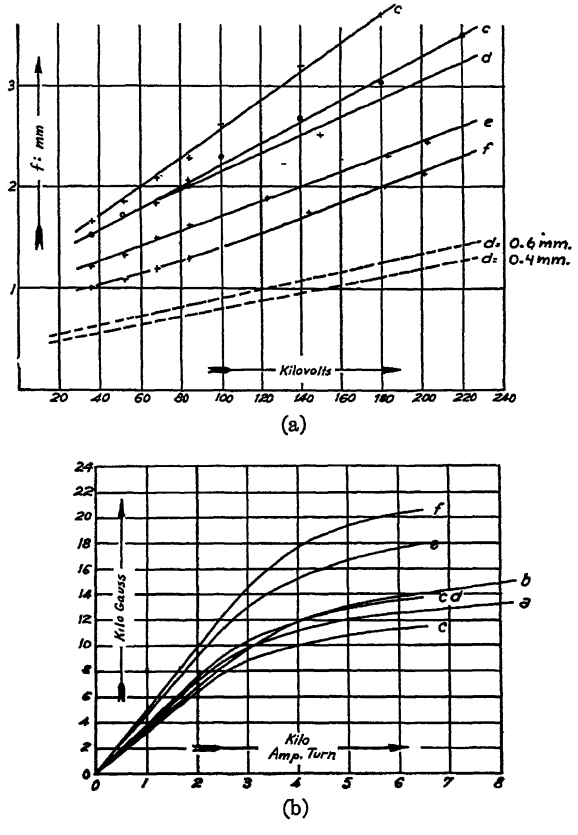


FIG. 37.—Optical (a) and magnetic (b) properties of Ruska's[99] objective.

$K^2 = ea^2B^2/8m\varphi_0$, we observe that f and g reach a minimum value. The lens is then insensitive to small variations of K^2, i.e., of the magnitizing current or of the gun voltage, which means the lens is "achromatic." This effect is much more pronounced for f (projectors) than for g (objectives) but is nevertheless observable and usable in the latter case; moreover, for such a value of K the pole pieces are nearly saturated and the magnetic heterogeneity is less pronounced. Therefore, objectives should be constructed to work under this condition.

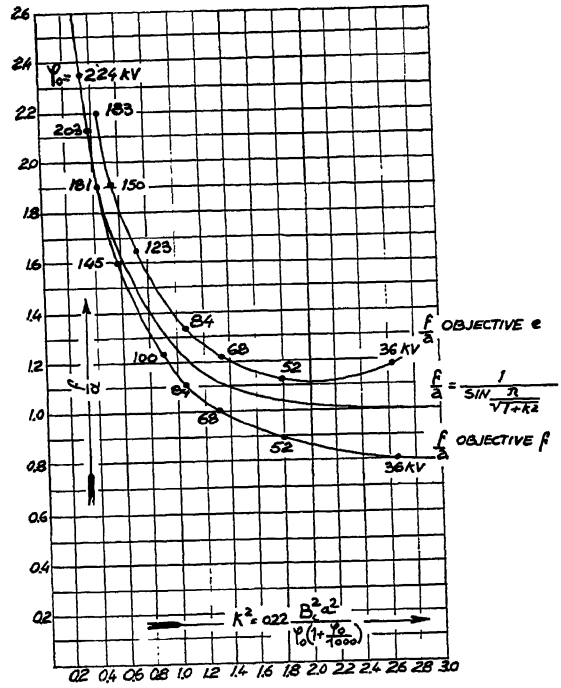


FIG. 38.—Focal length of Ruska's[99] objective, comparison with Glaser's model.

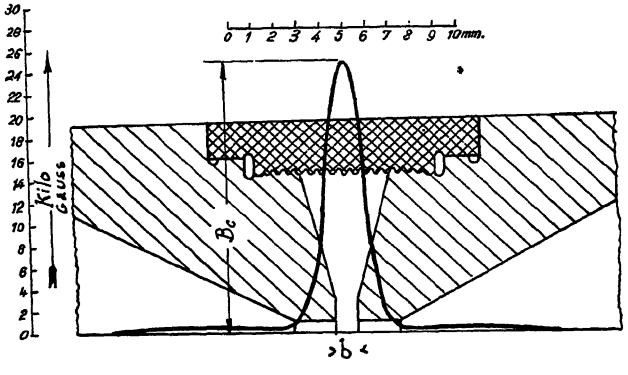
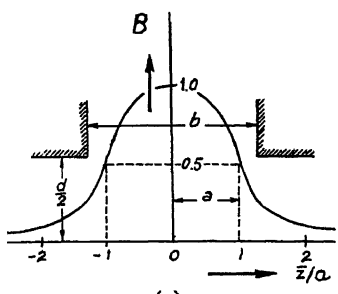
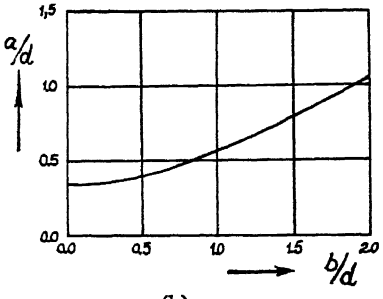


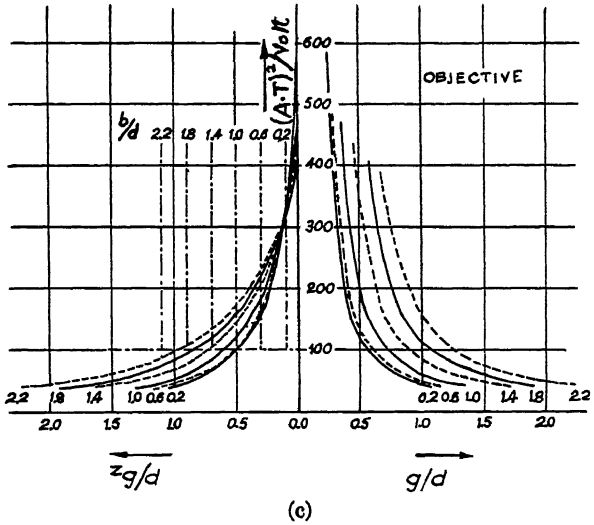
FIG. 39.—Ruska's[99] projector lens.



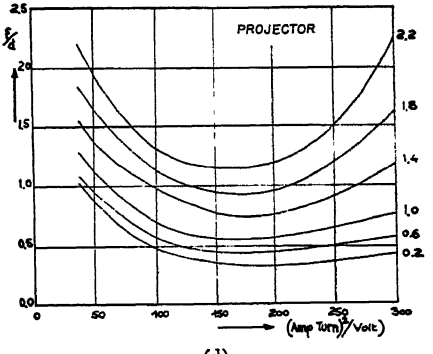
(a)



(b)



(c)



(d)

FIG. 40.—Le Poole's[100] graphs for the symmetrical magnetic lenses used in electron microscope.

Figure 36 shows the evolution of Ruska's[99] construction. Figure 37, a and b, gives the corresponding optical (a) and magnetic (b) properties (Glaser's model has properties intermediate to those of objectives *e* and *f*, as can be seen in Fig. 38).

The projector lens is symmetrical (Fig. 39) and its construction depends on only two parameters—the cone angle and the quotient of the gap width divided by the bore diameter. In any attempt to minimize *f*, which reduces the bulk of the instrument, the cone angle adjustment has a slight effect only, and may be taken between 70 and 75°. The optimum gap width depends on the electron velocity, on the bore diameter, and on the magnetizing current. The shape of $B(z)$ in such a lens is given in Fig. 40. Le Poole[100] has given convenient graphs to connect the half-width of the B curve to d, the diameter of the bore and b, the gap width (Fig. 40a and b). Figure (40c) gives the cardinal elements f and z_f (see Section I, 3, *b*) for such a lens used as projector (Fig. 40d) g and z_G, those for the same lens used as objective (see Section I, 3, *d* IC4).

BIBLIOGRAPHY

Modern Text Books

1. Cosslett, V. E. Introduction to Electron Optics. Clarendon Press, Oxford, 1946.
2. Zworykin, V. K., Morton, G. A., Ramberg, E. G., Hillier, J., and Vance, A. W. Electron Optics and the Electron Microscope. Wiley, New York, 1946.

REFERENCES

1′. Marton, L. *Rep. Phys. Soc. Progr. Phys.*, **10**, 204–252 (1946).
2′. Gabor, D. The Electron Microscope. Hulton Press, London, 1945.
3. Boersch, H. *Z. Naturforsch.*, **2a**, 615–633 (1947).
4. Recknagel, A. Beiträge zur Elektronenoptik (42–49) A. Barth, Leipzig, 1937.
5. Cartan, L. *J. phys. radium*, **18**, 111–120 (1937). Cartan, L., Comparat, P., and Thibaud, J. Technics of Today in Nuclear Physics. Gauthier-Villard, Paris, 1938.
6. Boersch, H. *Experientia*, **4**, 1–5 (1948).
7. Broadway, L. F., and Pearce, A. F. *Phys. Soc. Proc. Lond.*, **51**, 335–358 (1939).
8. Schaefer, H., and Walcher, W. *Z. Phys.*, **121**, 679–701 (1943).
9. Ramberg, E. G. *J. Appl. Phys.*, **13**, 582–594 (1942).
10. Recknagel, A. *Z. Phys.*, **120**, 331–362 (1943).
11. Mecklenburg, W. *Z. Phys.*, **120**, 21–30 (1942).
12. Kinder, E. *Z. Phys.*, **122**, 192–208 (1944).
13. Mahl, H., and Pendzich, A. *Z. techn. Phys.*, **24**, 38–42 (1943).
14. Glaser, W. *Z. Phys.*, **117**, 285–315 (1941).
15. Glaser, W., and Lammel, E. *Ann. Phys. Lpz.*, (5), **40**, 367–384 (1941).
16. Hutter, R. G. E. *J. Appl. Phys.*, **16**, 670–678 (1945).
17. Langmuir, D. B. *Proc. Inst. Radio. Engrs., N.Y. Wav. Electrons*, **25**, 977–991 (1937).
18. Pierce, J. R. *J. Appl. Phys.*, **10**, 715–724 (1939).

19. Pierce, J. R. *Bell Syst. Tech. J.*, **24**, 305–321 (1945).
20. Law, R. R. *Proc. Inst. Radio Engrs., N.Y. Wav. Electrons*, **25**, 954–976 (1937); **30**, 103–105 (1942).
21. Dosse, J. *Z. Phys.*, **115**, 530–556 (1940).
22. Jacob, L. *Phil. Mag.*, **28**, 81–98 (1939).
23. Klemperer, O., and Mayo, B. J. *J. Instn. Elect. Engrs.* (British), **95**, 135–142 (1948).
24. Pierce, J. R. *Proc. Inst. Radio Engrs., N.Y. Wav. Electrons*, **29**, 28–31 (1941).
25. Störmer, C. Numerical Integration Method for Ordinary Differential Equations. Congrès international des Mathematiciens, Strasbourg, 1920, pp. 242–257. Published in 1921 by the University Library, Toulouse.
26. Kryloff, A. N. *Mém. artillerie Franc.*, **6**, 353–423 (1927).
27. Mises von, R. *Z. angew. Math. u. Mechanik*, **10**, 81–92, (1 Feb. 1930).
28. Caratheodory, C. Geometrische Optik. Springer, Berlin, 1937.
29. Störmer, C. *Ann. Phys. Lpz.*, **16**, 685–696 (1933).
30. Becker, H., and Wallraff, A. *Arch. Elektrotech.*, **34**, 115–120 (1940).
31. Siegbahn, K. *Ark. Mat. Astr. Phys.*, **30A**, 1–12 (1943).
32. Sandor, A. *Arch. f. Elektrotech.*, **35**, 217–228 (4) and 259–287 (5) (1941).
33. Laplume, J. *Cahiers phys.*, No. 29–30, 55–66 (1947).
34. Gabor, D. *Nature, Lond.*, **159**, 303 (1947).
35. Hillier, J., and Ramberg, E. G. *J. Appl. Phys.*, **19**, 678–682 (1948).
36. Boersch, H. Communication No. 17. The electron velocity filter. 9th Electron Microscopy Conference, Cambridge, September, 1948.
37. Ramberg, E. G. *J. Opt. Soc. Amer.*, **29**, 79–83 (1939).
38. Bricka, M., and Bruck, H. *Ann. Radioélectricite*, **3**, 339–343 (1948).
39. Recknagel, A. *Jahrb. A.E.G. Forschung*, **7**, 15–22 (1940).
40. Bruck, H. *Cahiers phys.*, No. 24, 1–14 (1944).
41. Ardenne, M. von. *Z. Phys.*, **117**, 602–611 (1941).
42. Dosse, J. *Z. Phys.*, **117**, 722–753 (1941); **118**, 375–383 (1941).
43. Mahl, H., and Recknagel, A. *Z. Phys.*, **122**, 660–679 (1933).
44. Glaser, W., and Lammel, E. *Arch. Elektrotech.*, **37**, 347–356 (1943).
45. Hillier, J. *J. Appl. Phys.*, **17**, 411–419 (1946).
46. Le Rutte, W. A. 9th Electron Microscopy Conference, Cambridge, September, 1948. Communication B, to appear in *J. Appl. Sci. Res. Netherlands*.
47. Gabor, D. *Nature, Lond.*, **150**, 650–651 (1942).
48. Glaser, W. *Nature, Lond.*, **162**, 455–456 (1948).
49. Scherzer, O. *Z. Phys.*, **401**, 593–604 (1936).
50. Glaser, W. *Z. Phys.*, **116**, 19–33 (1940).
51. Recknagel, A. *Z. Phys.*, **117**, 67–73 (1941).
52. Rebsch, R. *Z. Phys.*, **116**, 729–733 (1940).
53. Siegbahn, K. *Phil. Mag.*, (7), **37**, 162–184 (1946).
54. Marton, L., and Bol, R. *J. Appl. Phys.*, **18**, 522–529 (1947).
55. Recknagel, A. *Jahrb. A.E.G. Forschung*, **7**, 15–23 (1940).
56. Ramberg, E. G. *J. Appl. Phys.*, **20**, 183–184 (1949).
57. Cotte, M. *Ann. Phys. Paris*, **10**, 333–406 (1938).
58. Wendt, G. *Z. Phys.*, **120**, 720–740 (1943).
59. Hachenberg, O. *Ann. Phys. Lpz.*, **2**, 516, 225–241 (1948).
60. Bruck, H. *C.R. Acad. Sci., Paris*, **224**, 1818–1820 (1947).
61. Glaser, W. *Z. Phys.*, **120**, 1–15 (1943).
62. Bertein, F. *C.R. Acad. Sci., Paris*, **224**, 560–562 (1941).
63. Bertein, F., Bruck, H., and Grivet, P. *Ann. Radioélectricité*, **2**, 249–253 (1947).

64. Bertein, F. *Ann. Radioélectricité*, **2**, 379–408 (1947); **3**, 49–62 (1948).
65. Bertein, F. *C.R. Acad. Sci., Paris*, **225**, 863–865 (1947); *J. phys. radium*, (8), **9**, 104–112 (1948).
66. Bruck, H., Remillon, R., and Roman, L. *C.R. Acad. Sci., Paris*, **226**, 650–652 (1948).
67. Cotte, M. *C.R. Acad. Sci., Paris*, **228**, 377–379 (1949).
68. Hillier, J., and Ramberg, E. G. *J. Appl. Phys.*, **18**, 48–71 (1947).
69. Bertein, F., and Regenstreif, E. *C.R. Acad. Sci., Paris*, **224**, 737–739 (1947).
70. Bruck, H., and Grivet, P. *C.R. Acad. Sci., Paris*, **224**, 1768–1769 (1947).
71. Bruck, H. *C.R. Acad. Sci., Paris*, **224**, 1628–1629 (1947).
72. Glaser, W. *Z. Phys.*, **121**, 647–665 (1943).
73. Bruck, H. *C.R. Acad. Sci., Paris*, **224**, 1553–1555 (1947).
74. Marechal, A. *Cahiers Phys.*, No. 26, 1–27 (1944); *Rev. Opt. Théor. Instrum.*, **26**, 257–277 (1947); **27**, 73–92 (1948); **27**, 269–287 (1948).
75. Bertein, F. *C.R. Acad. Sci., Paris*, **225**, 801–803 (1947).
76. Scherzer, O. *Optik*, **2**, 114–132 (1947).
77. Gabor, D. *Nature*, **161**, 777–778 (1948).
78. Hubert, P. *C.R. Acad. Sci., Paris*, **228**, 233–235 (1949).
79. Hutter, R. G. E. *J. Appl. Phys.*, **16**, 680–699 (1945).
80. Gundert, E. *Die Telefunken Röhre*, No. 19–20, 61–98 (1941).
81. Nicoll, F. H. *Proc. Phys. Soc., Lond.*, **50**, 888–898 (1938).
82. Mahl, H. *Jahrb. A.E.G. Forschung*, **7**, 43–54 (1940).
83. Grivet, P., and Regenstreif, E. 9th Electron Microscopy Conference, Cambridge, September, 1948. Communication No. 22.
84. Bruck, H., and Grivet, P. *Ann. Radioélectricité*, **2**, 244–249 (1947).
85. Chamson, P. *Ann. Phys., Paris*, (12) **2**, 333–413 (1947).
86. Plass, G. N. *J. Appl. Phys.*, **13**, 49–55 (1942); **13**, 524 (1942).
87. Regenstreif, E. *Ann. Radioélectricité*, **2**, 348–358 (1947).
88. Johannson, H. *Ann. Phys. Lpz.*, **18**, 385–413 (1933); **21**, 274–284 (1934).
89. Einstein, P. A., and Jacob, L. *Phil. Mag.*, **39**, 20–31 (1948).
90. Brüche, E. *Kolloid Z.*, **100**, 192–206 (1942).
91. Morton, G. A. *Rev. Mod. Phys.*, **18**, 362–378 (1946).
92. Moss, H. *J. Inst. Radio Engrs. (British)*, **5**, 10–26 (1945); 99–129 (1945).
93. Knoll, M., and Theile, R. *Fortschr. Hochfrequenztechnik*, **1**, 487–550 (1941).
94. Hillier, J., and Backer, F. *J. Appl. Phys.*, **16**, 469–483 (1945).
95. Ellis, S. G. *J. Appl. Phys.*, **18**, 879–891 (1947).
96. Borries, B. von. *Optik*, **3**, 321–377 and 389–412 (1948).
97. Soller, T., Starr, M. A., and Valley, E. G. M.I.T. Radiation Laboratory Series 110.22. Cathode Ray Tube Displays. McGraw-Hill, New York, 1948.
98. Proceedings of the Radiolocation Convention, 1946. *J. Instn. Elect. Engrs. (British)*, **93**, I, part. IIIa, 128–173; No. 5, 779–832.
99. Ruska, E. *Arch. Elektrotech.*, **38**, 102–130 (1944).
100. LePoole, J. B., and Mentz, M. *J. Appl. Sci. Res. Netherlands*, **1**, 3–17 (1947).

Field Plotting and Ray Tracing in Electron Optics
A Review of Numerical Methods

G. LIEBMANN

Research Laboratory, Associated Electrical Industries Limited, Aldermaston, Berkshire, England

CONTENTS

I. Introduction

"Electron Optics" is the study of the motion of charged particles, electrons or ions, in given electric or magnetic fields. Many of the fundamental investigations in electron optics have been based on simplified or approximate mathematical expressions for the electric or magnetic fields, such as to allow an easier mathematical treatment of the equations of motion of the particles. As long as the derivation of *general laws* was the object of the study, such simplifications were fully justified.

The main problem in applied electron optics is to arrange electric or magnetic fields in the best possible way to achieve a certain end; this end may be the attainment of high particle energies, or increased resolving power in electron microscopes, or improvements in cathode-ray tube performance, or simply the design of a new electronic device. It is, therefore, mostly necessary to study the fields produced by electrodes or pole pieces of a *given geometry*, and to determine the particle trajectories through them, or alternatively to *evolve the shapes* of electrodes or pole pieces which will lead to a desired trajectory.

A review of the methods developed for these purposes is the object of this article. Whilst the emphasis will be on the more recently published methods, an attempt at a more complete and rounded-off presentation of this particular branch of electron optics will be made. Methods which have been known for some time will therefore be included.

This review will not deal with the tracing of trajectories through time-varying fields, unless the period of variation is so slow compared with the time of flight of the particle that the fields may be considered as quasi-static, and the methods pertaining to static fields can be applied.

Most methods of trajectory tracing given below refer to axially symmetrical fields, as employed in electron lenses; the application of similar methods to plane fields, which occurs frequently, follows the same lines, and is not treated separately. Trajectories are determined for electrons only; the extension to the trajectories of ions will be obvious. Electromagnetic cgs units will be used.

The review falls logically into two main parts: the determination of fields and the tracing of electron trajectories through these fields. The automatic ray tracing methods, in which no separate information about the fields is sought, are treated in a separate further section.

II. Field Plotting

1. Properties of Static Electromagnetic Fields

We begin our discussion of the various methods of field determination by collecting in this section some fundamental equations for later reference.

The general properties of the electromagnetic field are expressed by Maxwell's equations. For static (or quasi-static) fields they reduce to:

$$\text{curl } \mathbf{H} = \frac{4\pi}{c}\mathbf{J} \tag{1}$$

$$\text{curl } \mathbf{E} = 0 \tag{2}$$

$$\text{div } \mu\mathbf{H} = 0 \tag{3}$$

$$\text{div } \epsilon\mathbf{E} = 4\pi\rho \tag{4}$$

We see from the structure of the eqs. (1–4) that the magnetic and electrostatic fields are superimposed without interaction. Our problem is the solution of these equations for given boundary conditions.

In view of eq. (2) we can write

$$\mathbf{E} = -\text{ grad } \phi, \tag{5}$$

where ϕ is the scalar electrostatic potential. Combination of eqs. (4) and (5) leads to Poisson's equation

$$\nabla^2\phi = -\frac{4\pi\rho}{\epsilon}. \tag{6}$$

If the field is free from electric charges, the last equation reduces to Laplace's equation

$$\nabla^2\phi = 0. \tag{7}$$

If $\mathbf{J} = 0$, one can represent the magnetostatic field by its scalar magnetic potential Ψ, with

$$\mathbf{H} = -\text{ grad}\,\Psi \tag{5a}$$

and

$$\nabla^2\Psi = 0 \tag{7a}$$

If $\mathbf{J} \neq 0$, we can write in view of eq. (3),

$$\mathbf{H} = \text{curl } \mathbf{A} \tag{8}$$

and

$$\text{div } \mathbf{A} = 0, \tag{9}$$

where $\mathbf{A}$ is the magnetic vector potential.

Equation (1) can then be written as

$$\text{curl curl } \mathbf{A} = \frac{4\pi}{c}\mathbf{J}. \tag{10}$$

This relation holds in any coordinate system. In a Cartesian system x,y,z, eq. (10) takes on the simple form

$$\nabla^2\mathbf{A} = -\frac{4\pi}{c}\mathbf{J}. \tag{11}$$

If we assume that all currents flow in the direction of the z-axis, we have $A_x = A_y = 0$ and $\frac{\partial A_z}{\partial_z} = 0$, and hence obtain Poisson's equation for a plane field:

$$\frac{\partial^2 A_z}{\partial x^2} + \frac{\partial^2 A_z}{\partial y^2} = -\frac{4\pi}{c} J_z. \tag{12}$$

Once this equation is solved the magnetic field components are found from

$$\left.\begin{aligned} H_x &= \frac{\partial A_z}{\partial y} \\ H_y &= -\frac{\partial A_z}{\partial x} \end{aligned}\right\} \tag{13}$$

Equations (5a) and (13) form a set of Cauchy-Riemann differential equations. The functions A_z and Ψ, both defining the magnetic field, are thus conjugate functions in the complex x,y-plane.

In axially symmetric fields we have $A_r = 0$, $A_z = 0$ and $\frac{\partial A_\theta}{\partial \theta} = 0$.

For the boundary conditions at the juncture of two media with dielectric constants ϵ_1 and ϵ_2, or permeabilities μ_1 and μ_2, one obtains:

$$\left.\begin{aligned} \epsilon_2 E_{2n} &= \epsilon_1 E_{1n} \\ E_{2t} &= E_{1t} \\ \mu_2 H_{2n} &= \mu_1 H_{1n} \\ H_{2t} &= H_{1t} \\ A_{2n} &= A_{1n} \\ \mu_2 A_{2t} &= \mu_1 A_{1t}. \end{aligned}\right\} \tag{14}$$

The suffix n denotes the normal component and the suffix t the tangential component of the field at the separating surface.

Hence, only the normal field component can exist at the surface of a good conductor or ferromagnetic material, whereas the vector potential at the surface of a ferromagnetic substance ($\mu \simeq \infty$) must be tangential.

2. *Analytic Methods*

a. Series Solutions. By the method of the "separation of variables" one can obtain solutions of Laplace's equation in x,y in the form:

$$\phi(x,y) = \sum_k (A_k \sinh kx + B_k \cosh kx)(C_k \sin ky + D_k \cos ky). \tag{15}$$

The corresponding solution of Laplace's equation in the cylindrical coordinates r, z is:

$$\phi(r,z) = \sum_k (A_k \sinh kz + B_k \cosh kz)(C_k J_0(kr) + D_k N_0(kr)), \quad (16)$$

where $J_0(kr)$ and $N_0(kr)$ are the zero-order Bessel and Neumann functions. The summations are taken over all possible values of k, the "separation constant."

The integration constants $A_k \ldots D_k$ are determined from the potential distribution along the boundaries. The solutions (15) and (16) are only useful in practice if it is possible to achieve the prescribed boundary values of the potential with a limited number of terms.

If the potential distribution is known to be symmetrical about an axis, one can express the potential everywhere in terms of the potential on the axis, and its derivatives:

$$\phi(x,y) = \sum_{n=0}^{\infty} (-1)^n \frac{1}{(2n)!} y^{2n} \phi(x)^{(2n)} \quad (17)$$

and

$$\phi(r,z) = \sum_{n=0}^{\infty} (-1)^n \frac{1}{(n!)^2 2^{2n}} r^{2n} \phi(z)^{(2n)} \quad (18)$$

Series solutions for the problem of equal diameter coaxial cylinders were obtained by Bertram with the help of the Fourier integral. In a first paper,[1] he replaced an intermediate series by an approximate expression which could be integrated, leading to this very useful approximate formula for the potential along the axis of two coaxial cylinders of unit radius, spaced by the distance s (Fig. 1):

$$\phi_0(z) = \frac{\phi_a + \phi_b}{2} + \left(\frac{\phi_a + \phi_b}{2.64}\right) \log_e \left[\frac{\cosh 1.32\left(z + \frac{s}{2}\right)}{\cosh 1.32\left(z - \frac{s}{2}\right)}\right] \quad (19)$$

In a second paper Bertram[2] tabulated series solutions $\phi(r,z)$ for boundary potentials of the form $\phi(1,z) = z^n$, $n = 0,1,2,3$. An application of Bertram's series to the problem of circular apertures and short cylinders, requiring a modified boundary potential, was recently given by G. Liebmann.[3]

A further discussion of the series solution for cylinders of unit radius is given in Chapter XI of the textbook by Zworykin *et al.*[4]

b. Conjugate Functions. A powerful method for finding solutions of Laplace's equation in x,y is the application of the theory of complex

functions. Let $f(w) = u(x,y) + iv(x,y)$, and $w = x + iy$. If $f(w)$ is any analytic function of w, then its real part $u(x,y)$ and its imaginary part $v(x,y)$ both represent solutions of Laplace's equation, the two solutions u and v being orthogonal. Thus if one can find an analytic function $f(w)$ of such kind that some of the curves u = constant (or v = constant) have the shape of the electrodes (or magnetic pole pieces) then the other lines u = constant represent the equipotential lines, and the orthogonal lines v = constant the lines of force (or flux lines). For instance, the function $f(w) = A \log w + B$ gives the solution for the field distribution between coaxial cylinders.

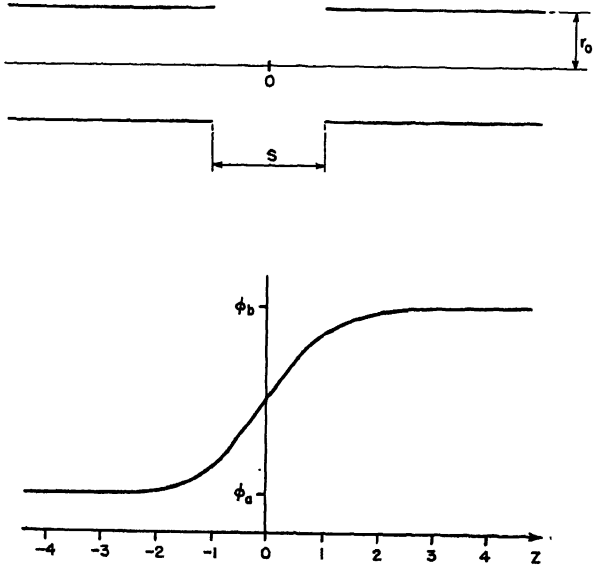


FIG. 1.—Potential distribution $\phi_0(z)$ along axis of equidiameter cylinders spaced distance s along axis, according to eq. (19) (Bertram).

Some potential fields represented by analytic functions useful in electron optics are discussed in Chapter XI of the book by Zworykin *et al.*[4] The place of the conjugate function is in cylindrical coordinates taken by Stokes's stream function. This was recently applied by Rüdenberg[5] to the development of desirable electrode shapes for electrostatic electron lenses.

c. Conformal Transformations. If the electrode arrangement is given, and the ensuing potential distribution is sought, it is often possible to transform the field in such a way that the original boundary is changed into a boundary which forms an equipotential line of a *known* potential distribution. It is shown in the theory of complex functions that any

analytic function $f(w)$, with $df(w)/dw \neq 0$, represents a "conformal" transformation which transforms equipotentials and flux lines in the f-plane into equipotentials and flux lines in the w-plane. Many such "mapping functions" have been investigated and are listed in the textbooks on the theory of complex functions and its applications (e.g., Churchill,[6] Ollendorff,[7] Rothe *et al.*,[8] and others). Once the potential distribution in the w-plane has been found, the transformation is reversed, yielding the potential distribution for the original electrode configuration in the f-plane. Well known applications of this method in electron optics are Vogdes and Elder's[9] derivation of an expression for the amplification factor and the maps of the electrostatic fields in radio tubes given in W. G. Dow's[10] textbook.

The important Schwarz-transformation, which maps the boundary of a closed polygon on to the real axis, and the interior of the polygon on to the upper half-plane, has been applied to many engineering problems because it allows the exact fitting of rectilinear boundary shapes. In electron optics, it has been applied by Herzog[11] to the screening problems in cathode ray tubes, and by Herne[12] to the derivation of the valve amplification factor for large grid wire diameters.

3. *Numerical Methods*

a. H. Liebmann's Procedure. Several numerical methods have been developed by which the potential distribution can be obtained for any boundary shape with any degree of accuracy. The earliest of these was given by H. Liebmann.[13] The plane region between the boundaries is divided by lines parallel to the x,y-axes into small squares of equal size (Fig. 2). Take any intersection point P_0 and its four nearest neighbors $P_1 \ldots P_4$. According to a well-known theorem of potential theory, the potential at the point P_0 is then the average value of the potentials at the points $P_1 \ldots P_4$:

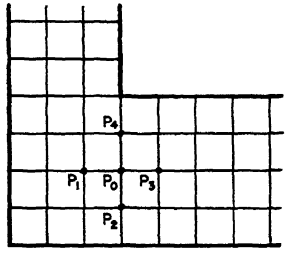


FIG. 2.—Diagram illustrating H. Liebmann's procedure.

$$\phi_0 = \tfrac{1}{4}(\phi_1 + \phi_2 + \phi_3 + \phi_4). \tag{20}$$

A computation according to H. Liebmann is carried out in this way: assume an arbitrary approximate potential distribution. Begin an improvement process at one boundary, using eq. (20). As this formula involves some exactly known boundary potential values, say ϕ_1, and some approximate assumed interior potential values, say ϕ_2, ϕ_3 and ϕ_4, the value ϕ_0 will be less arbitrary than the originally assumed potential

value. Carry on the averaging process over the whole region, and repeat the process over and over again until the potential values do not show any further changes. The described method is obviously very simple and requires little computational skill, but the convergence is very slow. Improvements were effected in its use by A. O. Mueller,[14] who increased the rate of convergence by "overshooting," and during the last few years, by Shortley and Weller,[15] who gave formulae for the simultaneous adjustment of "blocks" of points, and extended the method to irregular boundaries, and by Shortley *et al.*[16] Another proposal to increase the speed of the original method is due to Kormes,[17] who transfers the information on to punched cards which can be handled by modern business machines. Even in this way, the process remains fairly lengthy.

b. Southwell's Relaxation Technique. This method, which was developed by Southwell and his school, proves much faster and more powerful than the one just described. It is now widely used in many branches of physics and engineering. Only a very brief account can be given here; full details will be found in R. V. Southwell's[18] books. Shorter, concise presentations which deal with the principles of the relaxation method were published by Emmons[19] and more recently by Fox.[20] A more specific application to electron optics was made by Motz and Klanfer.[21]

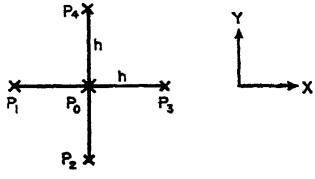


FIG. 3.—Diagram showing adjacent mesh points.

Consider a plane electrode structure, with rectilinear electrode boundaries parallel to the x,y-axes. Divide the interior into squares by equally distant lines, separation h, running parallel to the axes. Take one point P_0 (coordinates x_0,y_0, potential ϕ_0) at the intersection of two such dividing lines and call its neighbors P_1, P_2, P_3, P_4 (Fig. 3). Then

$$P_1P_0 = P_0P_3 = P_2P_0 = P_0P_4 = h$$

Using Taylor's series, we have for the potential ϕ_1 at P_1:

$$\phi_1 = \phi_0 - h\left(\frac{\partial\phi}{\partial x}\right)_0 + \frac{h^2}{2}\left(\frac{\partial^2\phi}{\partial x^2}\right)_0 - \frac{h^3}{6}\left(\frac{\partial^3\phi}{\partial x^3}\right)_0 + \frac{h^4}{24}\left(\frac{\partial^4\phi}{\partial x^4}\right)_0 \mp \cdots$$

Expressing the potentials ϕ_2, ϕ_3, ϕ_4 at the other neighboring points in the same way through the potential and its derivatives at the point P_0, and adding all four expressions, we obtain:

$$\phi_1 + \phi_2 + \phi_3 + \phi_4 - 4\phi_0 = h^2\left(\frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial y^2}\right) + \frac{h^4}{12}\left(\frac{\partial^4\phi}{\partial x^4} + \frac{\partial^4\phi}{\partial y^4}\right) + \cdots$$

All odd derivatives have cancelled out. Disregarding the higher order terms, as h can be assumed small, we have therefore:

$$\sum_{n=1}^{4} \phi_n - 4\phi_0 - h^2\nabla^2\phi_0 = 0 \tag{21}$$

provided $\phi_0, \phi_1 \ldots$ are the *exact* values arising from the boundary conditions and the prescribed values of $\nabla^2\phi = -4\pi\rho$. If we insert *approximate* values of ϕ, we shall have, instead, eq. (22):

$$\sum_{n=1}^{4} \phi_n - 4\phi_0 - h^2\nabla^2\phi_0 = \mathfrak{R}_0, \tag{22}$$

where $\mathfrak{R}_0 \neq 0$ is the *residual* at the point P_0. The residual $\mathfrak{R}$ is a measure of the deviation of the approximate solution from the exact solution. The improvement of a first arbitrary solution consists therefore in such adjustments to the potential value ϕ_0 that the residual $\mathfrak{R}_0$ is reduced. It is obvious from eq. (22) that a change of $\Delta\phi_0$ in ϕ_0 reduces the residual $\mathfrak{R}_0$ by $4\Delta\phi_0$; an increment of $\Delta\phi_0 = \frac{1}{4}\mathfrak{R}_0$ would therefore eliminate $\mathfrak{R}_0$. However, ϕ_0 takes the place of one of the ϕ_n in $\Sigma\phi_n$ in the four neighboring points $P_1 \ldots P_4$. The potential change $\Delta\phi_0$ at P_0 adds therefore the increment $\Delta\phi_0$ to the residuals $\mathfrak{R}_1 \ldots \mathfrak{R}_4$ at the neighboring points. Southwell expresses this symbolically by a "relaxation pattern" (Fig. 4), where the figures written in the circles represent the potential changes a unit increment in $\Delta\phi_0$ produces at these points.

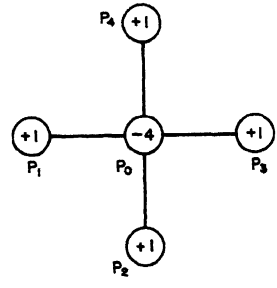


FIG. 4.—"Relaxation Pattern" (Southwell).

In this way one can easily see how the potential changes $\Delta\phi$ made at various points will influence the potential distribution. Mostly, one picks out those mesh points where the residuals are large, and reduces these below a desired limit. At the same time, it is often possible to foresee and to discount by "over-relaxation" those changes which later adjustments to neighbors will introduce. Various other short-cuts to lessen the labor of computation, like block and group relaxation, have been discussed in the literature.

In the case of Laplace's equation, the expression for the residual R_0 in Cartesian coordinates simplifies to

$$\mathfrak{R}_0 = \sum_{n=1}^{4} \phi_n - 4\phi_0 \tag{23}$$

For cylindrical coordinates r,z, one derives a similar equation

$$\mathfrak{R}_0 = \left[\phi_1 + \left(1 - \frac{h}{2r_0}\right)\phi_2 + \phi_3 + \left(1 + \frac{h}{2r_0}\right)\phi_4\right] - 4\phi_0, \qquad r_0 \neq 0, \tag{24}$$

where r_0 is the off-axis distance of point P_0; this goes over for points on the axis, where $\frac{1}{r}\frac{\partial\phi}{\partial r} \to \frac{\partial^2\phi}{\partial r^2}$, and $\phi_4 = \phi_2$ into:

$$\mathfrak{R}_0 = \phi_1 + 4\phi_2 + \phi_3 - 6\phi_0, \qquad r_0 = 0, \tag{25}$$

Curved boundaries can be treated by using "modified relaxation patterns," as was shown by Christopherson and Southwell.[22]

Recently Fox[23] pointed out that it is sometimes quicker to use a more complicated relaxation pattern operating at fewer points, by taking the fourth order term into account.

The ultimate accuracy obtainable by relaxation (or other numerical methods) depends on the mesh size used. It is easy to estimate the error caused by the finite mesh size and to extrapolate to a more accurate value if the process is carried out for two mesh sizes, say with h_a and $h_b = \frac{1}{2}h_a$, in each case until the residuals are small, and their sum over each reasonably large group of mesh points is small (the usual criterion for convergence for given mesh size). If the two potentials found in this way are ϕ_a and ϕ_b, then their difference $(\phi_a - \phi_b)$ gives a measure of the influence of the finite mesh size. An improved potential value ϕ^* can be obtained by extrapolation as

$$\phi^* = \phi_b - \tfrac{1}{3}(\phi_a - \phi_b); \tag{26}$$

this can be proved from the fact that $(\phi^* - \phi_b)$ is of the order of h^2.

4. Graphical Field Plotting

Graphical field plotting methods are based on the principle of the conservation of flux $\int_V \operatorname{div}(\epsilon E)dv = \int_V \operatorname{div}(\mu H)dv = 0$, and on the orthogonal properties of flux tubes and equipotential surfaces. These methods have been found very useful in the mapping of the magnetic flux in electrical machines, particularly as it is possible to map fields within current carrying regions. The ease with which approximate maps of the fields between *irregularly shaped electrodes* can be drawn makes them of some interest in electron optics.

Let the cross-sectional area of a tube of flux be ΔA, the potential increment between successive equipotential surfaces $\Delta\phi$, and their

distance Δl. From the conservation of flux we have then

$$\frac{\Delta l}{\epsilon \Delta A} = \text{constant} \cdot \Delta\phi \tag{27}$$

The cross-section ΔA can be written in the two-dimensional case as $\Delta W \times$ constant, ΔW being the width of the flux tube, whereas it will be $\Delta W \times r \times$ constant in the cylindrical case, r being the distance of the center of the curvilinear square from the axis (Fig. 5).

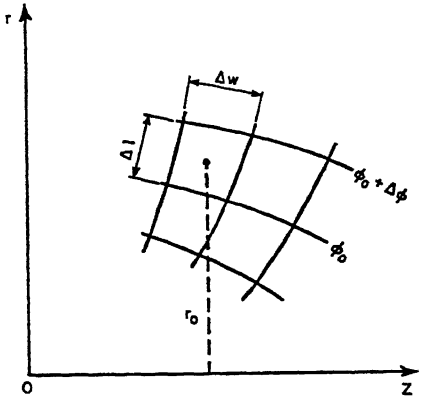


FIG. 5.—Elementary curvilinear square as used in graphical field plotting.

For a given potential increment $\Delta\phi$, the next equipotential line has therefore to be drawn such that (a) it is spaced by $\Delta l = k\epsilon\Delta W$ (or in cylindrical systems $\Delta l = k\epsilon r\Delta W$) from the preceding equipotential line, and (b) that it cuts the lines of flux at right angles. At the boundaries, the flux lines are of course perpendicular and the equipotential lines parallel to the boundaries. Such rules, as that the flux lines bisect

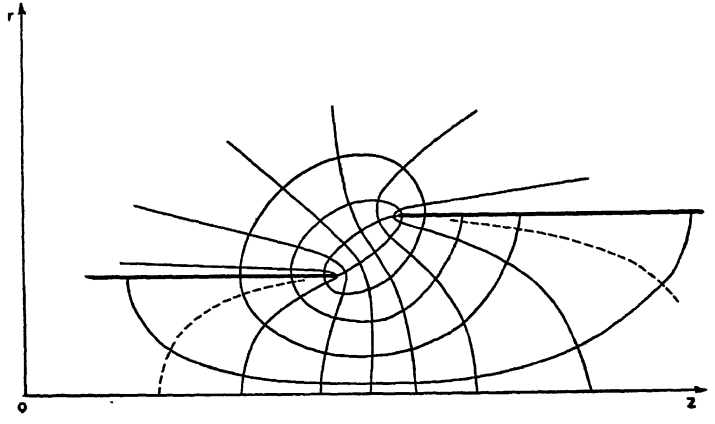


FIG. 6.—Field distribution between two axially spaced cylinders obtained by graphical field mapping.

angles of boundaries, and that equipotential lines approach projecting corners more closely, will be helpful in starting the sketching process, as will any other information drawn from known field configurations. It

will be found as a rule that in the first attempt it is not possible to fulfil correctly the requirement for the whole region that *all* areas between flux lines and equipotential lines are curvilinear squares obeying eq. (27). The field map has therefore to be redrawn with corrections, etc., until the required conditions are met everywhere. Accuracies of a few per cent can be obtained in this way. Further information will be found in the papers of Lehmann,[24] and of Kuhlmann,[25] who treated the axially symmetrical case, and in Moore's text book.[26] A recent discussion of the method was published by Willoughby,[27] and a brief review of more general graphical treatments was given by Poritsky.[28]

An application of the described method to the field between two cylinders is shown in Fig. 6.

5. *Experimental Methods*

a. The Electrolytic Tank. 1. *The Principle of the Method.* The idea of following equipotential lines with a probe to solve electric field problems is quite old; it was first used by Kirchhoff in 1845. An electrolytic tank was employed by Adams in 1875. The method has been further developed by numerous workers during the past forty years, and applied to many problems in electrostatics, hydrodynamics, and aerodynamics. We shall confine our discussion to the more recent work of interest in electron optics.

According to Ohm's law, the current density J is proportional to the electric field strength E:

$$\mathbf{J} = \sigma \mathbf{E} \tag{28}$$

where σ is the conductivity of the medium. Taking the divergence of eq. (1) we then have:

$$\operatorname{div} \mathbf{J} = \operatorname{div} \sigma \mathbf{E} = 0 \tag{29}$$

It is clear from the formal identity of eqs. (4) and (29) (for $\rho = 0$) that the flux tubes of electric current between electrodes immersed in a medium of conductivity σ are identical with the tubes of electric force between the same electrodes embedded in an insulating medium of dielectric constant ϵ, and that the orthogonal surfaces, i.e., the equipotential surfaces, are identical in the two cases. It follows from the linearity of the equations that a scaling up or down of the electrode structure or proportional changes of the electrode voltages leave the geometry of the pattern of flux tubes and equipotential surfaces unchanged.

To carry the analogy into effect, models of the electrodes are immersed into a tank filled with a weak electrolyte, e.g., tap water, taking advantage of symmetry conditions wherever possible, and appropriate voltages

are applied to them. The equipotentials can then be plotted with a probe of known potential.

A block diagram of the basic circuit is shown in Fig. 7.

The investigation of axially symmetric systems is often simplified if a tank with inclined insulating bottom is employed (Bowman-Manifold and Nicoll[29]). The two insulating planes, viz., the bottom of the tank and the free surface of the electrolyte, cut out a volume of electrolyte of sectorshaped cross section. The line of intersection of the two planes is the axis of the electron-optical system; the field distribution within the wedge of electrolyte is then the same, in view of the reflections in the insulating boundary planes, as that within a full cylinder. Another derivation of this principle, not based on symmetry considerations, but on the formal identity of Laplace's equation in cylindrical coordinates with eq. (29) was given by Pérès and Malavard.[30]

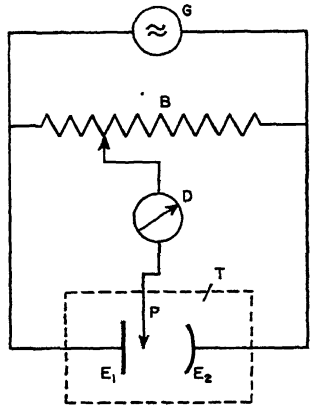


FIG. 7.—Basic circuit of electrolytic tank. (T, tank; E_1, E_2, electrodes; P, probe; G, generator; B, bridge potentiometer; D, balance detector.)

2. *Recent Electrolytic Tanks.* Details of improved electrolytic tanks were given by Hepp (1939),[31] by Himpan (1939, 1942),[32,33] by Zworykin *et al.*,[4] by Ertaud (1946),[34] by Willoughby (1946),[27] by Musson-Genon (1947),[35] and by Hutter (1947).[36]

As the design of these tanks is very similar, Hutter's tank only will be described, since it is the most recent one. His unit (Fig. 8) comprises a "deep" tank, 60 × 30 × 30 in. in size, and a "shallow" tank 5 × 30 × 30 in. Both tanks have insulated walls; the shallow tank can be tilted for use as an inclined-bottom tank. The model is supported by adjustable arms which slide on two flat rails running parallel with the longer side of the tank. The probe is the tip of a thin platinum wire projecting from a shielded glass tube; it is carried on a cross-slide which can be moved in two directions by hand operated gears. A marking stylus is rigidly connected to the probe slide and draws the equipotential lines on a large drawing board adjacent to the tanks.

In Himpan's[33] tank, the probe carrier, moved by a handle, can slide on a precision ground bar which swivels around a roller bearing pivot, combining ease of operation with accuracy of probe positioning. The probe location is indicated by a light spot marker on a glass plate which forms the lid of the tank. The equipotentials are then traced by hand.

While the vacuum tube voltmeter is used by most authors, Hepp[31] employs an additional phase meter to assist in the final balancing of his bridge potentiometer. Some workers favor a cathode ray tube balance detector as this indicates simultaneously amplitude and phase adjustment, and the absence of harmonics.

3. *Automatic Plotting.* Automatic marking of the equipotential lines on light sensitive paper by a light pointer attached to the probe was proposed by Ehrenfried.[37] The marker lamp is switched on through a relay operated by a vacuum tube voltmeter when the equipotential line—

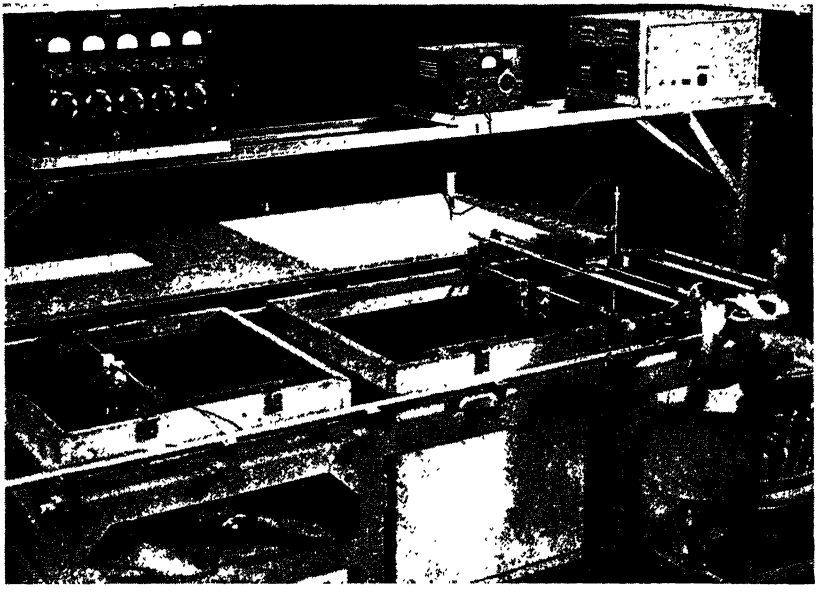


FIG. 8.—View of electrolytic tank (Hutter) (by courtesy of the *Journal for Applied Physics*).

to the potential of which the voltmeter had been adjusted—is crossed by the probe, which need only be moved about zig-zag wise, without balancing at each point.

Fully automatic plotting was recently achieved by Green.[38] In his arrangement, the motor driven probe travels at constant speed in the x-direction, being continuously adjusted in the y-direction by a servo system actuated by the voltage difference between a reference voltage and the probe voltage. When the probe reaches a "boundary strip," enclosing the area to be plotted, a switch operates a stepping relay which changes the reference voltage to that of the next equipotential, and

reverses the direction of travel of the probe slide. Therefore once started, the machine continues to follow in succession all the preselected equipotentials, which are plotted on a drawing board adjacent to the tank.

4. *Application of the Electrolytic Tank to Magnetic Problems.* Magnetic problems involving only the scalar magnetic potential are solved with the help of the electrolytic tank using the analogy discussed above. The application of the electrolytic tank to the determination of vector potential distributions was proposed by Pérès and Malavard[30] and by Peierls.[39]

From the discussion of the field properties, given in Section II, 1 (eq. (12) to (14)) follows that the vector potential distribution in two-dimensional systems can be plotted outside the current carrying conductors by feeding in currents proportional to J at the positions of the conductors and providing a "sink" for the current at infinity. In practice, it is only necessary to place the counter-electrode representing the sink near the wall of the tank a fair distance away from the region to be plotted, if it is shaped like an estimated flux line at this position. To satisfy the boundary conditions, eq. (14), iron ($\mu = \infty$) has to be represented by an insulator, e.g., a wax model of the iron circuit. The equipotential lines of the plot represent the magnetic field lines.

Peierls and Skyrme[40] recently showed that one can obtain the vector potential distribution A_θ for cases of axial symmetry from two field plots for a plane model. The practical application of Peierls's proposals is quite recent. Wilkinson[41] employed this principle to investigate the field distribution in the air gap of a betatron model, particularly in the vicinity of the exciting conductors. Bracher and Hide[42] used it to determine experimentally the optimum design of the pole faces of the new Birmingham proton synchrotron; Fig. 9 shows a view of the tank used and of the wax model.

5. *Field Plotting in the Electrolytic Tank under Space Charge Conditions.* The electrostatic potential distribution in the presence of continuously distributed space charge can be found experimentally in an electrolytic tank by a method developed by Musson-Genon.[35]

The space charge equation in x,y for a stream of electrons in a vacuum tube can be written as

$$\frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial y^2} = K\frac{i}{\sqrt{\phi}} \tag{30}$$

As the equation applying to the electrolytic tank is

$$\operatorname{div} \sigma\mathbf{E} = 0, \qquad \text{with} \qquad \sigma = ah.$$

$h = h(x,y)$ being the depth of the electrolyte, we can write

$$\operatorname{div} \mathbf{E} + \mathbf{E} \frac{\operatorname{grad} h}{h} = 0$$

This equation is formally identical with eq. (30), if we put

$$\mathbf{E} \operatorname{grad} \log h = -K \frac{i}{\sqrt{\phi}} \tag{31}$$

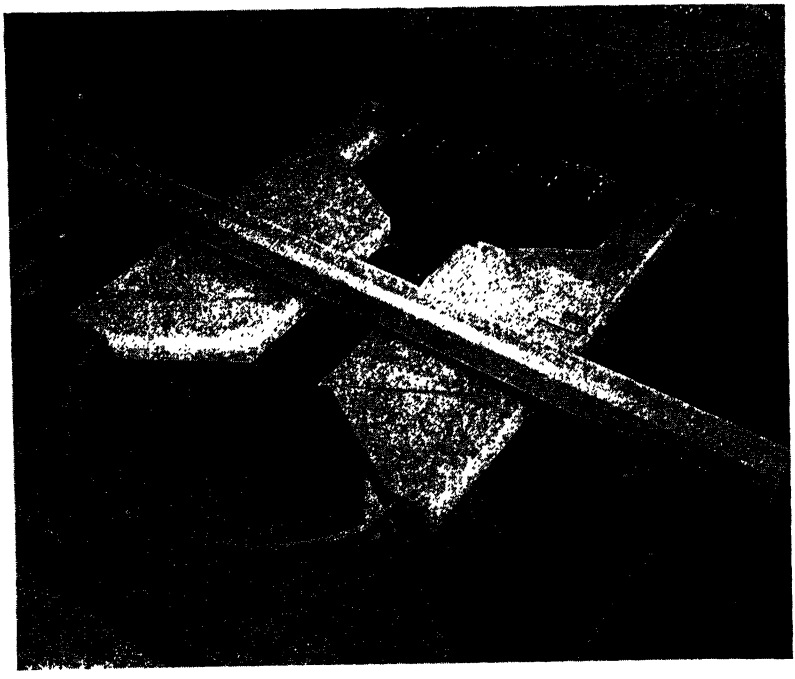


FIG. 9.—Photograph of Birmingham field plotting tank, showing wax model of magnet (Bracher and Hide). The current feeding conductors are seen in the background, the current "sink" is the bent strip in the foreground (by courtesy of Mr. Bracher).

Hence, if the depth h of the electrolyte is varied by shaping the bottom of the tank such that eq. (31) is fulfilled, a plot of the equipotentials at the surface of the electrolyte ($h = 0$) would give the desired potential distribution $\phi = \phi(x,y)$.

The difficulty that the equation for h also contains the unknown functions ϕ and $\mathbf{E} = -\operatorname{grad} \phi$ can be overcome by a method of successive approximations. As the first approximation, $h =$ constant. This yields potential and field strength distributions $\phi_0(x,y)$ and $E_0(x,y)$.

These are put into the approximate equation for h:

$$h = h_0 \exp\left[-K \int_{s_0}^{s} \frac{i}{E_0 \sqrt{\phi_0}} ds\right] \tag{32}$$

from which $h(x,y)$ is determined by numerical integration. Then a wax model of a "modulated" bottom, which establishes the desired height of the electrolyte, is made and inserted into the tank. The measurement of $\phi = \phi_1(x,y)$ and $\mathbf{E}_1 = -$ grad ϕ_1 is then carried out, from which an improved value of h can be found, and so forth. The correct conditions at the cathode are enforced by inserting there a conducting sheet under an angle of $3\pi/8$.

The method can be applied to the inclined bottom tank. If focusing occurs, changing the space charge distribution, a small number of trajectories have to be traced after each step of approximation; from this the redistribution of current density i, and hence of space charge, can be calculated.

6. *Optimum Working Conditions and Accuracy of Electrolytic Tanks.* The electrolytic tank cannot, for all its versatility, be classed as a precision apparatus. The estimates of its accuracy vary considerably. Willoughby,[27] for instance, stated recently: "considerable care is necessary to reduce errors to less than 5%." Several authors[43,44,45] gave accuracies of 2% to 3% whereas Hutter[36] claimed an accuracy of 1%. By selecting favorable electrode materials and electrolytes errors were reduced to the order of 0.5% by Theile and Himpan,[46] and by de Haller.[47] A recent investigation of the factors limiting the accuracy of measurements in electrolytic tanks by P. A. Einstein[48] points to a practical limit of 0.2%, if *extreme care* is taken to reduce *all* sources of error.

We shall discuss now the optimum working conditions and the main sources of error:

Mechanical accuracy. The machining and positioning of electrodes should be considered a mechanical precision job. Equally, the probe should be very fine and its position and depth must be accurately defined and be reproducible. Much of the work on tanks in recent years had as its aim the improvement of the mechanical accuracy.

Proximity of walls. Insulating walls represent reflectors, and conducting walls are equipotentials. Unless walls are made into an integral part of the model under investigation, it is necessary to keep the model well away from them. A flat-bottomed tank usable for strictly two-dimensional fields in which the wall proximity effect is greatly reduced was recently described by Makar *et al.*[49]

Electrode material. Many kinds of material have been used, copper, brass, or zinc being preferred, sometimes silver plated. Hepp,[31] working

with copper electrodes, found it advantageous to have them sand blasted. Investigations by Theile and Himpan[46] have shown that iron is better than any of these metals since the residual polarization effect ("transition resistance") that disturbs the field distribution in the tank is then reduced by a factor of 1.5 to 2 (see Fig. 10, which shows the error in the field distribution between plane parallel electrodes). A further improvement was obtained by Schmude and Schwenkhagen,[50] who applied a thin layer of colloidal graphite to clean brass surfaces; the graphite surfaces were then polished to a mirror-finish.

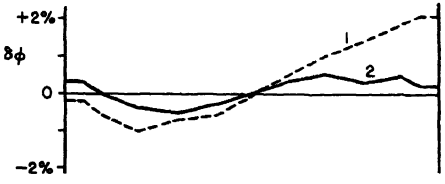


FIG. 10.—Error $\delta\phi$ in linear potential distribution due to residual polarization effects (Theile and Himpan). Curve 1: Ag electrodes in $n/2000$ H_2SO_4; curve 2: Fe electrodes in tap water (best of all investigated combinations of electrode material and electrolyte).

De Haller[47] gave this table for the "transition resistance" R_p for various electrode materials:

Material	R_p
Graphite, Pt sponge	Almost zero
Pt, polished	Very low
Au, Fe	Low
Ag	Noticeable
Cu	Fairly high
Ni, Zn, Sn	High
Stainless steel, Al	Very high

Probe material. The influence of the probe material on the accuracy obtainable is rather less than that of the electrode material. Polarization effects are only troublesome if the probe draws current, as in some measurements of gradients.

Electrolyte. Many kinds of weak electrolyte have been used. Theile and Himpan[46] found no appreciable difference in accuracy between tap water, $\frac{n}{1000}$ NaOH and $\frac{n}{2000}$ H_2SO_4. Schmude and Schwenkhagen used 0.5% NaOH or 0.5% $NaCO_3$ solutions with their graphited electrodes. Sometimes distilled water is preferred to reduce the current drain on the voltage source. Uniform temperature and uniform concentration are essential. According to de Haller,[47] a test of uniformity of

conductivity is advisable, carried out with the help of a test cell (Fig. 11) inserted into the tank.

Surface tension. Surface tension distorts the surface of the electrolyte at the electrodes, particularly if the liquid does not wet the electrodes evenly, and at the probe if this is not fine enough. In consequence, the field is distorted near the electrodes. Surface tension is also rather troublesome in inclined bottom tanks, reducing the accuracy of the measurements near the axis. The condition can be improved by tilting the tank to the contact angle of the electrolyte (Bowman-Manifold and Nicoll[29]) and by slightly grinding the insulated bottom near the axis, in a direction perpendicular to the axis (Klemperer[51]).

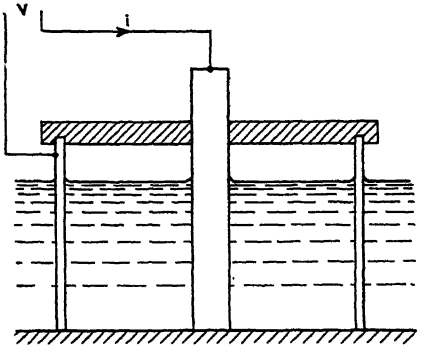


FIG. 11.—Test cell to check constancy of specific resistance of electrolyte in tank (de Haller).

Frequency of feeding voltage. To avoid serious polarization, AC has to be used. Even then polarization is not entirely eliminated; the residual effect decreases with $\sqrt{f}$ (Jones and Christian[52]; see also reference (49)), hence the higher the frequency f the better. However, stray capacity effects become more troublesome at higher frequencies, and a practical compromise is the use of frequencies in the range of 300 to 1000 cps.

Optimum voltage. Little information is available about this point; voltages of 4 to 50 volts rms are recommended in the literature.

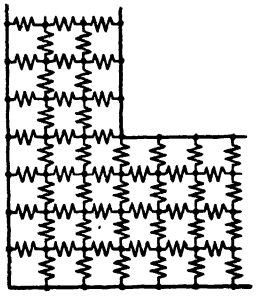


FIG. 12.—Principle of resistor network analogue.

Electrical disturbances. To obtain good accuracy, very sensitive detectors are necessary; the probe amplifiers and bridges used need therefore careful electrical screening (Himpan[32]). Willoughby[27] even recommends the use of a screened room. In his effort to increase the measuring accuracy, P. A. Einstein[48] found it necessary to use a Wagner earth bridge to obtain a true balance, mainly on account of the high probe impedance. The probe insulation must be very high to avoid errors due to variable leakage currents.

b. The Resistor Network. Instead of using the continuous distribution of a conducting medium in the electrolytic tank, one can represent a model of the electron-optical system by a network of resistances (Fig. 12).

Electrodes are simulated by short-circuiting (or open circuiting in plotting the magnetic vector potential distribution) the corresponding network points.

The operation of the network can be seen at once (Fig. 13) from Kirchhoff's law $\Sigma i_n = 0$, as $i_n = (\phi_n - \phi_0)/R$, with $n = 1, \cdots 4$, and R the value of the resistance units. Hence $\Sigma\phi_n - 4\phi_0 = 0$, which is identical with eq. (23), leaving a residual $\mathfrak{R}_0 = 0$ everywhere.

The usefulness of the resistor network in field plotting was only recognized a short time ago, when in 1943 Hogan[53] proposed this method. An apparatus following Hogan's proposals was constructed two years later by Redshaw.[54] In 1947 de Packh[55] described a resistor network for the approximate solution of Laplace's equation in cylindrical coordinates. Resistor networks for the plotting of field distributions were also briefly mentioned amongst the many types of impedance network analyzers proposed by Kron.[56] The drawback of these earlier resistance networks was their limited accuracy, which was given as between 2% and 5%.

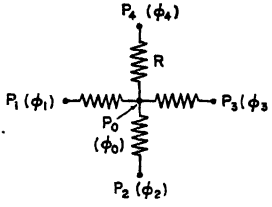


FIG. 13.—Resistor star at mesh point P_0.

G. Liebmann[57] succeeded recently in constructing a resistance network that gave a greatly improved accuracy. This network employs resistances graded to represent the case of axial symmetry; thus it may be looked upon as the analogue of an inclined bottom tank. There are 60 meshes in the z-direction and 20 meshes in the r-direction; the board is surrounded by an "infinity-strip" network which has the effect of moving the reflecting boundary further away from the axis. In problems where the whole size of the board could be utilized, and no high local field gradients appeared, accuracies of 1 to 2 parts in 10,000 were obtained, with only a few errors of the order of 3 to 5 parts in 10,000; these larger errors, due to local inaccuracies in the network were few and localized, and lent themselves to a quick elimination by subsequent relaxation.

If accuracies only of the order of 0.1% are required, no corrections at all are needed, and the field plotting can be accomplished very quickly, particularly if the knowledge of the field is only wanted along the axis. Boundaries which do not coincide with the rectangular mesh can be simulated to a high degree of accuracy by external shunting of the resistances at these boundary points. It is also possible to measure potentials within a mesh with the help of a clip-on potentiometer bridge.

By feeding in currents at the mesh points, and the use of an iteration process, Liebmann[57] could demonstrate that equations of the type $\nabla^2 f(r,z) = g(f,r,z)$ can be solved. As first step, an arbitrary distribution

f_0 of the sought function is assumed; e.g., f_0 may be a solution of Laplace's equation ($g = 0$) for the same boundary conditions. The currents are then set to represent $g(r,z) = g(f_0,r,z)$. This yields an improved function $f = f_1(r,z)$ which is now used instead of f_0, the fed-in currents being adjusted accordingly, and an improved function f_2 is found, and so on. The process is strongly convergent. The space charge equation $\nabla^2\phi = Ki\phi^{-\frac{1}{2}}$ is a special case of the type of equations that can be solved in this way, and Fig. 14 shows the successive approximations obtained in the experimental solution of the field distribution in a plane diode under space charge conditions.

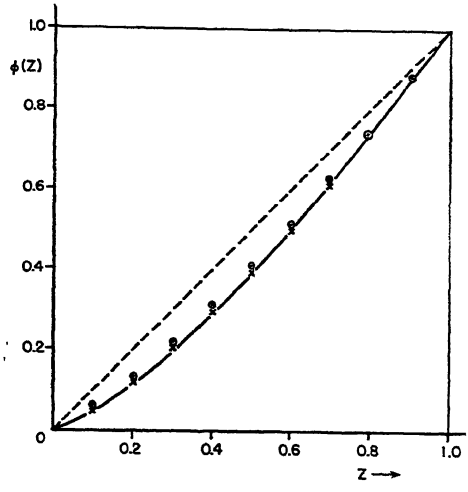


FIG. 14.—Normalized potential distribution in plane diode under space charge condition (by courtesy of *Nature, Lond.*).

Figure 15 is a photograph of the described resistance network.

c. Direct Measurement of Magnetic Fields. The plotting of magnetic fields by the methods discussed in the preceding sections applies only if $\mu = \infty$, or at least $\mu =$ constant. The appearance of iron saturation in magnetic electron lenses under some operating conditions often requires a direct measurement of the magnetic field distribution.

The methods used in electron optics follow standard practice (see review by McKeehan[58]); complications arise in certain types of electron lenses mainly on account of their small physical dimensions.

1. *The Flip Coil.* A coil of cross-sectional area A cm.^{2} and n turns, placed into a field of H oersteds, embraces a magnetic flux $F = AH$ lines. If the flux changes, an electromotive force $E = -10^{-8}n\frac{dF}{dt}$ volts

is induced, whence $H = \frac{10^8}{nA} \int Edt$. The time integral is usually measured with a ballistic galvanometer or a flux meter, the galvanometer-flip coil combination being calibrated in a standard coil.

The search coil is either pulled out of the lens field (Klemperer[59]) or blown out by an air jet (Simpson[60]), or the field is reversed, giving a deflection twice as great. This method was used by Marton[61] and by Dosse.[62] The search coil dimensions have to be very small for measure-

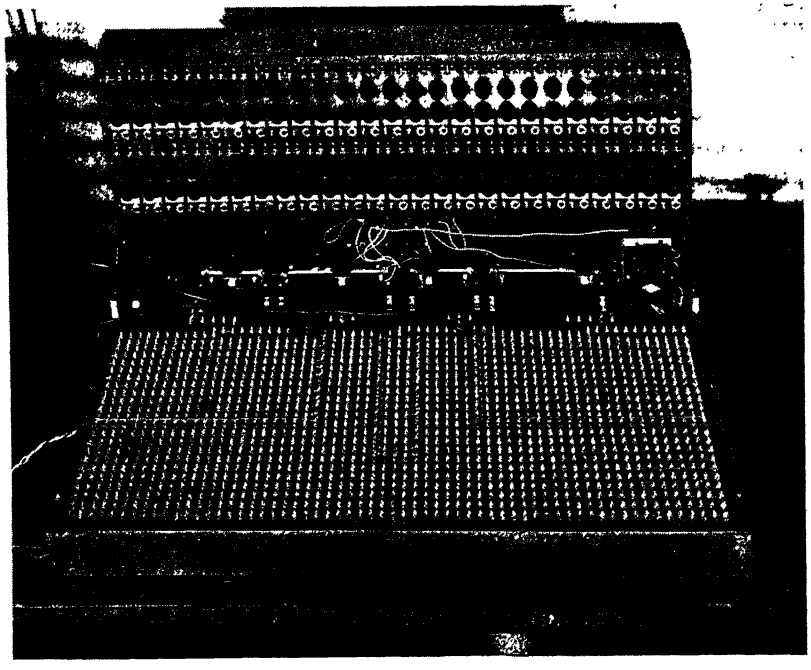


FIG. 15.—Resistance network for field plotting (Liebmann).

ments inside electron microscope lenses, and the mechanical precision in aligning search coil and lens has to be of a high order. Marton's coil had linear dimensions of 0.1 cm.; Dosse's coil was still smaller, being 0.04 cm. long and having a diameter of 0.03 to 0.04 cm., made from 100 turns of 0.002 cm. diameter wire. The reduction in coil diameter need not be as great if a set of coils with different diameters is used to plot $H(z)$ for different values of r, as used by Sandor.[63]

To overcome the errors due to the finite size of the coil in the measurement of inhomogeneous fields, Brown and Sweer[64] devised the "flux ball," a spherical search coil. For this shape the total flux is proportional to the field at the center of the coil. Brown and Sweer's coil had a diameter

of 3 inches. A much smaller "flux ball" coil, of $\frac{1}{4}$-inch diameter, was subsequently used by Williamson[65] to determine the field distribution in an electron microscope lens.

2. *The Rotating Coil.* A coil of area A and n turns, rotating evenly in a homogeneous magnetic field H such that the axis of rotation is perpendicular to the direction of the magnetic field, produces an emf: U = constant $\times$ $nA\omega H$, where ω is the frequency of rotation. Very small units, built like miniature armatures driven by synchronous motors, were described by Kohaut[66] and by Cole.[67]

3. *The Oscillating Coil.* An oscillating coil was employed by Klemperer and Miller,[68] and in improved form by Goddard and Klemperer,[69] to measure the field distribution in magnetic electron lenses of larger diameter. The coil of 0.2-cm. diameter moved to and fro through an angle of 180° at 10 cps, driven by a small reciprocating mechanism. The generated emf was amplified in a resonance amplifier.

4. *The Magnetic Balance.* A method for the measurement of the axial field distribution in small diameter magnetic lenses, which uses the very sensitive magnetic pendulum balance, was introduced by van Ments and Le Poole.[70] A very long, even coil is wound on a thin nonmagnetic bar which is suspended horizontally so that its axis coincides with the axis of the lens; the coil winding has somewhere a point of reversal P. The axial force acting on the coil is then proportional to the axial field strength H_p at the point P. The small horizontal displacement of the bar is read with a microscope. Figure 16 shows a photograph of the experimental arrangement.

5. *Electron-Optical Image Rotation.* The paraxial image formed by a magnetic lens is rotated by an angle

$$\theta = \left(\frac{e}{8m\phi}\right)^{\frac{1}{2}} \int_{z_0}^{z_i} H(z)dz,$$

where z_0 and z_i are the z-coordinates of object and image. This fact can be used, according to Zworykin *et al.*[4] to determine the axial field distribution by observing the "shadow image" of a straight edge, e.g., a fine stretched wire, on a fluorescent screen whilst the "object" is moved through the lens. $H(z)$ is then found by differentiating the measured θ curve with respect to z.

6. *Application of the Electron-Optical "Schlieren" Effect.* The electron-optical "Schlieren"-effect was recently applied by Marton[71] to the investigation of magnetic fields of very small extension. An electron optical "dark field" image is formed of an object, e.g. a steel wire, by placing a stop where the image of the electron source would be located. If the steel wire is now magnetized the fields near its surface will distort

its dark field image in a measurable way. Marton[72] subsequently modified the method by the superposition of a grid of fine wires in the plane of the stop; the magnetic field in a small area can then be determined from the distortion of the shadow images of these wires, super-

FIG. 16.—Magnetic balance for measuring axial magnetic field distribution in electron lenses (v. Ments and Le Poole) (by courtesy of Mr. Le Poole).

imposed on the image of the main object. Marton's method is also applicable to the investigation of electrostatic fields.

7. *Other Direct Methods.* Several other methods of magnetic measurement may become useful in electron optics owing to the small size of the probe.

A "Gaussmeter," based on the couple exerted on a small crystal of

siderose (CO_3Fe), due to the difference in its two principal susceptibilities, was described by Dupouy.[73]

Considerable progress in the application of the Hall effect was achieved recently by the use of germanium. Pearson[74] obtained an accuracy of $\pm 2\%$ for fields of 100 to 8000 Gauss with a probe $0.040 \times 0.5 \times 0.2$ in. Smaller probes appear feasible.

The incremental permeability of a very small "permalloy" transformer core was used by Gregg[75] to measure or monitor magnetic fields in betatrons.

Another novel apparatus for measuring localized magnetic fields is Kolin's[76] mercury jet magnetometer (see also Rhoderick[77]), in which the emf induced in a moving stream of mercury passing through the magnetic field is used as a measure of the field strength.

d. Direct Measurement of Electrostatic Field Distributions. An insulated probe gradually takes on the potential of the surrounding space. This principle was already applied by Maxwell to map electrostatic fields; potential equilibrium was indicated by a gold leaf electroscope. The method was improved in the course of years. Only the work of Semenoff and Walther,[78] and Walther and Inge,[79] who used an incandescent platinum probe, of Semenoff and Walther[80] who employed a capacity probe, and of Street[81] who used a vacuum tube voltmeter, will be mentioned here.

The advantage of these direct methods is that they allow the mapping of fields around apparatus under proper working conditions, while a serious disadvantage is the disturbance of the field distribution caused by the probe and its leads.

6. Evaluation of Field Plots

In several of the numerical trajectory tracing methods use is made of the values of the potential ϕ_0 on the axis and its derivatives with respect to z. For tracing a paraxial ray a knowledge of ϕ_0, ϕ_0', and often of ϕ_0'' is required, whereas the tracing of first order trajectories involves also ϕ_0''' and ϕ_0''''. In magnetic lenses $H_0 = -\psi_0'$ is needed for the paraxial ray, and $H_0' = -\psi_0''$ and $H_0'' = -\psi_0'''$ for the extraparaxial ray.

Differencing the axial potential is not accurate enough for the higher derivatives; the following method appears to give the greatest accuracy, as it makes use of as many experimentally or numerically determined points of the field distribution as possible. It is based on the series, eq. (18):

$$\phi(r,z) = \phi_0(z) - \left(\frac{r}{2}\right)^2 \phi_0''(z) + \frac{1}{4}\left(\frac{r}{2}\right)^4 \phi_0''''(z) \cdots$$

First derivative: difference ϕ_0 up to third differences, and find ϕ_0' from $\phi_0' = \frac{1}{h}(\Delta'\phi_0 - \frac{1}{6}\Delta'''\phi_0)$, h being the interval.

Second derivative (Gundert[82]): plot on a large scale $\phi(r,z)$ as function of $t = \left(\frac{r}{2}\right)^2$, for fixed value of z, using r values up to $r \simeq 0.3$ to 0.4 lens radius. These points should all lie on a straight line through the value $\phi_0(z)$; the slope gives the value of $-\phi_0''(z)$.

Third derivative: difference $\phi(r,z)$ with respect to z up to third differences, for fixed r, $r \simeq 0$ to 0.3 lens radius, and find $\frac{\partial\phi}{\partial z} = \Delta_z'\phi - \frac{1}{6}\Delta_z'''\phi$. Then plot $\frac{\partial\phi}{\partial z}$ as function of $\left(\frac{r}{2}\right)^2$, for fixed value of z. The intercept of the resulting straight line with the axis gives $\phi_0'(z)$ (a "smoothed" value, if the plotted points $\frac{\partial\phi}{\partial z}$ show any appreciable scatter around the straight line), and its slope gives $-\phi_0'''(z)$.

Fourth derivative: difference ϕ_0''' with respect to z; then $\phi_0''''(z) = \frac{1}{h}\Delta'\phi_0'''$. Higher differences would be meaningless.

Field plots need careful evaluation if a good accuracy in subsequent ray tracing is desired. An accuracy of one part in 1000 in the field determination appears necessary to obtain the paraxial focal length with 1% accuracy for strong electron lenses (i.e., focal length of the same order of magnitude as lens diameter), although this requirement can be relaxed for weak lenses roughly in proportion to the refracting power. A still higher accuracy is needed for the correct determination of the lens aberrations as ϕ''' is the dominant term in aberration calculations.

7. *Construction of Desired Field Distributions*

A problem frequently arising in the applications of electron optics is the determination of electrode shapes which produce a desired field distribution, e.g., along the axis. An experimental approach to this problem is Hutter's,[36] who equipped his electrolytic tank for this purpose with six probes, each connected to a vacuum tube voltmeter. These probes are arranged along the given equipotential line, or alternatively are connected to such voltages as desired at their respective locations. The electrode shapes are then modified until all indicators are balanced simultaneously; this can be done most easily in an inclined bottom tank.

A graphical method to achieve the same purpose, based on the difference equation equivalent to Laplace's equation, was proposed by Balachowsky.[83] The desired equipotential line, AE in Fig. 17, is made the

starting line of the process. If the lines AF and EF are drawn parallel to the z and r axes, the potential ϕ_F at F, the point of intersection, is given by

$$\phi_F = \phi_1 - \left(\frac{1/db + 1/ac}{1/c^2 + 1/rc + 1/b^2}\right)(\phi_0 - \phi_1) \qquad (33)$$

The notation used will be clear from Fig. 17, r being the distance of the point F from the axis. The graphical construction is continued until an equipotential line is reached which can be made into an electrode.

An analytic method for the solution of the problem, if the field along the axis is prescribed, was given by Glaser.[84] It requires the solution of an integral equation for the number of ampere turns/centimeter of the winding (or of the charge/unit length in the electrostatic case). The "kernel" of the integral equation is obtained through the Fourier transform of some known geometric functions.

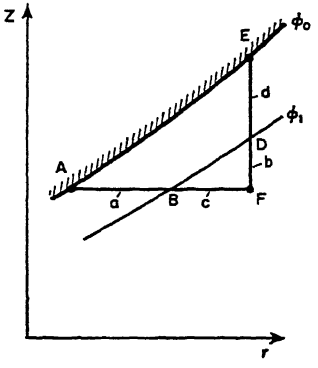


FIG. 17.—Balachowsky's method of constructing potential distributions.

III. RAY TRACING

1. Graphical Ray Tracing

A number of graphical methods for the determination of electron trajectories through given electrostatic fields, i.e., methods using purely geometric constructions, have been described in the literature. The simplest and quickest, but on the whole least accurate, method is the application of Snell's law, proceeding from one equipotential surface to the next one in a straight line, and changing direction in the ratio $\sqrt{\phi_2}/\sqrt{\phi_1}$ at each surface. This method was discussed by Jacob,[43] who considers that the final error in the position of the electron trajectory may be as great as 50%.

A simple protractor for the construction of approximate electron trajectories through a field in which the equipotential lines differ always by the same amount $\Delta\phi$, based on Snell's law, was recently described by Graffunder.[85] The potential of the nth equipotential line is then $n\Delta\phi$, and

$$(\sin \alpha_2/\sin \alpha_1) = [(2n - 1)/(2n + 1)]^{\frac{1}{2}}$$

Hence, if a series of concentric circles with radii $r_0\sqrt{2n - 1}$, $n = 1,2,3 \ldots$ is drawn on a sheet of celluloid (see Fig. 18), and the edge AA is

placed along the tangent of the equipotential line at point O, a line drawn parallel to the normal in O through the intersection P_1 of the trajectory in medium 1 with circle $r_0\sqrt{2n-1}$ will intersect the circle $r_0\sqrt{2n+1}$ in P_2. P_2O is the new direction of the trajectory in medium 2; this trajectory can be drawn easily as a straight edge is provided on the protractor which can swivel around point O.

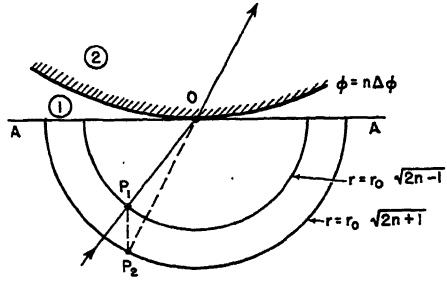


FIG. 18.—Principle of Graffunder's protractor.

Greater accuracy is obtained in Salinger's[86,87,88] "circle method," in which the electron path between subsequent equipotential surfaces is replaced by the osculating circle of the curved path, with radius $R = 2\phi/\text{grad}\ \phi$, provided the radius of curvature is not too large for accurate drawing with compasses.

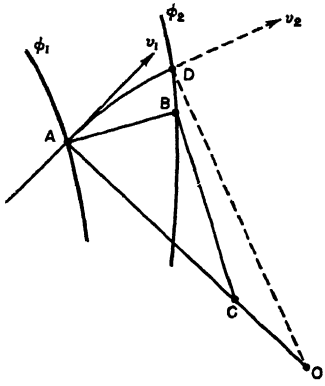


FIG. 19.—Construction of electron trajectory by the "circle method" (Salinger).

With reference to Fig. 19 the construction of the electron path by the circle method is as follows: let the direction of the incident ray be v_1, the ray intersecting the equipotential line ϕ_1 in the point A. Construct a perpendicular to ϕ_1 at A; this intersects the next equipotential line ϕ_2 in the point B. Erect perpendiculars to $\mathbf{v}_1$ at A and to $\overline{AB}$ at B, calling the point of intersection of the perpendiculars C. Then $\overline{AO} = R = [(\phi_2 + \phi_1)/(\phi_2 - \phi_1)]\,\overline{AC}$ determines the center O of the circle with radius R, lying on the extension of the line AC. Draw a circle with radius R about the center O from the point A until it meets the equipotential ϕ_2 in the point D. The circle segment AD is then the electron trajectory between ϕ_1 and ϕ_2. The line OD is also the perpendicular to the new velocity vector $\mathbf{v}_2$ in the point D on ϕ_2, etc.

The graphical construction giving, under the usual conditions, greatest accuracy appears the "parabola method," described fully by Zworykin and Morton.[88] Here, the path between two equipotential surfaces is replaced by the osculating parabola; the points of the parabolic trajectory can be constructed with compasses, but the construction appears more exact than that of the original "circle method" in cases where the curvature of the trajectory is not great. Thus the circle method and the parabola method are complementary, the one being most accurate where the other one is least exact, and vice versa.

The construction of the electron trajectory by the parabola method is as follows (see Fig. 20): let the initial velocity vector be again v_1; this cuts the equipotential line ϕ_1 in point A. Erect a perpendicular to ϕ_1 in point A, intersecting the equipotential line ϕ_2 in point B. Extend AB

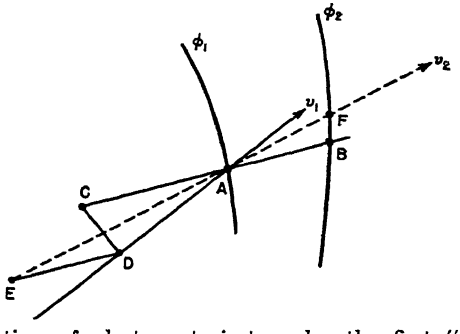


FIG. 20.—Construction of electron trajectory by the first "parabola method" (Zworykin and Morton).

back from A to point C by the distance $\overline{AC} = [(\phi_2 + \phi_1)/(\phi_2 - \phi_1)]\,\overline{AB}$. Drop a perpendicular from C on to the extended line of the vector v_1, intersecting this in the point D. Draw the line DE parallel to AB through point D, making $\overline{DE} = \overline{AB}$. The straight line from point E through point A is the velocity vector v_2, interesecting the equipotential line ϕ_2 in F. Point F is the intersection point of the electron trajectory with the equipotential line ϕ_2. The electron path between A and F would be an arc of a parabola.

A slightly simpler construction, using the properties of the parabolic electron path, was given by Hepp.[31] He employs an auxiliary diagram to find the direction of the chord of the parabolic trajectory A_1A_2 between two subsequent equipotential lines ϕ_1 and ϕ_2. To construct the electron path between points A_1 on ϕ_1 and A_2 on ϕ_2 (Fig. 21a) proceed thus: draw an auxiliary diagram, Fig. 21b, consisting of two concentric circles C_1 and C_2 of radii $\sqrt{\phi_1}$ and $\sqrt{\phi_2}$ around center O, a suitable unit of length to measure $\sqrt{\phi}$ having been chosen. Draw the line OB_1 parallel to the

velocity vector $\mathbf{v}_1$ of the incident electron ray, then draw the line B_1B_2 parallel to the normal A_1N to the equipotential surface ϕ_1 in the point A_1. The direction of the line OB_2 gives the direction of the new velocity vector $\mathbf{v}_2$ which can be used to continue the construction, proceeding to the next equipotential surface, etc. Bisect the distance B_1B_2 in the point M, and draw the line OM. The parallel to OM through the point A_1 represents the chord A_1A_2 of the parabolic electron trajectory, intersecting the equipotential line ϕ_2 in the point A_2.

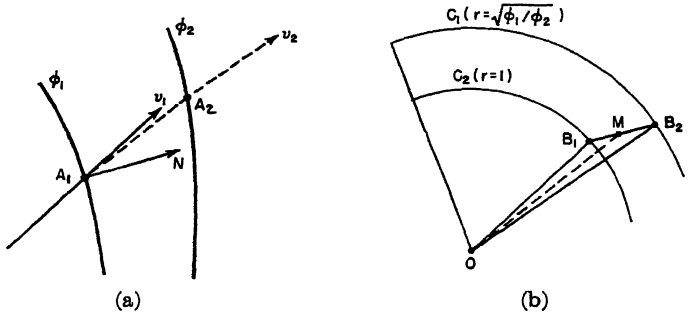


FIG. 21.—Construction of electron trajectory by the second "parabola method" (Hepp).

Selgin[89] recently developed another graphical method, based on the relations

$$\cot \alpha = \tfrac{1}{2}(\cot \theta_1 + \cot \theta_2),$$
$$\cot \theta_2 = [(\phi_2/\phi_1)(1 + \cot^2 \theta_1) - 1]^{\frac{1}{2}},$$

which can be proved from a consideration of the geometry of the parabolic electron path. α is the angle of the electron path with the x-axis at the point where it cuts the equipotential line ϕ_2, and θ_1 and θ_2 are the angles of the initial and final velocity vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ with the x-axis, which lies in the direction of the electrostatic field gradient. The graphical construction employs a chart of curves of cot θ_2 as a function of cot θ_1, which are hyperbolae, for several selected values K_1, K_2 . . . of the potential ratio (ϕ_2/ϕ_1). The axes of the chart are rotated by 45° with respect to the x-y-axes. The construction proceeds as follows (Fig. 22): the chart is placed against the equipotential line ϕ_1 such that the y-axis is tangent to ϕ_1 in the point A, the intersection of the electron path with ϕ_1, and that the point A coincides with the ordinate value $y = \sqrt{2}$. The initial velocity vector $\mathbf{v}_1$ is now extended until it cuts the "base line" of the chart in point B. The corresponding point C on the curve for cot θ_2, for the appropriate value (ϕ_2/ϕ_1), is obtained by following a line parallel to the cot θ_2 axis of the chart; the intersection of the orthogonal

line through C with the base line is D. The direction of the new velocity vector v_2 is then given by the direction of the straight line DA; the point of intersection F of the parabolic electron path from A with the equipotential line ϕ_2 is determined as the intersection of the straight line EA with the line ϕ_2, the point E bisecting the distance BD. Selgin's construction can be applied, with little modification, to retarding fields.

The tracing of the path of an electron starting from rest can often be made more accurate by using a thorem due to Rajchman.[90] According to this theorem the radius of curvature of the electron trajectory at a point of zero electron velocity is three times the radius of curvature of the line of force of the electrostatic field at this point.

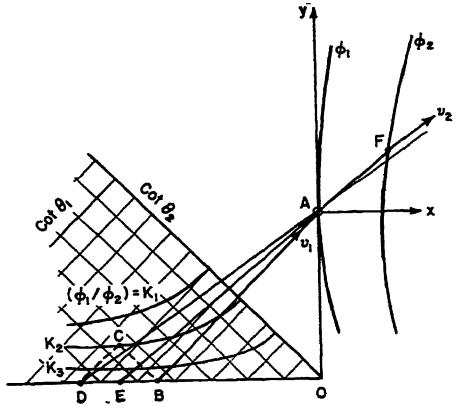


FIG. 22.—Construction of parabolic electron trajectory by Selgin's method.

Musson-Genon[91] discussed the accuracy of the various graphical methods at some length. He devised further graphical methods of constructing electron trajectories; some of these he considered more accurate but requiring a rather longer time to execute than the methods just described, and needing the help of an optical protractor designed by him.

Although Musson-Genon[91] takes a more optimistic view of the possibilities of graphical ray tracing, its accuracy appears rather limited to this writer. Even with great care in the execution of the constructions, errors of several per cent are likely in the final result. The graphical methods cannot, therefore, compete with the newer numerical methods when a high degree of precision is required, and they are of little use in the investigation of lens aberrations, where small differences between trajectories have to be found accurately. They are, however, valuable in tracing trajectories through complicated electrode structures, particularly structures that are not rotationally symmetric, provided limited

accuracy can be tolerated. A typical example of such application is the tracing of trajectories through the electrode system of certain types of electron multipliers.[92]

The graphical methods described apply only to space-charge-free electrostatic fields; they can be used for magnetic lenses if the lens field is replaced by Störmer's "equivalent potential" $U = \phi - \frac{e}{2m} A^2$, where $A = A_\theta$ is the magnetic vector potential (Dosse[93]). However, it appears that this method is inaccurate and laborious, and therefore of little practical interest.

A graphical method for the determination of electron trajectories through combined longitudinal magnetic fields and transverse electric fields varying slowly with time was given by Weimer and Rose.[94] This type of field arises in certain television pick-up tubes. Points on the trajectories are determined graphically as end points of a vector of constant length rotating with uniform velocity about a center, while the center of the circle described by the vector is translated in a parabolic path, both movements being superimposed.

2. Trigonometrical Ray Tracing

Klemperer and Wright[95] discussed a method of tracing an electron trajectory through electrostatic lenses which is the exact analogy of the trigonometric ray tracing used in light optics. The refraction of the ray passing with a given inclination at a given height through an equipotential surface, considered as a refracting surface of given curvature and given ratio n'/n of refractive indices, is determined with the help of the formulae developed for this purpose by the optical designers (see, e.g., Conrady's textbook[96]). The ray is then supposed to proceed in a straight line to the next equipotential surface, representing the next refracting surface to which the computing formulae are applied, etc. If about 15 to 20 steps are used, a fair degree of accuracy can be expected, but the method appears now superseded for most applications by the numerical methods mentioned later on.

3. Numerical Ray Tracing

a. The Basic Equations. The underlying idea of all numerical ray tracing methods is the approximation of the particle trajectory by a step-by-step solution of the differential equations of motion of the charged particle, which passes through the given electrostatic or magnetic fields with a given initial velocity. The various methods to be discussed differ merely in the technique employed, and in their suitability for particular problems.

It is obvious that the accuracy of a numerical method depends on the integration interval $\Delta z = z_{n+1} - z_n$ (see Fig. 23). The more efficient the process, the greater the interval may be for a prescribed accuracy. Similar considerations to those given at the end of Section II, 3, *b* apply to the extrapolation from computations repeated for different lengths of interval (for a further discussion of this point, see Duncan[97]).

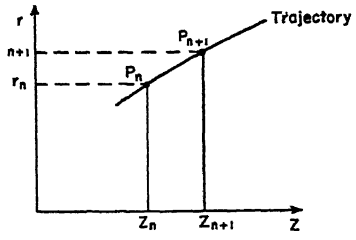


FIG. 23.—Diagram illustrating numerical step-by-step tracing of electron trajectories.

The general equation of motion of an electron in an electromagnetic field is

$$\mathbf{F} = m\mathbf{a} = -e(\mathbf{v} \times \mathbf{H}) + e \operatorname{grad} \phi \quad (34)$$

where $\mathbf{F}$ is the force acting on the electron, m its mass, $\mathbf{a}$ its acceleration, $-e$ its charge, and $\mathbf{v}$ its velocity. It proves convenient, in axially symmetric fields, to replace the magnetic field $\mathbf{H}$ by the vector potential $A_0 = A$ (see Section II, 1). The equations of motion can then be written in component form, dots denoting differentiation with respect to time (see Brüche and Scherzer[98]):

$$\left.\begin{aligned} \ddot{z} &= +\left(\frac{e}{m}\right)\frac{\partial U}{\partial z} \\ \ddot{r} &= +\left(\frac{e}{m}\right)\frac{\partial U}{\partial r} \\ \dot{\theta} &= +\left(\frac{e}{m}\right)\left(\frac{A}{r} + \frac{C}{r^2}\right) \end{aligned}\right\} \quad (35)$$

with

$$U = \phi - \frac{e}{2m}\left(A + \frac{C}{r}\right)^2 \quad (36)$$

A, ϕ and therefore U, the "equivalent potential," are functions of z and r; C is a constant determined by the initial angular momentum of the particle. As $C \neq 0$ only for skew rays, it is usually permissible to put $C = 0$. Elimination of the time from eqs. (35) with the help of the energy relation $\frac{m}{2}(\dot{z}^2 + \dot{r}^2 + r^2\dot{\theta}^2) = e\phi$ yields the *general ray equation*

$$r'' - \frac{1 + r'^2}{2U}\left[\frac{\partial U}{\partial r} - r'\frac{\partial U}{\partial z}\right] = 0 \quad (37)$$

and a first order differential equation for the rotation of the electron about the axis:

$$\theta' = \left(\frac{e}{2m}\right)^{\frac{1}{2}} \frac{A}{r} (1 + r'^2)^{\frac{1}{2}} U^{-\frac{1}{2}}. \tag{38}$$

Primes denote differentiation with respect to z.

Equations (37) and (38) can be interpreted as describing the path of an electron moving in a plane which rotates about the axis by the angle θ.

If only electrostatic fields are present ($A \equiv 0$) eq. (37) can be written as

$$r'' - \frac{1 + r'^2}{2\phi}\left[\frac{\partial \phi}{\partial r} - r' \frac{\partial \phi}{\partial z}\right] = 0 \tag{37a}$$

For $r \to 0$, $r' \to 0$, eq. (37) reduces to the *paraxial ray equations;* then for electrostatic lenses:

$$r'' + \frac{r'}{2}\left(\frac{\phi'_0}{\phi_0}\right) + \frac{r}{4}\left(\frac{\phi''_0}{\phi_0}\right) = 0, \tag{39}$$

and for magnetic lenses:

$$r'' + \frac{eH_0^2}{8m\phi} r = 0 \tag{39a}$$

In deriving (39a) from (37) use was made of the series development of the vector potential A:

$$A(r,z) = \frac{r}{2} H_0(z) - \frac{r^3}{16} H_0''(z) \pm \cdots$$

The paraxial image rotation in magnetic lenses is then given by

$$\theta' = \left(\frac{e}{8m\phi}\right)^{\frac{1}{2}} H_0 \tag{40}$$

Equations (37) to (40) are to be integrated between the initial position z_0 and the final position z_i, which may be the image of z_0, for the given initial values of r_0, r_0' and $\theta_0 = 0$ at $z = z_0$.

For the solution of eqs. (38) and (40) only a simple numerical integration need be performed, by using any of the known formulae for numerical quadrature. The solution of the differential equation (37) will be discussed in the next sections.

b. Numerical Ray Tracing Based on the Paraxial Ray Equation. The earliest method for the numerical integration of electron trajectories in magnetic fields is probably due to Störmer[99] who developed it for his investigations of the Aurora Borealis. Although it is of general interest, it was apparently not used in later electron optical investigations, and it will therefore not be discussed here.

Maloff and Epstein[100] appear to be the first writers who applied numerical methods to the solution of the paraxial ray equation (39).

They write $\frac{\phi_0'}{2\phi_0} = -2A$ and $\frac{\phi_0''}{4\phi_0} = -B$, assuming A and B to be constants over each of the z-intervals into which they divided the lens. The solution of eq. (39) can then be written as

$$r = \frac{r_n' - r_n m_2}{2\sqrt{A^2 + B}} \exp(m_1 z) - \frac{r_n' - r_n m_1}{2\sqrt{A^2 + B}} \exp(m_2 z),$$

where r_n and r_n' are the known values of r and r' at the beginning of the nth interval, and $m_1 = A + \sqrt{A^2 + B}$ and $m_2 = A - \sqrt{A^2 + B}$; z is reckoned from the beginning of the nth interval. From the above formula for $r = r(z)$, r_{n+1} at the end of the interval and, by differentiation, r_{n+1}' at the end of the interval can be obtained and used as the initial values for the next interval, and so forth.

In a modified form of their method, Maloff and Epstein[100] write $\frac{\phi_0'}{2\phi_0} = -(a + bz)$ and $\frac{\phi_0''}{4\phi_0} = -(p + qz)$, a, b, p, and q being constants in the nth interval, and then put r in the form of a Taylor series; the higher derivatives of r needed in this series are obtained by differentiating eq. (39).

An approximate path integration method is due to Gans.[101] He assumes that the curve representing the axial potential $\phi = \phi_0$ as function of z can be replaced by straight line sections. Then for each section $\phi'' = 0$, and eq. (39) reduces to $r'' + \frac{\phi'}{2\phi} r' = 0$, which can be integrated to give $r' = C \times \phi^{-\frac{1}{2}}$, and integrated once more between the points P_n and P_{n+1}, the end points of the nth straight line section giving $r_{n+1} = r_n + (2C_{n+1}/\phi_{n+1})(\phi_{n+1}^{\frac{1}{2}} - \phi_n^{\frac{1}{2}})$. At the junction of two segments $\phi'' = \infty$, but as ϕ, ϕ', r and r' all remain finite and continuous, it is permissible to integrate across the junction. Gans's method is rather quick but yields only approximate trajectories. Gans had already proposed to improve his method by replacing the straightline sections by parabola sections. This "parabola method" was further developed by Recknagel,[102] but is cumbersome when applied numerically.

Another numerical method, which is probably the simplest and quickest one as it only involves numerical evaluation of fairly simple integrals but which is then restricted to electrostatic lenses, is due to Spangenberg and Field.[87] It is based on the principle that electrons will move in paths normal to the surfaces of constant action, the action function a being given by $a = \int \phi^{\frac{1}{2}} ds$, or $\left(\frac{\partial a}{\partial r}\right)^2 + \left(\frac{\partial a}{\partial z}\right)^2 = \phi$. The action function a can be written as a power series in r:

$a = p + qr^2 + sr^4 + \cdots$. Then $p'^2 = \phi_0$ and $16q^2 + 8p'q' = -\phi_0''$. Numerical integration of the last equation over the nth interval, from P_n to P_{n+1}, yields

$$q_{n+1} = (q_n - \tfrac{1}{8}\phi_0^{-\frac{1}{2}}\phi_0''z)(1 + 2q_n\phi_0^{-\frac{1}{2}}z).$$

The slope of the electron path is $r' = 2q\phi_0^{-\frac{1}{2}}r$, and the electron path can be determined by further numerical integration, provided the ray does not cross the axis within the lens.

The paraxial image formation in certain magnetic lenses was investigated by Siday,[103] who integrated eq. (39a) numerically under the assumption that over the interval from P_n to P_{n+1} the value of r, appearing as a factor in eq. (39a), can be considered constant, i.e., $r = \frac{1}{2}(r_{n+1} + r_n)$. Then a first integration yields

$$r_{n+1}' - r_n' = \tfrac{1}{2}(r_{n+1} + r_n)\int_{z_n}^{z_{n+1}} kH^2(z)dz,$$

the integral being evaluated numerically. As further the approximate equation $r_{n+1} = r_n + \frac{1}{2}\Delta z(r_{n+1}' + r_n')$ applies, two equations for the two unknowns r_{n+1} and r_{n+1}' are available, the values of r_n, r_n' and $H(z)$ being known. Thus, the trajectory can be followed through the lens step by step, from P_n to P_{n+1}, etc. The assumption r = constant over the integration interval makes the method more suited for thick but rather weak lenses; for strong lenses, the steps Δz have to be made very small to approach the required condition, involving considerable computational effort, if a fair degree of accuracy is desired.

Dosse[104] applied the fact that the electron describes a parabolic path in a uniform transverse field to the numerical determination of electron trajectories through electrostatic and magnetic lenses. He considered the parabolic path through a thin section of an electrostatic lens with a potential distribution $\phi(r,z) = \phi_0(z) - \frac{1}{4}r^2\phi_0''(z)$ (see eq. (18)) where the forward velocity component of the electron is given by $v_z = \left(\frac{2e}{m}\phi_0\right)^{\frac{1}{2}}$ and the radial velocity component by $v_r = \sum \Delta v_r$. The radial acceleration is $\frac{e}{m}E_r$, with $E_r = -\frac{\partial\phi}{\partial r} = \frac{r}{2}\phi_0''$, and therefore $\Delta v_r = \frac{e}{m}E_r\Delta t = \frac{e}{2m}\phi_0''r_{n+\frac{1}{2}}\Delta t$. Hence, $\Delta z = v_z\Delta t$ and $\Delta r = v_r\Delta t + \frac{1}{4}\frac{e}{m}\phi_0''r_{n+\frac{1}{2}}\Delta t^2$. Thus starting from point P_n with coordinates r,z the point P_{n+1} of the trajectory is reached with coordinates $r + \Delta r$, $z + \Delta z$. The value $r_{n+\frac{1}{2}}$, required in the expression for r is approximately the half-way value of r between P_n and P_{n+1}, obtained from $r_{n+\frac{1}{2}} = r_n + \frac{1}{2}v_r\Delta t$. Trajectories

through magnetic lenses are calculated in the same way, after expressing the magnetic field through the equivalent potential U (eq. (36)). From the paraxial trajectories, the aberrations can then be calculated by numerical integration of expressions given by Glaser.[105]

The paraxial ray equation (39) or the corresponding equation for the paraxial trajectory through combined electrostatic and magnetic fields, derived from eq. (37), was put by Picht[106] into a very simple form, through the transformation $\rho = r\phi_0^{\frac{1}{2}}$:

$$\rho'' = -T\rho, \tag{41}$$

with

$$T = \frac{3}{16}\left(\frac{\phi_0'}{\phi_0}\right)^2 + \frac{eH^2}{8m\phi_0}.$$

By introducing the new variable $d = \frac{\rho'}{\rho}$, eq. (41) can be transformed into a linear differential equation:

$$d' = T + d^2 \tag{42}$$

T and d being functions of z.

Equations (41) and (42) were taken by Ramberg[107] as the starting point for the investigation of several types of strong electron lenses by numerical ray tracing. Equation (42) is integrated from point to point with the help of Taylor's series, for given initial conditions, yielding $d = d(z)$. The actual electron trajectory $r(z)$ is then obtained by further integration from the relation $\rho = \rho_0 \exp\left[-\int_{z_0}^{z} d\,dz\right]$, and finally from $r = \rho\phi^{-\frac{1}{2}}$. A difficulty arises where the trajectory approaches the axis as $d(z) \to \infty$ for $z \to z_i$; this would require very small integration intervals but the substitution $-\rho'/\rho = \tan y$, proposed by G. Archard, overcomes the difficulty. Numerical comparison on a few test cases has shown that the method outlined by Ramberg is capable of accurate results if the z intervals are chosen small enough but it is slower, for equal accuracy, than some of the methods discussed in the next section.

The lens aberrations are obtained as in the case of Dosse's parabola method by inserting the r-values of the paraxial trajectory into Glaser's[105] or Scherzer's[108] aberration formulae, and performing the required numerical integration.

An interesting method of numerical tracing of the paraxial electron path through a magnetic lens was recently given by van Ments and Le Poole.[70] These authors derived series solutions for the distance r of the trajectory from the axis at a point z, and for the reciprocal focal length $1/f$ in terms of the lens excitation k^2:

$$r = r_0(1 - a_1k^2 + a_2k^4 - a_3k^6 + \cdots)$$

and

$$1/f = b_1k^2 - b_2k^4 + b_3k^6 - \cdots,$$

where the $a_1, a_2, a_3 \ldots$ and $b_1, b_2, b_3 \ldots$ are functions of the coordinate z, and

$$k^2 = \frac{ea^2H_m^2}{8m\phi},$$

a being the half-width of the field and H_m the maximum field.

Thus, once the functions $a_1, a_2, a_3 \ldots$ and $b_1, b_2, b_3 \ldots$ are found for the various values of z, the value of k^2 can be determined which would bring the focus to a particular value of z by equating the series expression for r to zero, using the $a_1, a_2, a_3 \ldots$ values referring to this particular value of z. By using the same z value to determine the $b_1, b_2, b_3 \ldots$ values in the series for $1/f$, the focal length can then be determined. Alternatively, the trajectory can be plotted for given k^2 using the series for r. This method is therefore very useful if the dependence of lens properties on the excitation is of main interest, but the evaluation of the coefficients $a_1, a_2, a_3 \ldots, b_1, b_2, b_3 \ldots$ for the many z-coordinates needed and the subsequent use of the series are fairly laborious if the whole trajectory is required. The coefficients are given by these integrals:

$$b_1(z) = \int_z^\infty \left(\frac{H(\zeta)}{H_m}\right)^2 d\zeta,$$

$$a_1(z) = \int_z^\infty b_1(\zeta)d\zeta,$$

$$b_2(z) = \int_z^\infty a_1(\zeta)\left(\frac{H(\zeta)}{H_m}\right)^2 d\zeta,$$

$$a_2(z) = \int_z^\infty b_2(\zeta)d\zeta, \text{ etc.}$$

Over the mostly used range of excitations ($0.2 \leq k^2 \leq 1$), coefficients up to a_4 and b_4 are needed for an accuracy of $\frac{1}{2}\%$.

The most straightforward method of integrating the paraxial ray equations numerically was given by Cosslett,[109] who applied the Taylor series directly by putting:

$$r_{n+1} = r_n + r_n'\Delta z + \tfrac{1}{2}r_n''(\Delta z)^2,$$

and

$$r_{n+1}' = r_n' + r_n''\Delta z.$$

r_n'' is found from the initial values r_n and r_n' at the beginning of the nth interval ($z = z_n$) with the help of the paraxial ray equations (39) and (39a), and is then inserted in the above series expressions, whence r_{n+1} and r_{n+1}' are determined. The method is quite fast and fairly accurate, except in the case of strong electron lenses, where the interval Δz must be chosen rather small to avoid systematic errors. Its simplicity proves

very attractive in all computations that do not require the highest accuracy.

c. Numerical Ray Tracing Based on the General Ray Equation. Methods have been given for the numerical integration of the general ray equation, or the general equations of motion; the aberrations are then directly obtained from the trajectories. Such methods, due to Motz and Klanfer, and Goddard and Klemperer, are therefore applicable outside the range of the second order theory.

Before describing them, a recent extension by G. Liebmann[110] of the direct Taylor series approach to the second order approximation of the general ray equation (37) will be mentioned. In this, the Taylor series in terms of the equivalent potential $U(r,z)$ are taken up to and including the fourth order terms, and are then developed in terms of the axial potential ϕ_0 and the axial magnetic field strength H_0, and the derivatives of ϕ_0 and H_0. The result can be written in the form of a set of recurrence formulae:

$$r_{n+1} = Q_1 r_n + Q_2 r_n',$$
$$r_{n+1}' = Q_3 r_n' + Q_4 r_n,$$

where the coefficients $Q_1 \ldots Q_4$ are functions of ϕ_0 and H_0; for instance:

$$Q_1 = 1 - \frac{1}{8}\left(\frac{\phi_0''}{\phi_0} + \frac{eH_0^2}{2m\phi}\right)(\Delta z)^2 + \frac{1}{48}\left(\frac{\phi_0'}{\phi_0}\right)\left(\frac{\phi_0''}{\phi_0} + \frac{eH_0^2}{2m\phi_0}\right)(\Delta z)^3.$$

The deviation of the extra-paraxial trajectory from the paraxial trajectory can be expressed by a similar set of recurrence formulae:

$$\Delta r_{n+1} = Q_1 \Delta r_n + Q_2 \Delta r_n',$$
$$\Delta r_{n+1}' = Q_3 \Delta r_n' + Q_4 \Delta r_n' + Q_5.$$

The coefficients $Q_1 \ldots Q_4$ are the same as before, but the new coefficient Q_5 contains the values of r_n and r_n' of the paraxial trajectory, and the higher derivatives of ϕ_0 and H_0. The coefficients $Q_1 \ldots Q_4$ are very simple expressions for magnetic lenses, depending only on $(eH_0^2/2m\phi)$ and Δz. It is claimed for this method that it is so far the fastest numerical method for the assessment of first order aberrations of magnetic lenses.

The exact general ray equation for electrostatic lenses, eq. (37a), was integrated numerically by Motz and Klanfer,[21] using a general method for the solution of second order differential equations given by Comrie. The potential $\phi(r,z)$ and its derivatives $\frac{\partial \phi}{\partial z}$ and $\frac{\partial \phi}{\partial r}$ have to be tabulated, for suitable intervals of Δz and Δr, for the region through which the trajectory passes. The integration procedure is then as follows:

Tabulate the values of $r, f = 2\Delta z r', F = (2\Delta z)^2 r''$ and the first, second,

and third central differences d', d'', d''' of r (e.g., $d_{\frac{1}{2}}' = r_1 - r_0$, $d_1'' = d_{\frac{3}{2}}' - d_{\frac{1}{2}}'$, . . .) and the first and second central differences Δ', and Δ'' of F for at least four initially known points of the trajectory $P_0 \ldots P_3$. To proceed from point P_3 to the next point P_4 on the trajectory, estimate the next second difference Δ_3'' from the table (e.g., $\Delta_3'' \simeq \Delta_2''$, or what the trend of the differences may indicate) and calculate d_3'' from this formula: $d_3'' = \frac{1}{4}(F_3 + \frac{1}{12}\Delta_3'' \ldots)$. Then build up $d_{\frac{7}{2}}'$ and r_4 from the differences d_3'', $d_{\frac{5}{2}}'$ and from r_3. This value of r_4 is provisional and is checked by calculating $f_2' = 2\Delta z r_2'$ from the formula $f_2 = 2d_2' - \frac{1}{6}(2d_2''') + \cdots$; if agreement is not good enough, adjust r_4 and the differences to the correct value. Then form f_4 by using an extrapolation formula: $f_4 = f_2 + F_3 + \frac{1}{6}\Delta_3'' - \cdots$, and compute $F_4 = (2\Delta z)^2 r''$ from eq. (37a), and differences $\Delta_{\frac{7}{2}}'$ and Δ_3''; this completes the advance of the trajectory by one interval Δz from point P_n to P_{n+1}. Then proceed to the next point by estimating next difference Δ_4'' etc., following the procedure just described.

The extent of aberrations in the lens can be judged by tracing several trajectories and comparing them with a paraxial trajectory.

The equations of motion (35) were directly integrated by Goddard[111] and Goddard and Klemperer,[69] with the help of a formula due to W. E. Milne:

$$z_{n+1} = z_n + z_{n-2} - z_{n-3} + \frac{\Delta t^2}{2}(5\ddot{z}_n + 2\ddot{z}_{n-1} + 5\ddot{z}_{n-2})$$

$$r_{n+1} = r_n + r_{n-2} - r_{n-3} + \frac{\Delta t^2}{2}(5\ddot{r}_n + 2\ddot{r}_{n-1} + 5\ddot{r}_{n-2})$$

The values of $\ddot{z}_n$, $\ddot{r}_n$, etc. are taken from eqs. (35). The required field terms $\frac{\partial U}{\partial z}$ and $\frac{\partial U}{\partial r}$, or in the case of pure magnetic lenses, A, $A\frac{\partial A}{\partial z}$ and $A\frac{\partial A}{\partial r}$ are tabulated, according to Goddard and Klemperer,[69] for suitable intervals Δr and Δz, for those regions through which the trajectory passes. The four initial values needed (z_{n-3}, z_{n-2}, z_{n-1}, z_n) are obtained from a Taylor series. The final solution is given in parametric form.

Goddard and Klemperer used in part of their work a more accurate but more complicated integration formula of Milne, which is correct to sixth differences, but the extra work arising from the use of this formula will be rarely rewarded owing to the much lower accuracy with which the field data are usually known.

Whilst trajectory plotting by Goddard and Klemperer's method is slower than by some of the other described methods, it appears to be the best numerical method yet devised for rays that pass through the electron-

optical system so far from the axis that the system can no longer be described in terms of paraxial rays and primary aberrations.

4. Application of the Differential Analyzer

Most work in the past was done either by graphical or by numerical ray tracing, and the advances made in automatic computing techniques in recent years are only now beginning to be applied to electron optics. A differential analyzer was used by Blewett *et al.*[112] for the determination of the paths of electrons in a split anode magnetron, and in the study of transit-time effects in triodes. The machine was set up for the direct integration of the equations of motion of the electron in two dimensions; the necessary information about the field was continuously fed in by hand by an operator who read the required values of potentials and field strengths from "field maps" at the position indicated by the plotting stylus, representing the position of the electron in the field. The field maps were prepared by the usual plotting techniques, as described in Section II of this review.

IV. Automatic Trajectory Tracing

1. The Rubber Membrane

Several ingenious devices have been described that automatically reproduce the trajectories of electrons through models of electrode systems, without separate field plotting and subsequent ray tracing.

The device which has so far proved most useful is Kleynen's[113] rubber membrane. It can be shown that the height $h = h(x,y)$ of a rubber membrane, which is stretched tightly over a horizontal frame and whose surface is slightly deformed by placing models of the electrode system under it, satisfies Laplace's equation (7), provided $\frac{\partial h}{\partial x} \ll 1$ and $\frac{\partial h}{\partial y} \ll 1$. Hence, the height distribution h represents a relief model of the potential distribution if the heights of the electrodes are adjusted to be proportional to the boundary potentials. Kleynen could also prove that point masses sliding without friction over the stretched membrane under the force of gravitation satisfy, in first approximation, the equations of motion of charged particles in an electrostatic field; hence, such point masses will trace out the trajectories of the particles. The sliding point masses are replaced, in practice, by small rolling steel balls. The trajectories appear as a series of small dashes, produced by photographing the movement of the highly reflecting steel balls in the light of a high pressure mercury arc lamp operated on 50 cps. AC. The length of the individual dashes gives an indication of the relative particle speed.

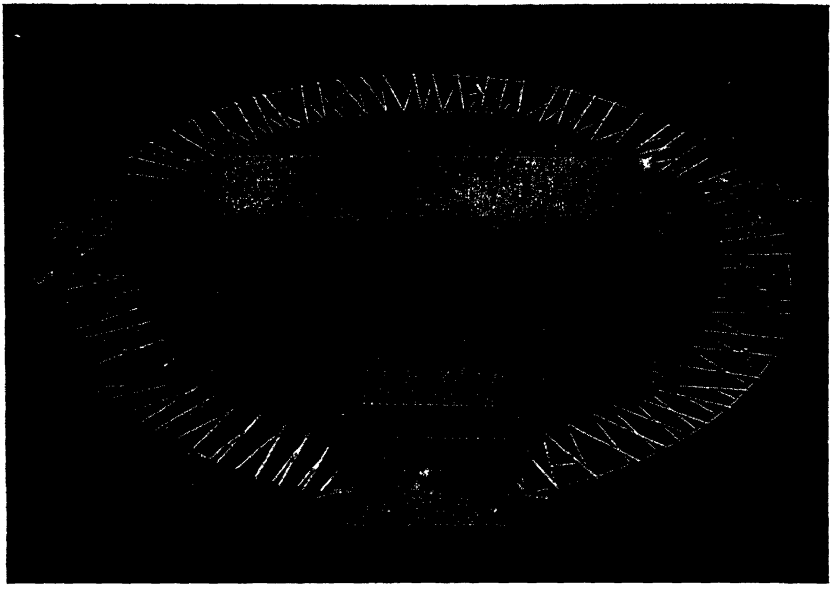


FIG. 24.—Kleynen's rubber sheet model (*A*, anode, *C*, cathode, *G*, grid of amplifier tube) (by courtesy of *Philips Technical Review*).

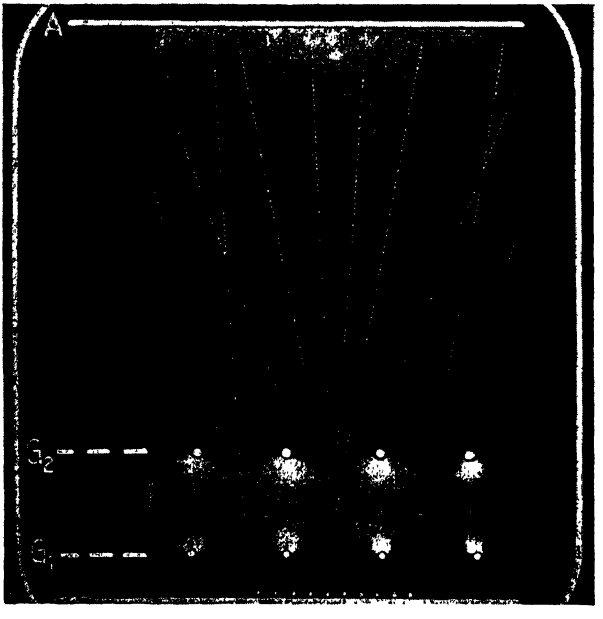


FIG. 25.—Electron trajectories in amplifier tube obtained with the rubber sheet model (Jonker) (by courtesy of *The Wireless Engineer*).

Figure 24 shows a photograph of Kleynen's arrangement, Figure 25 reproduces a typical result obtained with its help.[114] Deviations from the stated ideal conditions introduce errors in the trajectories; errors reported by Kleynen were of the order of 1% to 7%. The rubber sheet cannot represent axially symmetric systems, but can be adapted to the approximate investigation of space charge problems by successive readjustments of the height of the membrane. It was applied to many problems of radio tube design, e.g., to demonstrate the control grid action[113,114,115] and to investigate the current distribution in multigrid tubes,[114,116,117] to the development of secondary emission amplifier tubes[118] and photomultipliers,[92,119] and to the study of electron trajectories in the deflection system of cathode ray tubes.[120] An oscillating rubber membrane for the investigation of transit-time phenomena in high frequency radio tubes was devised by Hollmann.[121]

2. Automatic Trajectory Tracing in Electrolytic Tanks

Several automatic trajectory plotters have been described in which the potential and the local potential gradients are measured simultaneously in an electrolytic tank. The probe carrier is then moved by hand or automatically in such a way that it follows the path which an electron would take under the measured velocity and force conditions, and a stylus attached to the probe carrier traces directly the electron trajectory.

Two such plotters, based on the graphical "circle method," are due to Gabor[122] and to D. B. Langmuir.[123] The probe carriage is moved in such a way that the radius of curvature of its path is given by $2\phi/\text{grad}\ \phi$. The fullest description of these early automatic ray tracing machines was given by Myers.[124]

A semi-automatic plotting method using a single probe was given by Marvaud.[125] The probe traces the path by first making a small excursion in the x-direction, and then returning to the trajectory along the y-direction. The amount Δy necessary for the correct end position of each complete step is obtained from an integration of the velocity components measured in the tank.

A new, fully automatic trajectory plotter based on the principle of continuous integration of the equations of motion of the electron with the help of a mechanical differential analyzer was recently announced by Sander *et al.*[126] The two probe voltage differences measuring the longitudinal and radial field components are each integrated twice by mechanical integrators in cascade which are coupled to the bridge potentiometers. The first integration gives the velocity components, the second integration the displacements of the electron in the x- and y-directions. These

displacements are coupled back into the movement of the probe carrier so that it follows the electron trajectory.

The accuracy of all these trajectory plotting devices is limited by the errors arising from the follow-up mechanisms, and by the factors restricting the accuracy of measurements in electrolytic tanks; these apply particularly to the measurements of the field gradients (D. B. Langmuir[123]).

3. *The Hodoscope*

An apparatus for demonstrating the paths of electrons in magnetic electron lenses, called a "hodoscope," was constructed by Loeb.[127] It uses the fact that a very thin, flexible wire carrying a current J, will take on the shape of the electron trajectory; the radius of curvature of uses the property of a very thin, flexible wire carrying a current J, of taking on the shape of the electron trajectory; the radius of curvature of the trajectory is $R = \frac{mv}{eH}$, whereas the radius of the figure taken on by the thin wire is $R = \frac{T}{JH}$, T being the tension on the wire. Hence $\frac{mv}{e} = \frac{T}{J}$, provided the wire is so thin that gravitational or elastic forces can be disregarded. Figure 26 gives a diagrammatical view of Loeb's apparatus. A silver wire of 0.002 cm. diameter is attached at points A and C; a guide hole B is placed vertically under the point A. Both end points of the wire can be adjusted in height and lateral position by micrometer screws. The tension T, which has to be known for correlation with the electron velocity v, is found by exciting vibrations in the straight line part of the wire AB with the help of the small auxiliary coil E, and measuring the frequency. Conjugate object and image points are characterized by instability of the filament position, as an infinite number of trajectories of equal length is possible between conjugate points. The accuracy appears to be of the same order as that obtained in electrolytic tanks.

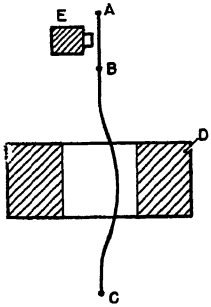


FIG. 26.—Principle of Loeb's hodoscope.

4. *Other Automatic Ray Tracing Methods*

The oldest and best known method of making the electron paths through electron optical systems visible is to fill the envelope containing the electron optical system with a suitable gas, e.g., argon. Some beautiful experiments of this kind are shown in Brüche and Scherzer's book.[98]

An ingenious mechanical model for demonstrating the paths of electrons moving in a magnetic field under the action of transverse

electrostatic fields was described by Rose.[123] He showed that the axis of a precessing gyroscope which is subjected to lateral forces executes the same movement, within first order approximation, as an electron would under the stated conditions. The analogy is formally complete, within the given limits, if the uniform magnetic field is represented by the spin velocity of the gyroscope, and the electric fields acting on the electron are simulated by the field of a permanent magnet acting on one pole of a thin long magnet attached to the gyroscope axis. Figure 27 shows a photograph of an electron trajectory in a split-anode magnetron obtained with Rose's apparatus.

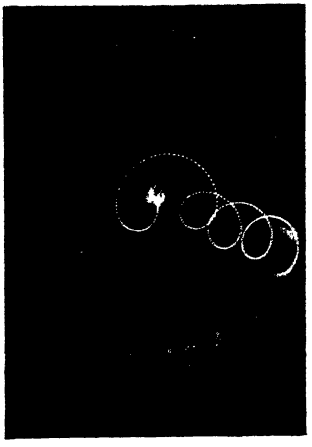


FIG. 27.—Path of an electron in a magnetron, obtained with Rose's analogue (by courtesy of *Journ. of Appl. Phys.*).

Millikan's oil drop experiment was adapted by Jacob[129] to show trajectories of charged particles through electrostatic electron lenses, operated in air of atmospheric pressure. The viscosity of the air makes considerable corrections necessary, and the method is therefore of low accuracy.

V. CONCLUSION

In this last section we shall try to assess the practical value and field of application of the many different methods of field plotting and trajectory tracing discussed in this review.

Most electron-optical problems fall into one of two categories. To the first category belong all those problems where only an approximate answer is required. For instance, it is generally sufficient for the designer to predict the focal length, or the focusing ratio, of a cathode-ray tube within, say, 10%. The exact adjustment of the focusing to the required value is usually carried out by the adjustment of some voltage or current or other parameter within the control of the operator of the device. To this category also belong most radio tube and electron-multiplier problems, where the electrode configuration, and hence the field distribution, may be very complicated. Further, the presence of space charge may often make the results of dubious value. For the solution of problems in this class most of the methods described in this review can be applied, the choice of method depending partly on the kind of equipment available, on the specific features of the problem in hand, and to a certain extent on personal preference. Electrolytic tanks or resistor networks

will be very useful for obtaining the required field maps with sufficient accuracy, but numerical relaxation with a fairly coarse mesh, or graphical field mapping if the electrode structure is very irregular, will be equally successful, and will frequently take a comparable time for similar accuracy. If the problem is to find only the paraxial properties of axially symmetric systems, the numerical ray tracing methods of Cosslett or Liebmann or Goddard (using relatively large intervals) will be fastest and most reliable. If complicated fields are involved, Hepp's or other graphical methods of ray tracing may be indicated.

For the solution of this class of problem an automatic plotting device, if available, will be very useful, particularly if many sets of trajectories have to be plotted or numerous modifications of electrode arrangements have to be investigated, where advantage can be taken of the high operating speed of some of these instruments.

To the second and much more difficult category belong those problems where the second order imaging properties play a part. Most of the features which determine the *quality* of an electronic device, e.g., the current density or resolving power, belong to this category. The operator of the electronic device usually has no control over these features, and it is therefore of great importance that the designer can arrive at the best design parameters. Only the most accurate of the methods described are capable of giving information which is of any value for the determination of the second order properties. Amongst the field plotting methods, the usefulness of the resistor network is outstanding. Failing this, only the relaxation method taken to an advanced stage of refinement, or the fitting of mathematical functions to approximate measured field plots will be sufficiently accurate. For the ray tracing only the more accurate numerical methods can be used with confidence.

ACKNOWLEDGEMENTS

The author wishes to thank all those who have assisted him by providing unpublished information or contributing photographs of apparatus, or by giving permission to reproduce figures or information from their publications. He further expresses his indebtedness to his colleagues for their critical reading of the manuscript and helpful suggestions, and to Dr. T. E. Allibone, F.R.S., for permission to publish this article.

REFERENCES

1. Bertram, S. *Proc. Inst. Radio Engrs., N.Y.*, **28**, 418 (1940).
2. Bertram, S. *J. Appl. Phys.*, **13**, 496 (1942).
3. Liebmann, G. *Phil. Mag.*, **39**, 281 (1948).
4. Zworykin, V. K., Morton, G. A., Ramberg, E. G., Hillier, J., and Vance, A. W. Electron Optics and the Electron Microscope. Wiley, New York, 1945.
5. Rüdenberg, R. *J. Franklin Inst.*, **246**, 311 (1948).

6. Churchill, R. V. Complex Variables and Applications. McGraw-Hill, New York, 1948.
7. Ollendorff, F. Potentialfelder der Elektrotechnik. J. Springer, Berlin, 1932.
8. Rothe, R., Ollendorff, F., and Pohlhausen, K. Theory of Functions as Applied to Engineering Problems. Technology Press, M.I.T., Cambridge, Mass., 1948.
9. Vogdes, F. B., and Elder, F. R. *Phys. Rev.*, **24**, 683 (1924).
10. Dow, W. G. Fundamentals of Engineering Electronics. Wiley, New York, 1937.
11. Herzog, R. *Arch. Elektrotech.*, **29**, 790 (1935).
12. Herne, H. *Wireless Engr.*, **21**, 59 (1944).
13. Liebmann, H. *S.B. Bayer. Akad. München*, 385 (1918).
14. Mueller, A. O. *Arch. Elektrotech.*, **17**, 501 (1926).
15. Shortley, G. H., and Weller, R. *J. Appl. Phys.*, **9**, 334 (1938).
16. Shortley, G., Weller, R., Darby, P., and Gamble, E. H. *J. Appl. Phys.*, **18**, 116 (1947).
17. Kormes, M. *Rev. Sci. Instrum.*, **14**, 248 (1943).
18. Southwell, R. V. Relaxation Methods in Engineering Science. Clarendon Press, Oxford, 1940; Relaxation Methods in Theoretical Physics. Clarendon Press, Oxford, 1946.
19. Emmons, H. W. *Quart. Appl. Math.*, **2**, 173 (1944).
20. Fox, L. *Quart. J. Mech. Appl. Math.*, **1**, 253 (1948).
21. Motz, H., and Klanfer, L. *Proc. Phys. Soc., Lond.*, **58**, 30 (1946).
22. Christopherson, D. G., and Southwell, R. V. *Proc. Roy. Soc.*, **A168**, 317 (1938).
23. Fox, L. *Proc. Camb. Phil. Soc.*, **45**, 50 (1949).
24. Lehmann, Th. *Lumière Élect.*, **8**, 103 (1909); *Rev. Gén. Élect.*, **11**, 347 (1923).
25. Kuhlmann, K. *Arch. Elektrotech.*, **3**, 203 (1914).
26. Moore, A. D. Fundamentals of Electrical Design. McGraw-Hill, New York, 1927.
27. Willoughby, E. O. *J. Instn. Elect. Engrs.*, Part III., **93**, 275 (1946).
28. Poritsky, H. *Trans. Am. Inst. Elect. Engrs.*, **57**, 727 (1938).
29. Bowman-Manifold, M., and Nicoll, F. A. *Nature, Lond.*, **142**, 39 (1938).
30. Pérès, J., and Malavard, L. *Bull. Soc. Franç. Élect.* **8**, 715 (1938).
31. Hepp, G. *Philips Tech. Rev.*, **4**, 223 (1939).
32. Himpan, J. *Telefunkenröhre*, No. 16, 198 (1939).
33. Himpan, J. *Elektrotech. Z.*, **63**, 349 (1942).
34. Ertaud, A. L'Optique Électronique. Ed. L. de Broglie. *Rev. Optique*, Paris, 1946, p. 105.
35. Musson-Genon, R. *Ann. des Télécomm.*, **2**, 298 (1947).
36. Hutter, R. G. E. *J. Appl. Phys.*, **18**, 797 (1947).
37. Ehrenfried, A. D. *Am. J. Phys.*, **12**, 371 (1944).
38. Green, P. E. *Rev. Sci. Instrum.*, **19**, 646 (1948).
39. Peierls, R. E. *Nature, Lond.*, **158**, 831 (1946).
40. Peierls, R. E., and Skyrme, T. H. R. *Phil. Mag.*, **40**, 269 (1949).
41. Wilkinson, K. J. R. *B.T.H. Activ.*, **19**, 421 (1948).
42. Bracher, D. F., and Hide, G. S., to be published.
43. Jacob, L. *Phil. Mag.*, **25**, 570 (1938).
44. Hill, R. D. *J. Sci. Instrum.*, **22**, 221 (1945).
45. Bruck, H., and Romani, L. *Cahiers phys.*, No. 24, 15 (1944).
46. Theile, R., and Himpan, J. *Telefunkenröhre*, No. 18, 50 (1940).
47. de Haller, P. *Proc. Sixth Intern. Congr. Appl. Mech.*, Paris, Sept. 1946; *Sulzer Tech. Rev. Switz.*, No. 3/4, 11 (1947).

48. Einstein, P. A., to be published.
49. Makar, R., Boothroyd, A. R., and Cherry, E. C. *Nature, Lond.*, **161**, 845 (1948).
50. Schmude, H., and Schwenkhagen, H. *Telefunkenröhre*, No. 24/25, 47 (1942).
51. Klemperer, O. *J. Sci. Instrum.*, **21**, 88 (1944).
52. Jones, G., and Christian, S. M. *J. Am. Chem. Soc.* **57**, 272 (1935).
53. Hogan, T. K. *J. Instn. Engrs., Aust.*, **15**, 89 (1943).
54. Redshaw, S. C. *Proc. Instn. Mech. Engrs.*, **159**, 25 (1948).
55. de Packh, D. C. *Rev. Sci. Instrum.*, **18**, 798 (1947).
56. Kron, G. *Elect. Engng. N.Y.*, **67**, 672 (1948).
57. Liebmann, G. *Nature, Lond.*, **146**, 149 (1949).
58. McKeehan, L. W. *J. Opt. Soc. Amer.*, **19**, 213 (1929).
59. Klemperer, O. *Phil. Mag.*, **20**, 545 (1935).
60. Simpson, J. A. *Rev. Sci. Instrum.*, **11**, 430 (1940).
61. Marton, L. *Phys. Rev.*, **55**, 672 (1939); *Proc. Inst. Radio Engrs., N.Y.*, **32**, 546 (1944).
62. Dosse, J. *Z. Phys.*, **117**, 437 (1941).
63. Sandor, J. *Arch. Elektrotech.*, **35**, 217 and 259 (1941).
64. Brown, W. F., and Sweer, J. H. *Rev. Sci. Instrum.*, **16**, 276 (1945).
65. Williamson, K. I. *J. Sci. Instrum.*, **24**, 242 (1947).
66. Kohaut, A. *Z. techn: Phys.*, **18**, 198 (1937).
67. Cole, R. H. *Rev. Sci. Instrum.*, **9**, 215 (1938).
68. Klemperer, O., and Miller, H. *J. Sci. Instrum.*, **16**, 121 (1939).
69. Goddard, L. S., and Klemperer, O. *Proc. Phys. Soc., Lond.*, **56**, 378 (1944).
70. van Ments, M., and Le Poole, J. B. *Appl. Sci. Res.* **B1**, 3 (1947).
71. Marton, L. *J. Appl. Phys.*, **19**, 687 (1948).
72. Marton, L. *J. Appl. Phys.*, **19**, 863 (1948).
73. Dupouy, G. *Ann. Phys. Paris*, **14**, 549 (1931).
74. Pearson, G. L. *Rev. Sci. Instrum.*, **19**, 263 (1948).
75. Gregg, E. C. *Rev. Sci. Instrum.*, **18**, 77 (1947).
76. Kolin, A. *Rev. Sci. Instrum.*, **16**, 209 (1945).
77. Rhoderick, E. H. *Rev. Sci. Instrum.*, **17**, 453 (1946).
78. Semenoff, N., and Walther, A. *Z. Phys.*, **17**, 67 (1923).
79. Walther, A., and Inge, L. *Z. Phys.*, **19**, 192 (1923).
80. Semenoff, N., and Walther, A. *Z. Phys.*, **19**, 136 (1923).
81. Street, R. *J. Sci. Instrum.*, **23**, 203 (1946).
82. Gundert, E. *Telefunkenröhre*, No. 19/20, 61 (1941).
83. Balachowsky, M. *Bull. Soc. Franç. Élect.*, **6**, 181 (1946).
84. Glaser, W. *Z. Phys.*, **118**, 264 (1942).
85. Graffunder, W. *Bull. Ass. Suisse Élect.*, **38**, 79 (1947).
86. Salinger, H. *Electronics*, **10**, 50 (October, 1937).
87. Spangenberg, K., and Field, L. M. *Proc. Inst. Radio Engrs., N.Y.*, **30**, 138 (1942).
88. Zworykin, V. K., and Morton, G. A. Television—The Electronics of Image Transmission. Wiley, New York, 1940.
89. Selgin, P. J. *Electronics*, 124 (September 1948).
90. Rajchman, J. A. *Arch. Sci. Phys. Nat.* 5th Ser., **20**, 231 (1938).
91. Musson-Genon, R. *Ann. Télécomm.*, **2**, 254 (1947).
92. Zworykin, V. K., and Rajchman, J. A. *Proc. Inst. Radio Engrs., N.Y.*, **27**, 558 (1939).
93. Dosse, J. *Z. techn. Phys.*, **17**, 315 (1936).

94. Weimer, P. K., and Rose, A. *Proc. Inst. Radio Engrs., N.Y. Wav. Electrons*, 1273 (1947).
95. Klemperer, O., and Wright, W. D. *Proc. Phys. Soc. Lond.*, **51**, 296 (1939).
96. Conrady, A. E. Applied Optics and Optical Design. Oxford Univ. Press, London, 1929.
97. Duncan, W. J. *Phil. Mag.*, **39**, 493 (1948).
98. Brüche, E., and Scherzer, O. Geometrische Elektronenoptik. Springer, Berlin, 1934.
99. Störmer, C. *Arch. Sci. Phys. Nat.* 4th Ser., **24**, 221 (1907).
100. Maloff, I. G., and Epstein, D. W. *Proc. Inst. Radio Engrs., N.Y.*, **22**, 1386 (1934); Electron Optics in Television. McGraw-Hill, New York, 1938, pp. 81–88.
101. Gans, R. *Z. techn. Phys.*, **18**, 41 (1937).
102. Recknagel, A. *Z. Phys.*, **104**, 381 (1937).
103. Siday, R. E. *Proc. Phys. Soc., Lond.*, **54**, p. 266 (1942).
104. Dosse, J. *Z. Phys.*, **117**, 722 (1941).
105. Glaser, W. *Z. Phys.*, **83**, 104 (1933); *Z. Phys.*, **116**, 19 (1940).
106. Picht, J. *Ann. Phys. Lpz.*, **15**, 926 (1932).
107. Ramberg, E. G. *J. Appl. Phys.*, **13**, 582 (1942).
108. Scherzer, O. *Z. Phys.*, **101**, 593 (1936).
109. Cosslett, V. E. Introduction to Electron Optics. Clarendon Press, Oxford, 1946, pp. 37–38.
110. Liebmann, G. *Proc. Phys. Soc., Lond.*, **62B**, 753 (1949).
111. Goddard, L. S. *Proc. Phys. Soc., Lond.*, **56**, 372 (1944).
112. Blewett, J. P. *Proc. Inst. Radio Engrs., N.Y. Wav. Electrons*, **36**, 69 (1948).
Kron, G., Maginniss, F. J., and Peterson, H. A. *Proc. Inst. Radio Engrs., N.Y. Wav. Electrons*, **36**, 70 (1948).
Hahn, W. C., and Blewett, J. P. *Proc. Inst. Radio Engrs., N.Y. Wav. Electrons*, **36**, 74 (1948).
Whinnery, J. R., and Jamieson, H. W. *Proc. Inst. Radio Engrs., N.Y. Wav. Electrons*, **36**, 76 (1948).
113. Kleynen, P. H. J. A. *Philips Tech. Rev.*, **2**, 321 (1937).
114. Jonker, J. L. H. *Wireless Engr.*, **16**, 344 (1939).
115. Fremlin, J. H. *Phil. Mag.*, **27**, 709 (1939).
116. Jonker, J. L. H. *Philips Tech. Rev.*, **5**, 131 (1940); *Tijdschr. Ned. Radiogenoot.*, **10**, 113 (1943); *Philips Research Repts.*, **1**, 19 (1945); *ibid.*, **1**, 331 (1946).
117. Fremlin, J. H. *Electronic Engng.*, **16**, 103 (1943).
118. Jonker, J. L. H., and van Overbeek, A. J. W. M. *Wireless Engr.*, **15**, 150 (1938).
119. Pierce, J. R. *Bell Lab. Rec.*, **16**, 305 (1938).
120. Hollmann, H. E., and Thoma, A. *Z. tech. Phys.*, **19**, 475 (1938).
121. Hollmann, H. E. *Proc. Inst. Radio Engrs., N.Y.*, **29**, 70 (1941).
122. Gabor, D. *Nature, Lond.*, **139**, 373 (1939).
123. Langmuir, D. B. *Nature, Lond.*, **139**, 1066 (1937).
124. Myers, L. M. Electron Optics. Chapman and Hall, London, 1939, pp. 136–142.
125. Marvaud, J. *C.R. Acad. Sci. Paris*, **226**, 476 (1948).
126. Sander, K. F., Oatley, C. W., and Yates, J. G. *Nature, Lond.*, **163**, 403 (1949).
127. Loeb, J. *L'Onde Élect.*, **27**, 27 (1947).
128. Rose, A. *J. Appl. Phys.*, **11**, 711 (1940).
129. Jacob, L. *Nature, Lond.*, **159**, 475 (1947).

Cathodoluminescence

G. F. J. GARLICK

Physics Department, University of Birmingham, England

CONTENTS

I. Introduction

1. The Aspects of Cathodoluminescence Under Review

During the past ten years many papers have been published on the subject of cathodoluminescence, that is, the luminescence emitted from substances when bombarded by electrons. At the present time the extensive use of cathode ray tubes in television, radar, and scientific researches of all kinds has familiarized many users with the practical function of the luminescent screens of the tubes. In this review we shall be concerned with the types and characteristics of the materials which are used for these screens. For the most part the researches on cathodoluminescence have been made on the conventional materials in practical use and have been largely prompted by the practical demands of the particular application of the cathode ray tube screen. As a result there exists a wealth of experimental data on specific screen materials under the practical conditions of operation, or under laboratory conditions which simulate them, but no comprehensive and quantitative understanding of the physical mechanisms involved in cathodoluminescence. The present theoretical concepts of cathodoluminescence are largely qualitative and in the present instance we shall consider them separately from the experimental studies.

The structure of this review is briefly as follows. In the remainder of the introduction we discuss the more common terminology of the subject and then consider the classes of materials which are of practical interest. We end the section with a brief summary of the practical requirements of luminescent solids for use as cathode ray tube screens.

In the second section the experimental studies of cathodoluminescence are described. They include studies of the electrical characteristics of cathodoluminescent solids and of their stability under electron bombardment. The third section is devoted to a consideration of the theoretical interpretation of cathodoluminescence phenomena. We shall conclude the review with a discussion of the present trends of cathodoluminescence research. The presentation of the subject is that of a physicist working in the fundamental aspects of luminescence. However, as discussed in the final section, the usefulness of unsystematic empirical methods in the practical development of phosphors for cathode ray tubes is constantly diminishing and it is felt that we need to adopt a new approach in such research if better materials are to be its outcome. For this reason we must here point to the need for understanding more about the basic processes in electron excitation of phosphors.

2. *The Concepts of Luminescence*

Luminescence is a generic term which covers many phenomena in which energy is converted into visible radiation at ordinary temperatures. When visible emission is obtained by absorption of ultraviolet light in a substance it is referred to as *photoluminescence:* other prefixes are usually self explanatory, for example, *chemiluminescence.* We sometimes refer to the luminescence emitted during excitation of a substance as *fluorescence.* Emission which continues for longer than 10^{-8} seconds after excitation ceases will be called *phosphorescence* or *afterglow.* There has been some discussion in the past about the definition of these terms.[1] The above definitions specify the physical conditions rather than the nature of the processes involved and so the controversial aspect of the distinction between them is avoided. Phosphorescence is often referred to as *persistence* in radar publications. Some phosphorescent solids emit luminescence when warmed in the dark after excitation and this is known as *thermoluminescence.* Recent studies of thermoluminescence have furnished valuable evidence on the nature of the phosphorescence processes in solids.[2] The solids which are of practical interest are known as *phosphors.* We shall describe briefly the various classes of phosphors used in cathode ray tubes as screen materials.

3. *Classes of Cathodoluminescent Solids*

All the phosphors used for cathode ray tube screens are prepared by careful heat treatment of their basic constituents. By this means they become luminescent materials and are formed into microcrystals (usually 1–100 μ in size). In most of these phosphors the essential constituent for luminescence is a small but controlled amount of a specific impurity

added before heating to the matrix material, for example, green emitting zinc silicate phosphors contain about 0.1 to 1.0% of manganese. The impurities which produce the luminescence are known as activators. In some phosphors no impurity is deliberately added but a stoichiometric excess of one constituent, produced by heating, serves as the luminescence activator. In all these phosphors it has been established that the seat of the luminescence emission is in the centers formed by the activator, although such centers may include other atoms or ions as well as that which gives rise to the characteristic emission. We may summarize the chief classes of phosphor used for cathode ray tube screens as follows:

a) Oxide phosphors; particularly zinc oxide.
b) Sulfide phosphors; zinc and cadmium sulfides, sulfo-selenides and tellurides of zinc and cadmium.
c) Silicate phosphors, zinc and zinc-beryllium silicates.
d) Tungstates; calcium, magnesium, cadmium, and zinc tungstates.
e) Fluorides; zinc and magnesium fluorides.

There are other phosphors not included in these classes but they are not in common use and many are still in the developmental stage. We now consider briefly the phosphors of each class:

a. Oxide Phosphors—Zinc Oxide. Zinc oxide is the only phosphor of this class of practical usefulness. It is usually made by heating zinc oxide to about 1000°C. No deliberate impurity is added but a stoichiometric excess of zinc must be produced by using a reducing atmosphere. This phosphor has been investigated recently by Leverenz.[3] It has a characteristic green emission and the spectrum is given in Fig. 1. This figure also includes the spectrum of specimens made without the reducing atmosphere, in which case the emission is in the ultraviolet. Both spectra are for electron excitation. The afterglow decay of zinc oxide is very rapid and does not persist for more than a microsecond at ordinary temperatures.

b. Sulfide Phosphors. Zinc sulfide and zinc-cadmium sulfide phosphors have been perhaps the most extensively investigated phosphor systems. The most important impurity activators for these phosphors are silver, copper, and manganese, each producing characteristic emission as shown by the spectra of Fig. 2. Zinc sulfide may also be self activated as in the case of zinc oxide. Increase in cadmium sulfide content of mixed phosphor systems shifts the emission towards the red, as shown in the figure, except in the case of manganese activated specimens. The effect of preparation and constitution on the luminescence characteristics of these phosphors has been exhaustively studied in recent years.[4,5,6,7] Copper activated zinc and zinc-cadmium sulfides possess the longest after-

glows of this class of phosphor and have found extensive application as cathode ray tube screens in radar systems.[3,8]

Copper and silver activated zinc and zinc-cadmium selenides and tellurides are similar in their physical behavior to zinc and zinc-cadmium

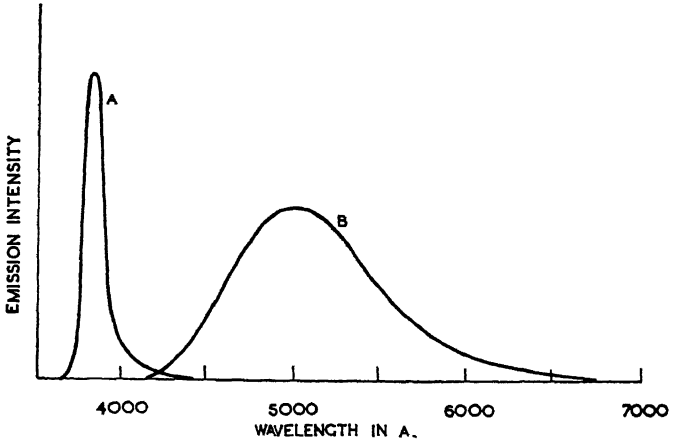


FIG. 1.—Cathodoluminescence spectra of zinc oxide phosphors (Leverenz). A—Phosphor prepared at 1000°C. in ordinary atmosphere, B—Phosphor prepared at 1000°C. in reducing atmosphere.

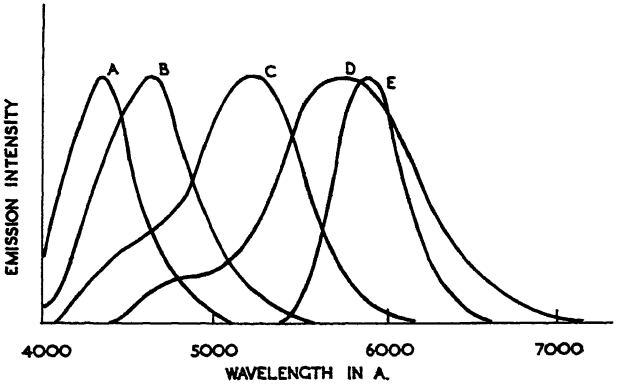


FIG. 2.—Cathodoluminescence spectra of zinc sulfide phosphors (Leverenz). A-ZnS-Ag (Hexag. form), B-ZnS-Zn, C-ZnS-Cu, D-ZnS(80%)CdS(20%)-Cu, E-ZnS-Mn(2%).

sulfides. Zinc sulfo-selenides are of interest as television screens in cathode ray tubes. Their emission can be made "near white" while their afterglow is negligible. These phosphors are mentioned later in Section IV, 1, *a*.

c. Silicate Phosphors. The most well known phosphors of this class are the zinc and zinc-beryllium silicate phosphors activated by manganese. Spectra for these phosphors measured by Schulman[9] are given in Fig. 3. The physical characteristics of these phosphors have been recently studied by many workers.[10,3,11] Other phosphors of this class include cadmium, magnesium, and calcium silicates. Most of these phosphors have very little long afterglow but possess a short lived exponential decay of emission with a half life of a few milliseconds.

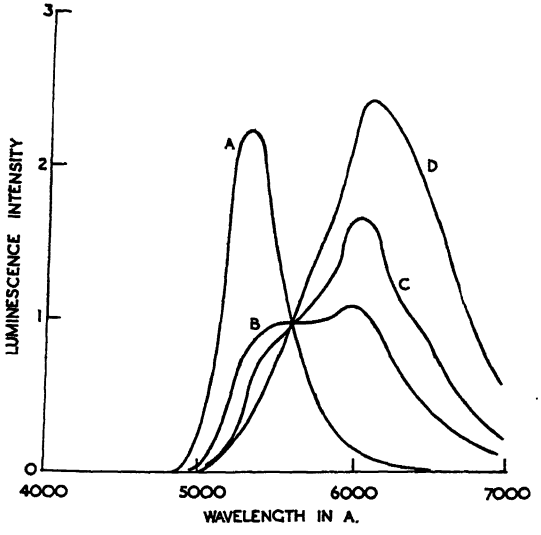


FIG. 3.—Luminescence spectra of zinc beryllium silicate phosphors (Schulman). Molar constitution:

A—ZnO (2.0) MnO (0.122) SiO_2 (1.18)
B—ZnO (1.8) BeO (0.2) MnO (0.058) SiO_2 (1.11)
C—ZnO (1.8) BeO (0.2) MnO (0.111) SiO_2 (1.17)
D—ZnO (1.6) BeO (0.4) MnO (0.132) SiO_2 (1.19).

d. Tungstate Phosphors.[3,12,13] The important phosphors of this class are the tungstates of calcium, magnesium, cadmium, and zinc; they are prepared by heat treatment without inclusion of an impurity. Their cathodoluminescence spectra are given in Fig. 4.[3] The emission is supposed to take place in the excess WO_4^{--} ions. Thus the spectra are very similar. The exception is calcium tungstate but this has a different crystal structure from the others and so the difference in spectrum can be ascribed to the change in the surroundings of the tungstate ions in this lattice.

e. Fluoride Phosphors.[3,14] Manganese activated zinc and magnesium fluorides have found important uses in cathode ray tubes for radar.

Their emission spectra are broad bands similar to those of other manganese activated phosphors and occur in the yellow and orange spectral regions for the zinc and magnesium fluorides respectively. Their unique property is the possession of afterglow which decays exponentially, having a half life of nearly 100 milliseconds: the decay rate is not affected by temperature. They show marked deterioration of luminescence efficiency when bombarded with electrons of low energy (0–5 kv).

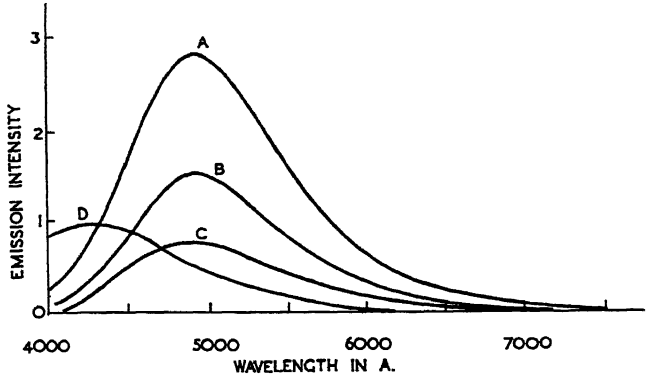


FIG. 4.—Cathodoluminescence spectra of tungstate phosphors (Leverenz).
A-$MgWO_4$, B-$CdWO_4$, C-$ZnWO_4$, D-$CaWO_4$.

4. The Practical Requirements of Cathodoluminescent Materials

Certain standards of performance are demanded of phosphors for use in cathode ray tubes. These may be summarized as follows:

i. The phosphor must have a high luminescence efficiency with the desired spectral characteristics under the given operating conditions.

ii. The afterglow characteristics must be suited to the particular application. This is important in television and radar.

iii. The secondary electron emission of the phosphor must be adequate under the given conditions.

iv. The phosphor must have a high stability under prolonged electron bombardment.

These practical requirements have determined to a large extent the type of research carried out in the field of cathodoluminescence, which we shall describe in the next section.

II. EXPERIMENTAL STUDIES OF CATHODOLUMINESCENCE

1. Luminescence in Relation to Electron Energy and Density

To study the luminescence emission due to electron excitation different methods of exciting the phosphor may be used. In one case the

electrons fall in a defocused beam on a given phosphor area whereas in another case the beam is focused and the area of phosphor under observation is scanned in raster fashion. In the latter case the excitation conditions for a screen element are much more complex than in the former case but the periodic pulsed excitation may be a closer simulation of practical conditions. However, the general form of the variation of emission with electron current and density may not be very different for the two cases.

a. Variation of Luminescence with Electron Current Density. At sufficiently low current densities the luminescence intensity (L) of conventional cathode ray tube screens is proportional to the current density (i) but as the latter increases a saturation effect is observed. The value of current density at which this occurs is different for different phosphors but in all cases is approximately independent of the electron energy (over the usual working range of 0–10 kv). For zinc silicate the saturation is evident at about 10 μa/sq. cm. but for zinc sulfide phosphors it is absent at 100 μa/sq. cm. We may compare the saturation effects in different phosphors by finding the current density at which the light output per unit current density has decreased to 80% of its value at lower current densities: Table I includes such values for the phosphors listed. Figure 5, due to Strange and Henderson,[15] shows the variation of emission with current density at two different electron voltages for a manganese activated zinc silicate phosphor. Comparison of the ratio of the intensities for given current densities at the two different voltages shows that the saturation effect is independent of the electron voltage. This fact was previously established for zinc silicate phosphors by Nottingham in 1937 (see ref. 20 of Table I).

TABLE I. Cathodoluminescence characteristics of phosphors.

Phosphor	Current density for 20% light loss in μa/sq. cm.	Index n	V_0 in volts	V_L in kilovolts	Ref. No.
Zn_2SiO_4-Mn	10	1.5	.....	.....	19
Zn_2SiO_4-Mn	10	2.0	100	6.5	20
Zn_2SiO_4-Mn	10	2.0	100	6.5	21
Zn_2SiO_4-Mn	25	1.53–1.6	0	9.0	15
ZnS-Ag	200	2.8	100	> 10.0	21
ZnS-Ag	150	1.8–2.0	.....	6.0	19
ZnS-Ag	500	1.36	150	> 10.0	15
ZnS-CdS-Ag	10	variable	.....	> 10.0	21
ZnS-Mn	70	1.0	200–300	> 10.0	15
$CaWO_4$	30	2.0	100	5.0	21

b. Variation of Luminescence with Electron Energy. For a given current density the luminescence intensity of a phosphor screen increases with the energy of the incident electrons (measured as the cathode to screen potential V) and this variation is represented by the empirical formula:

$$L = k(V - V_0)^n \tag{1}$$

where k, V_0 and n are constants for a given screen. The index n varies from 1 to 3 for different phosphors. The potential V_0 is known as the

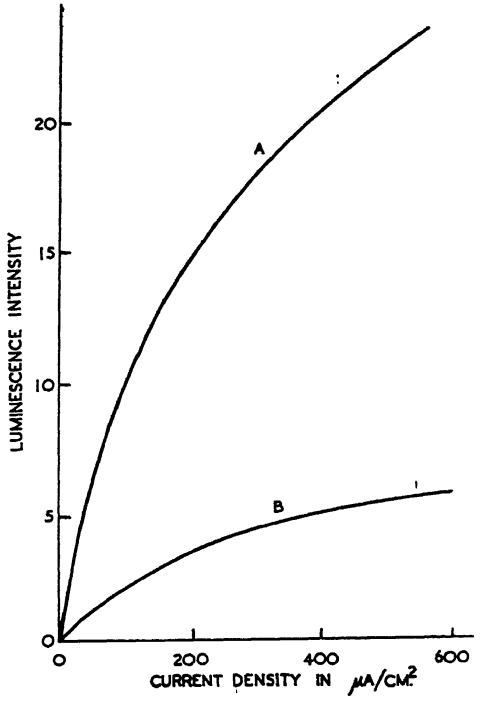


FIG. 5.—Variation of cathodoluminescence with electron current density (Strange and Henderson). A—10 kv, B—4 kv.

dead voltage of the screen. It depends not only on the phosphor screen but also on the preparation of the screen and the tube processing. As long ago as 1909 Lenard and Saeland[16] ascribed it to surface contamination of the phosphor. Their view seems to be confirmed by the work of Strange and Henderson[15] who, by careful preparation of zinc silicate screens, obtained a specimen with no dead voltage. We would, however, expect the phosphor to have a small negative "dead" voltage due to its internal potential.[17] The validity of eq. 1 has been demonstrated by

various workers. The index n is found to be independent of current density even when saturation effects are present. We have listed the values of the constants V_0 and n for different phosphors, as obtained by these workers, in Table I. We have also included the values of V_L, the potential at which the secondary electron emission from the screen becomes too small to prevent the charging up of the screen and at which

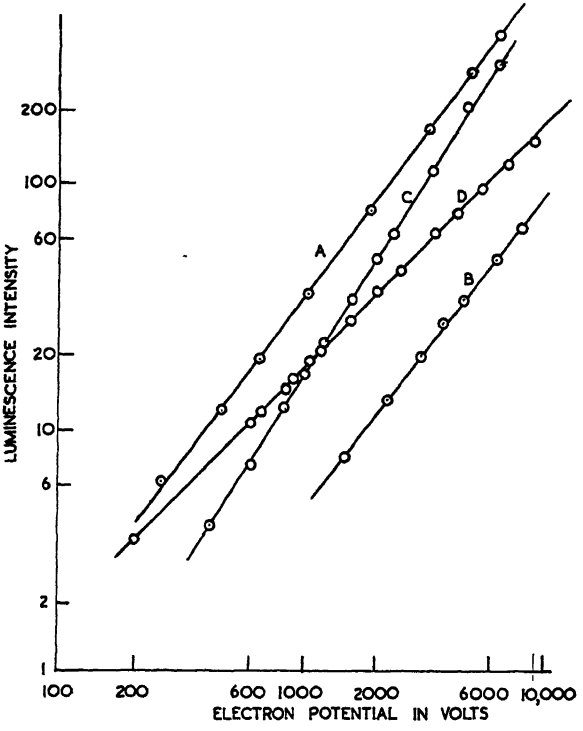


FIG. 6.—Variation of cathodoluminescence with electron potential at fixed current densities (Strange and Henderson).

A—ZnS-Ag:10 μamp./cm.^{2}, B—ZnS-Ag:550 μamp./cm.^{2},
C—Zn_2SiO_4-Mn:5 μamp./cm.^{2}, D—ZnS-Mn:10 μamp./cm.^{2}

the above eq. 1 breaks down as a result. We shall discuss the significance of V_L in the next section. All the results listed in Table I are for excitation of the phosphor screens with a stationary, defocused electron beam. The values of the various parameters change slightly if faster excitation is used. For example, Piore and Morton[18] find a change of the index n from 1.95 to 1.84 for a particular specimen when the excitation is changed from static to raster conditions. We give in Fig. 6 curves obtained by

Strange and Henderson which show the validity of eq. 1 for various phosphors and for different current densities.

2. *The Rise and Decay of Luminescence*

The form of the rise and decay of luminescence is important in both television and radar. In the case of television it is the first few milliseconds of these processes which matter and so we consider first the initial stages.

a. Initial Decay Processes.[15,19,22] After suitable excitation most phosphors have a very rapid initial decay of phosphorescence which takes place in a few microseconds. It is followed by a slower decay component which may last several milliseconds. In both these initial stages the phosphorescence decays exponentially with time but each component has a different decay rate. Strange and Henderson[15] have called the faster component the α process and the slower component the β process. With respect to these processes they divide phosphors into two main categories. One of these is known as the "Ag" category since silver activated zinc sulfide phosphors are typical of it. The other is known as the "Mn" category since manganese activated phosphors are typical of it. Phosphors of the Ag class have a β process which only lasts for microseconds while those of the Mn class have β processes lasting for milliseconds. The relative percentages of the α and β processes present at different current densities of excitation for a silver activated zinc sulfide phosphor are given in Table II, below, due to Strange and Henderson.[15] The table shows how the percentage of the α process increases with current density but also shows that the decay rates for both α and β processes do not vary markedly with the current density.

TABLE II. Initial decay processes in silver activated zinc sulfide phosphor.

Current density in $\mu a/cm.^2$	Decay constant of α process in $sec.^{-1}$	Decay constant of α process in $sec.^{-1}$	$\%\alpha$ Proc.	$\%\beta$ Proc.
5	1.10^5	$1.3.10^4$	50	50
12	1.10^5	$1.45.10^4$	63	37
60	1.10^5	$1.65.10^4$	75	25
180	$1.2.10^5$	$4.1.10^4$	87	13
700	$1.8.10^5$	$5.0.10^4$	90	10
5000	$4.6.10^5$	10^5	95	5

Except after high current densities of excitation, there is no appreciable amount of the α process in phosphors of the Mn class. In Fig. 7 we give decay curves for a manganese activated zinc silicate phosphor at two very different excitation densities. The initial rapid decay of the α

process can be seen in curve A. When the manganese content of such phosphors is increased beyond a certain amount the decay constant of the β process begins to increase. Strange and Henderson[15] have measured the change in the decay constant for the β process with manganese content using cathode ray excitation while Fonda has made similar measurements using ultraviolet excitation (2537 A.). These two sets of inde-

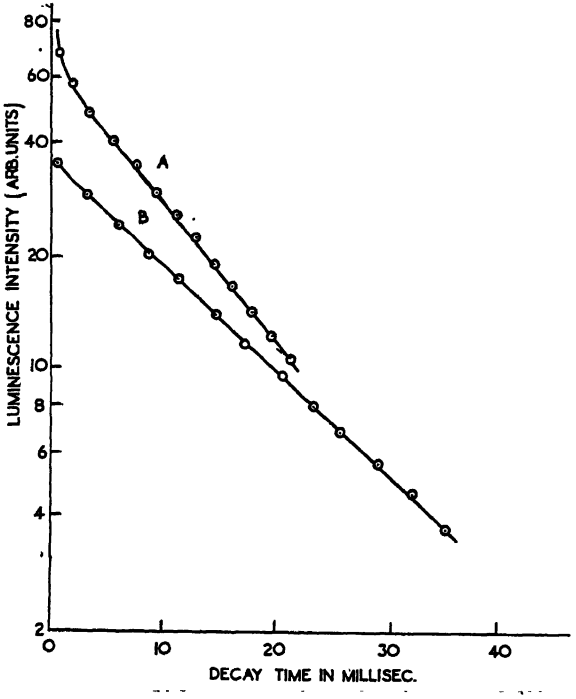


FIG. 7.—Initial decay processes for a Zn_2SiO_4-Mn phosphor (Strange and Henderson).

A—Current density 200 μa/cm.^{2}
B—Current density 1 μa/cm.^{2}
} Voltage 5 kv.

pendent measurements are in good agreement for phosphors having the same manganese content.[24] A theoretical explanation of the increase in decay rate with activator content and at high temperatures has been given by Kröger and his associates.[25] We shall not discuss their hypothesis.

b. Growth Processes.[15] Two components similar to the decay components are observed in the rise of luminescence during excitation. The rates of rise do not, however, depend on the decay constants alone but vary with the current density. They appear to be independent of the electron energy.

c. Long Afterglow Due to Cathode Ray Excitation. Phosphors showing long afterglow when excited by ultraviolet radiation also show it when excited by electrons but the two decay characteristics are quite different. In Fig. 8 we compare the decay of the afterglow of a ZnS-CdS-Cu phosphor screen after ultraviolet excitation with that after electron excitation. The decay is slower for ultraviolet excitation than for electrons although the initial brightness is the same in each case. In addition the light sum over the whole decay is much greater for ultravoilet excitation. It is also found that for a given initial brightness the decay intensity saturates at a much earlier time for electron excitation than for ultraviolet excitation (by saturation we infer that the brightness

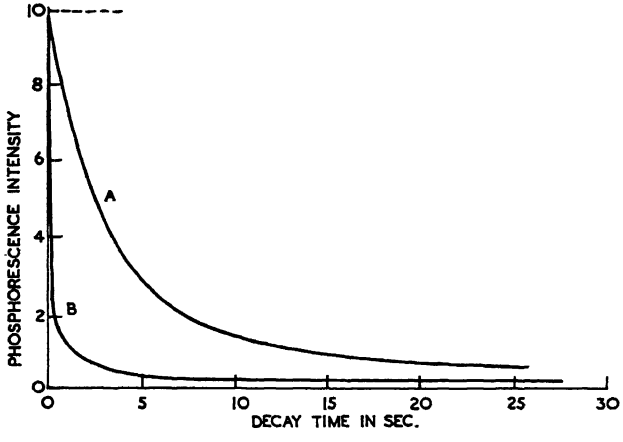


FIG. 8.—Long afterglow decay of a ZnS-CdS-Cu phosphor excited to a given initial brightness.

A—3650 A. ultraviolet excitation, B—Electron excitation (6 kv).

of decay at a given time is independent of further increase in excitation intensity). We shall discuss these comparisons in the next section with respect to their theoretical interpretation.

3. The Cathodoluminescence Efficiency of Phosphors

When phosphors, such as zinc sulfide, are excited by ultraviolet radiation of a wavelength giving optimum conversion of ultraviolet energy to visible light it is found that one quantum of luminescence is emitted for each quantum of ultraviolet radiation absorbed. For phosphors emitting in the blue-green region the conversion efficiency can be of the order of 70 or 80%. This happens when the ultraviolet light is absorbed directly in the impurity centers. When it is of shorter wavelength and is absorbed in the crystal lattice of the phosphor the efficiency

is only of the order of 10% of the optimum value. A similar value of 10% for the conversion efficiency is the maximum usually found for cathode ray excitation when the electrons are absorbed in the bulk material.[26] The efficiency under cathode ray excitation increases with electron voltage, as would be expected from eq. 1 if the value of the index n is greater than unity (efficiency measured as light output per unit power $V.i$). The similar efficiency values for ultraviolet excitation at short wavelengths and for cathode rays has a theoretical significance which we shall discuss in Section III, 2.

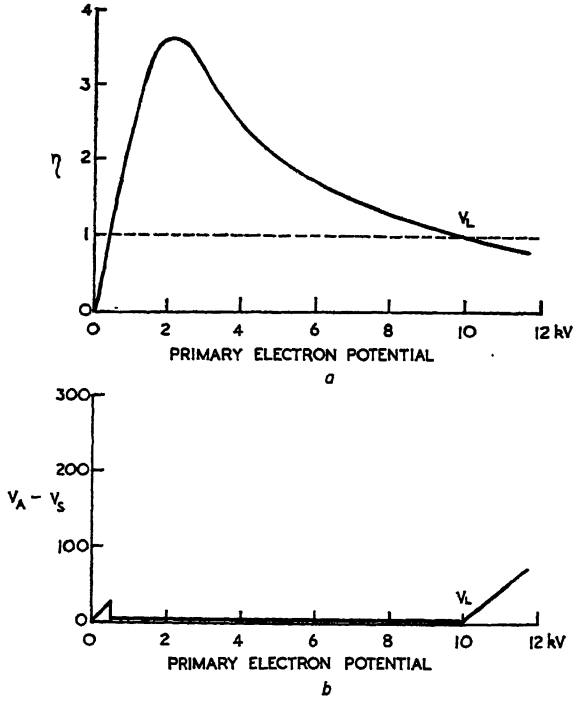


FIG. 9.—(a) Variation of secondary electron emission ratio (η) with primary electron potential for a phosphor. (b) Variation of screen potential (V_s) relative to anode potential (V_A) with primary electron potential.

4. *The Secondary Electron Emission of Phosphors*

a. Dependence of Secondary Emission on Primary Electron Energy. The primary electrons incident on a phosphor screen produce secondary electrons inside the crystal grains. Some of these escape from the phosphor while others are used in the luminescence process. If the secondary emission of the screen is inadequate the screen will charge up and repel

the primary electrons, reducing their energy and so reducing the luminescence intensity. The change in secondary emission with primary electron potential is shown by Fig. 9a. The ordinate gives the ratio (η) of secondary emission current to primary current. The useful operating region for the phosphor is between the two points at which η becomes unity. The upper voltage limit we have already called V_L. It is usually known as the *sticking potential.* Figure 9b shows how the potential of an isolated phosphor screen varies with the electron voltage. The points at which η is unity are shown by the change in the screen potential. This method of indicating the secondary emission of a phosphor by measuring the screen potential can be used for conventional sealed off cathode ray tubes. The method of making the potential measurements has been described by Nelson.[27] The secondary emission characteristics of phosphors and other insulators have been investigated by many workers.[28,29,30] In general the values of η and V_L are much larger for crystalline insulators than for metals. We shall discuss this in Section III.

b. Secondary Emission in Relation to Luminescence. We have already seen in the previous paragraph that the useful working range of the conventional type of cathode ray tube screen with respect to electron voltage is limited by the sticking potential. In order therefore to use phosphors in high voltage cathode ray tubes the screen must be prevented from charging up. This is achieved in practice by providing the screen with a conducting backing, usually a layer of evaporated aluminum. The technique and performance for these backed screens has been described by Epstein and Pensak.[31]

5. The Fatigue and Deterioration of Phosphors under Electron Bombardment

It is found that there are two distinct types of deterioration of phosphors due to electron bombardment, one which occurs in certain types of phosphors at low electron voltages and one which occurs in all phosphors at high electron voltages. We shall consider these effects separately below.

a. Phosphor Deterioration at Low Voltages (0–5 kv). The decrease in luminescence efficiency due to prolonged electron bombardment is often referred to as *electron burn.* Its occurrence at low electron voltages is most marked in phosphors of the Mn class which have relatively high activator concentrations. The decrease in efficiency is often accompanied by a visible darkening of the phosphor. Extensive experiments have been made by Grotheer[32] who has attempted to relate the decrease in luminescence efficiency to the amount of visible darkening, assuming

the latter to be due to the formation of colloidal metal in the phosphor. His results are not conclusive; in Fig. 10 we give curves from his results showing the progress of electron burn with bombardment for different phosphors using 1600 volt electrons. It will be noticed that the degree of burn decreases as the decay constant for the specimen increases. The degree of burn for a given phosphor and a given excitation decreases as the electron voltage increases. This must be associated with the increase

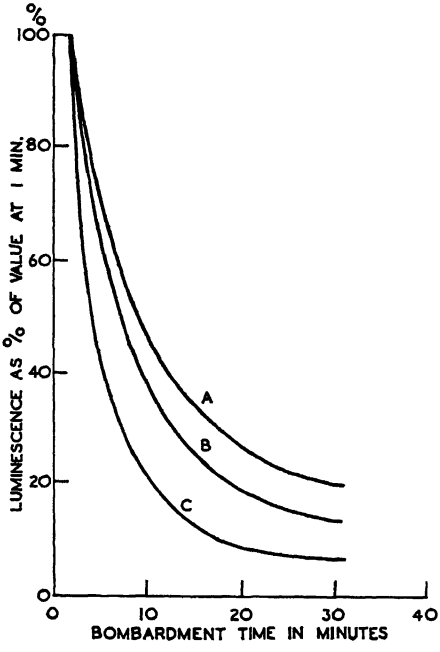


FIG. 10.—Deterioration of phosphor efficiency with electron bombardment at low voltages (1600 v:700 μa/cm.^{2}) (Grotheer).

A—ZnS-Cu, B—$CaWO_4$, C—Zn_2SiO_4.

in electron penetration into the phosphor and the consequent lowering of the excitation density per unit volume.

b. Phosphor Deterioration at High Voltages.[33] At high voltages phosphors of the Mn class are more stable under electron bombardment than those of the Ag class. At the high voltages used in television projection kinescopes the screen deterioration is due to a destruction of the phosphor material. The bombarding electrons behave as reducing agents and produce colloidal metal in the screen, other nonmetallic constituents often evaporating from it.

c. Effect of Electron Burn on Secondary Emission.[27] Prolonged electron bombardment of phosphor screens is found to lower the sticking potential and also the secondary emission ratio (η). Nelson has obtained results for a zinc silicate phosphor using 6 kv electrons which are given in Fig. 11. They show how the screen potential varies with respect to the final anode potential of the cathode ray tube after different times of bombardment. After 700 hours excitation the sticking potential, originally greater than 10 kv, has decreased to less than 7 kv.

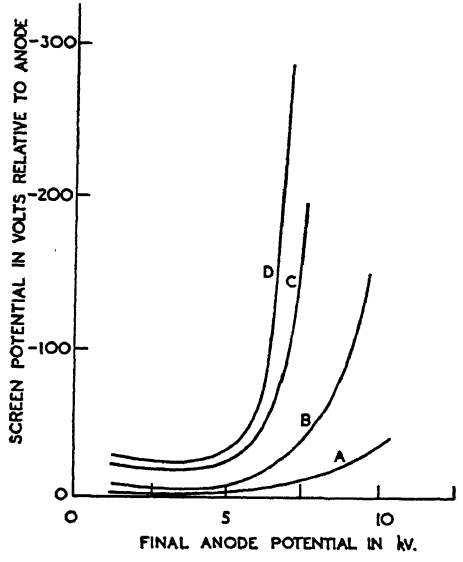


FIG. 11.—Effect of electron burn on the secondary emission of a phosphor (Nelson).
A—Initial curve for $ZnBeSiO_4$-Mn, B—After 300 hours,
C—After 500 hours, D—After 765 hours.

d. Ionic Bombardment. Phosphor screens are also affected deleteriously by ion bombardment. The ions are from other parts of the tube (glass walls, electrodes etc.) and are accelerated to the screen by the applied voltage. They produce the ion spot in the center of the screen so well known to users of cathode ray tubes. Positive ion contamination of the tube may diminish the cathode emission by bombardment.

III. THEORETICAL STUDIES OF CATHODOLUMINESCENCE

1. Theoretical Model for Crystalline Phosphors

a. Electronic Energy States in Crystals. As stated in Section I, 3, the luminescence emission of phosphors is due to special centers formed by

the activator. When the whole of the luminescence process is confined to these centers it is easy to understand.[34] However, excitation of phosphors of the conventional classes by short wavelength ultraviolet radiation or by particles frees electrons from the lattice atoms and from the luminescence centers. These electrons then move through the phosphor crystals and give rise to the observed photoconductivity. In this case emission of luminescence will depend on the return of excited electrons to luminescence centers. We thus see that the whole of the luminescence process from absorption to emission is complex. In order to understand it we consider the energy states of the electrons taking part in the process in their normal and excited states in the crystal lattice and in the activator centers. To do this we apply what is known as the *collective electron* or energy band model of an insulating crystal. The term collective is used because the crystal is considered as one large unit and the energies of its electrons are treated collectively. It was first used to explain luminescence in photoconducting phosphors by Johnson[35] in 1939 and is based on the original wave-mechanical treatment of perfect crystals due to Bloch.[36] It postulates that the electrons of the matrix lattice atoms of a crystal are not localized on those atoms but are shared by the whole crystal or, to be more precise in the case of ionic crystals, shared by all ions of the same type. The energies of these electrons falls into bands of permitted energy states separated by bands of forbidden states. In an insulator the allowed bands are either completely filled or completely empty. We have depicted the energy bands for such a crystal and for the valence electrons only in Fig. 12. The highest filled band indicated is in the case of the alkali halide and other simple ionic crystals occupied by electrons belonging to the negative ions. The empty band above the forbidden region is due to the energy states of the positive ions.[37] Although the electrons in the lower band are free to move through the crystal they cannot contribute to conduction in an applied field of ordinary strength since they cannot take up new energy states in the band. They contribute to conduction when raised into the empty band above by suitable energy absorption (short wavelength ultraviolet radiation etc.). This band is therefore called the conduction band. The removal of electrons from the filled band creates positive holes there that can also contribute to a current and play a part in luminescence processes.

b. Lattice Imperfections and Impurities in Crystals. In phosphors and in most crystalline materials the lattice is far from perfect, and its defects give rise to localized energy states for electrons. These may lie in the forbidden energy region. When such states occur near the conduction levels they can capture electrons which move in the conduction band; therefore, they are known as electron traps. Other lattice defects

may constitute centers in which excited electrons, captured from the conduction band, can lose their energy in a series of vibrational energy exchanges with the surroundings. For this reason we depict these centers in the left hand side of Fig. 12. The activator centers also provide localized levels for electrons and we show them in the figure as having a ground state and one excitation state for simplicity. In general these states have many levels. The activator centers may also create electron traps in their neighborhood and in recent studies it has been shown that these traps rather than those attributed to lattice defects

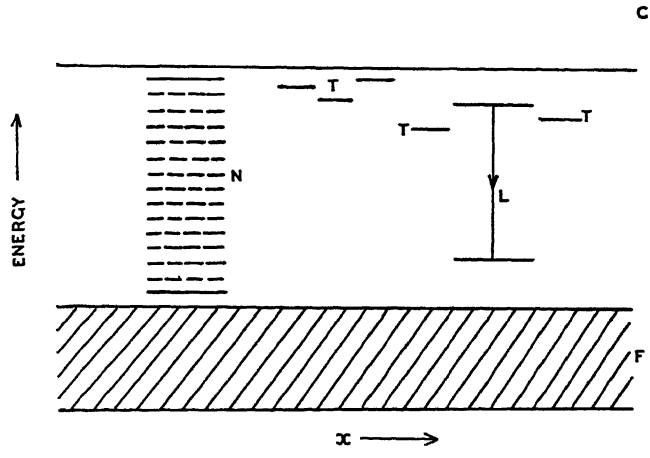


FIG. 12.—Electron energy model for a crystalline phosphor.
C—Conduction band, F-Filled band. L-Luminescence center.
N—Nonradiative center, T-Electron traps.

are important in the luminescence process.[38] Electron traps give rise to the phosphorescence and thermoluminescence characteristics of phosphors since by their storage of electrons and the subsequent release of the electrons by thermal agitation they delay the emission process. In general there is a distribution of trapping levels in phosphors and recently methods have been devised to determine it from experiments.[2]

2. Comparison of Optical and Electron Excitation of Phosphors

The above theoretical model may now be used to interpret the cathodoluminescence characteristics of phosphors and we begin with a consideration of luminescence efficiency under optical and electron excitation. The theoretical picture receives experimental support from ultraviolet absorption measurements on conventional cathode ray tube

phosphors.[7,10,39] The first strong absorption band in the ultraviolet is due to electron transitions from the highest filled band into the conduction band (this may not be the case in all phosphors, e.g., the alkali halides[37]). The longer wavelength regions may contain absorption bands due to the lattice defects, electron traps[40] and luminescence centers. If we plot the efficiency of luminescence excitation against the ultraviolet wavelength then a curve is obtained with a maximum at the edge of the ultraviolet absorption band.[41] At shorter wavelengths the efficiency falls to about one seventh of its maximum value although the absorption in the phosphor is much higher and the energy of the absorbed quanta greater. Thus when electrons are excited into the conduction band they need not lose their energy radiatively in the luminescence process. We infer that they are captured in nonradiative centers. From the above evidence and that obtained independently in other experiments of the author (to be published) we must assume that the number of nonradiative centers in phosphors is several times the number of luminescence centers. Cathodoluminescence involves the production of free electrons in the conduction band by the primary electron beam and so we expect the luminescence efficiency to be of the same order as that for optical excitation by ultraviolet radiation of short wavelength. In both cases the absorption coefficients are relatively high ($\geq 10^4$ mm.$^{-1}$) compared with those for long wavelength ultraviolet radiation in phosphors (~ 1 mm.$^{-1}$). Thus the density of excitation in the phosphor is relatively high. When electron densities in the conduction band are high then direct recombination of these electrons with positive holes in the filled band may be more likely than for conditions of low electron densities. In view of the lattice defects which exist in phosphors it is more likely that electrons will recombine with localized centers (not luminescence centers) to which the positive holes have migrated and thus return to their ground states without emission of luminescence. It is probable from recent unpublished studies that only electrons in the conduction band, and not those excited within luminescence centers due to the activator impurity, can reach these nonradiative centers. This will happen under cathode ray or short wavelength ultraviolet excitation. Hence we find similar efficiencies for these different excitations.

3. Mechanisms of Electron Penetration and Secondary Emission in Phosphors

a. The Penetration of Electrons into Solids. For many years there has been a great interest in the way in which an electron or other charged particle loses its energy when it penetrates a solid. The early experiments were carried out by using thin metal foils and electrons of about

50 kv energy. They led to the well known Thomson-Whiddington law connecting the depth of penetration of the electron with its initial velocity and its velocity at the depth x. We shall restate this law using electron energies instead of velocities. If V_i is the initial energy and V the energy at depth x then the relation is given as follows:

$$x = a(V_i^2 - V^2) \tag{2}$$

where a is a constant for a given material. Its detailed form was derived by Bethe.[42] It is found to be approximately proportional to the inverse of the density of the material. We do not know whether this law applies to electrons of relatively low energy (0–10 kv). Other studies by Stinchfield (see III), based on experiments by Terrill[43] give relations for the electron current variation with depth. At the beginning of penetration the current density i is related to its initial value i_0 and the penetration depth x thus:

$$i = i_0 \cdot \exp(-\alpha x) \tag{3}$$

where α is a constant dependent on the material and on the electron energy. From Terrill's experiments its variation with energy (mean energy values used) is obtained and a revised relation to cover all ranges of depths is given as follows:[44]

$$i = i_0 \exp\left[-k \cdot \frac{x/x_0}{\{1 - (1 - x/x_0)^{\frac{1}{4}}\}^4}\right] \tag{4}$$

Since these relations are obtained from experiments on metals and for electrons of relatively high energies, we do not know whether they apply to the electron energies in cathodoluminescence studies or to insulating materials exclusive of phosphors.

b. Secondary Electrons and Secondary Emission. We must include in the discussion the processes of secondary electron production and secondary emission from solids when bombarded by electrons. We have given the general dependence of secondary emission on the energy of the incident primary electrons in the curve of Fig. 9a. When primary electrons penetrate into a solid they produce secondary electrons along their path. The density of these electrons will vary along the primary electron path in a similar way to that in gases. The amount of secondary electron production should be approximately proportional to the inverse square of the primary electron energy at the particular place in its path. Thus the greatest number of secondary electrons will be produced towards the end of the path of the primary electron in a solid. A schematic form of the variation of secondary electron production with path length is given in Fig. 13. It is obvious that as the incidence energy

of the primary electrons is increased the bulk of the secondary electron production will occur at greater distances from the surface of the solid. From these considerations we can give an explanation of the form of Fig. 9a. When the incident electrons have very low energies they produce very few secondary electrons, which also have very low energies. These secondaries will have great difficulty in escaping from the surface of the solid since their energy is of the same order as that required to overcome the surface forces Thus the secondary emission will be very small, insufficient to prevent the material from charging up. As the primary electron energy is increased the surface forces will become less

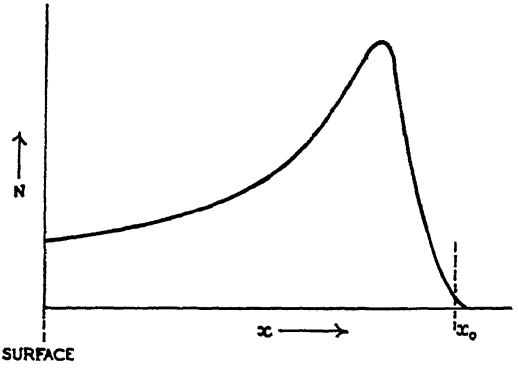


FIG. 13.—Variation of secondary electron production (N) with penetration (x) of primary electrons into matter.

important and the secondary emission will rise with this increase, as shown in the figure. Now, we must explain the maximum in the curve and the subsequent decrease in emission. Since the secondary electrons are mainly produced at greater depths from the surface, as the energy of the incident electrons increases a penetration will be reached for the primary electrons such that the secondaries produced will be unable to reach the surface. This will cause a decrease in emission as observed and will occur when the depth at which secondaries are produced becomes comparable with their mean free path in the solid. Thus we expect the emission to become small as the primary electron velocity is increased and we also expect that the sticking potential will depend on the mean free path. We consider this further in the next paragraph.

c. Secondary Electrons in Metals and in Phosphors and Insulating Crystals. In Section II, 4, *a* we already stated that the secondary emission ratio η and the sticking potential V_L are generally greater for insulators than for metals. These differences are explained by the

different behavior of secondary electrons in the two types of solid. In metals, primary and secondary electrons not only make collisions with lattice atoms or ions but more so with the large number of electrons in the conduction band of the metal. The mean free path of the secondary electrons will be largely determined by their collisions with the 'electron gas' and will be relatively independent of the metal temperature since the gas is degenerate. The usual value for the mean free path is of the order of 10^{-5} to 10^{-6} cm.[45] In an insulator such as a crystalline phosphor there are normally no electrons in the conduction band and even under electron bombardment the number of secondary electrons in this band rarely approaches that in metals. Thus collisions between electrons can usually be neglected and the mean free path of secondary electrons of more than a few volts energy is determined by their collisions with lattice ions and irregularities such as impurity centers. The mean free paths for such electrons in insulating crystals is thus somewhat larger than that in metals although of the same order. Thus in an insulator more electrons can reach the surface from greater depths below it than in metals and this explains the large values of emission and sticking potential found for insulators. The secondary emission of insulators is dependent on temperature since the increased lattice vibrations shorten the mean free paths of conduction electrons. An increase in the number of lattice defects also decreases the secondary emission.[29]

d. Complexity of Secondary Emission Processes in Phosphors. The fundamental treatment of secondary electron emission data for phosphors is rendered very difficult by the complex structure of the phosphor screens normally used. In both metals and insulators the secondary emission varies markedly with the angle of incidence of the primary electrons. This variation is shown schematically in Fig. 14. In Fig. 15 we give a diagrammatic explanation of this variation. Although for a given primary electron energy the range in the solid is fixed the orientation of the electron beam with respect to the surface will determine the depth of the end of the path below the surface. Thus for glancing angles secondary electrons will be produced at places much nearer to the surface than for normal incidence and the secondary emission will be correspondingly greater. In a phosphor screen the individual crystals and their faces are orientated at all angles to the electron beam and so the conditions for secondary emission are complex. In general the primary electrons will mostly enter the phosphor at non-normal incidence. This complexity of the phosphor surface makes theoretical treatment of both secondary emission and luminescence processes very difficult. For the purpose of fundamental research much simpler systems such as those for single crystals are needed in experiment.

e. Electron Penetration and the Production of Luminescence. The chief aim of the above discussion in paragraphs *a*, *b*, *c*, and *d* has been to provide a basis for consideration of the relation between the emission of luminescence and the energy and current density of the primary electron

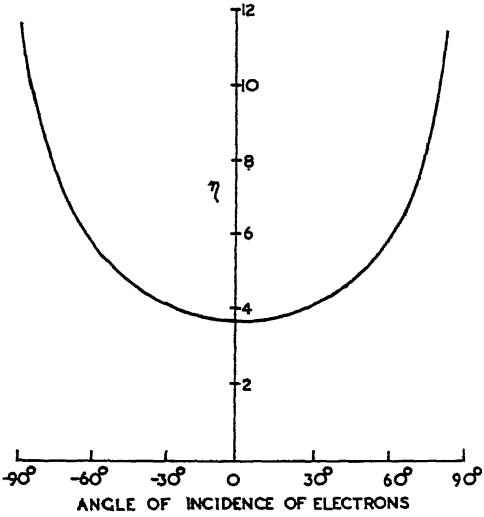


FIG. 14.—Effect of angle of incidence of primary electrons on secondary emission from solids. (η is secondary emission ratio.)

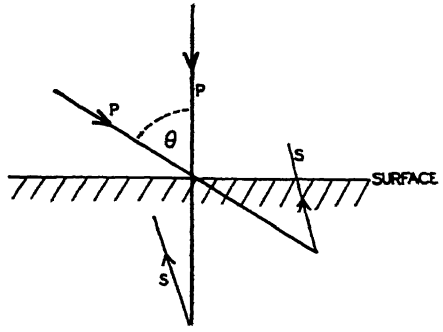


FIG. 15.—Diagrammatic explanation of Fig. 14.
P—Primary electrons, S—Secondary electrons,
θ—Angle of incidence of primary electrons.

beam. In particular we wish to explain why the simple formula given in eq. 1. holds for phosphor screens and how its constants depend on the phosphor type and the particular constitution of each specimen. It will be evident that energy is transferred from the primary electrons to

the luminescence centers by means of the secondary electrons produced in the phosphor. Some of these electrons are, as we have seen, used in the secondary emission process. The remainder will lose energy in the phosphor until they have thermal velocities and energies (0.025 ev at room temperature). They are then captured either by luminescence centers to cause emission or by nonradiative centers in which case their energy is dissipated thermally in the phosphor lattice. The proportion of the nonradiative centers is large compared with the luminescence centers, as we have seen in Section III, 2 (ratio 7:1). Using this general picture of the whole process we now consider the attempts made to derive eq. 1 from theoretical considerations. The first of significance is due to Fano.[46] He attempted to show that such a relation with an index n equal to 2 could be derived from theory. However, he assumed that the value of 2, which holds experimentally for some zinc silicate specimens,[20] is universal for phosphors. Thus his rather elaborate treatment and assumptions do not lead to more than a particular form of eq. 1. He also assumed that the energy of excitation was handed over to the luminescence centers, not by electron capture but by an exciton wave.[47] Some of his assumptions are, however, of interest. For example, he considers that electrons produced in surface layers of phosphors or reaching them from the interior are unavailable for luminescence processes. This is a reasonable assumption since nonradiative processes are more likely to occur in these regions where the lattice distortions and defects are a maximum. His explanation of the power law of eq. 1 (he neglects the dead voltage in his treatment) is based on the assumption that the probability of luminescence increases as the secondary electrons are produced at greater distances from the surfaces of the phosphor crystals. This is equivalent to saying that the ratio of nonradiative centers to luminescence centers varies with the distance from the crystal surface. We have represented this schematically in Fig. 16 showing the probability of radiative energy loss by secondary electrons as a function of their distance from the surface. We do not know yet what the exact form of the variation is but it will obviously help to determine the value of the index n in eq. 1.

More recently Strange and Henderson[15] have discussed the difficulties attendant on the theoretical derivation of eq. 1. They begin by assuming that in the excitation—emission process energy is conserved but they invalidate their arguments by neglecting the energy lost in nonradiative transitions and also by secondary electron emission. Since nonradiative processes must decrease with primary electron energy, according to experimental results, they cannot be assumed to be constant and therefore omitted from consideration of energy transfer. These authors have

attempted to show that whatever electron current vs. penetration relations are used, values of the index n different from unity cannot be obtained. However, this treatment involves their fallacious use of the principle of conservation of energy. They have made some important observations in their discussion. For instance, it is difficult on the basis of the model which we have established above to explain why the index n should not be affected by current saturation effects and why the latter should be independent of the electron energy. We must await experiments on simpler phosphor systems to settle these various points.

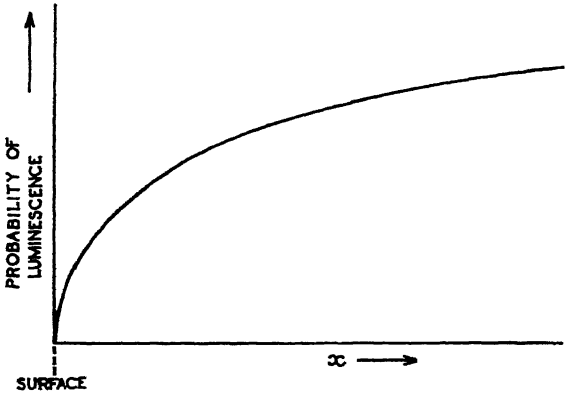


FIG. 16.—Variation of probability of luminescence with distance (x) from a phosphor surface.

4. Theoretical Interpretation of Rise and Decay Processes

a. Initial Decay Processes. An exponential decay of phosphorescence which has a rate independent of intensity of excitation or phosphor temperature is characteristic of electron transitions within the luminescence centers of a phosphor. It is likely that the initial α and β decay processes in cathodoluminescent materials, distinguished by Strange and Henderson,[15] are of this nature. There is independent evidence from decay measurements after ultraviolet excitation of phosphors to confirm this assumption for the β decay in phosphors of the Mn class. For example, we may consider the close agreement between the results of Strange and Henderson[15] and those of Fonda[23] for the decay process of willemite phosphors after cathode ray and ultraviolet excitations, respectively. At present there are no corresponding measurements for the Ag types of phosphor. The increase in the proportion of the α process with current density has been attributed by Strange and Henderson to the increase in collisions of the second kind in phosphors. We would rather suggest

that at high current densities the luminescence centers are more distorted by the high density of secondary electrons around them. Since the low probability of the electron transition in the luminescence centers determines the rate of the process and depends on the symmetry of the center configuration, we would expect the decay rate to increase if high concentrations of excited electrons in the crystal distorted the centers. The effect would be less marked in phosphors of the Mn class, as is observed, since the centers in these phosphors are more self contained and structurally shielded from lattice disturbances and the effects of high electron concentrations.

b. The Initial Rise of Luminescence. When we consider only the β process in the rise of luminescence with time during excitation we find that its rate increases with current density and is not determined solely by the decay constant. The theoretical derivation of the form of the rise of luminescence with time has been given by several workers[15,19] but we give the relation derived by Strange and Henderson. The luminescence intensity L at a time t after excitation begins with a current density of i is given by:

$$L = \frac{A \cdot B \cdot i \cdot N}{(A \cdot i + B)} \{1 - \exp - (A \cdot i + B)t\} \tag{5}$$

where A is the probability of a luminescence center being excited, B is the decay constant for the β process and N is the number of luminescence centers per unit volume. It is assumed that a very thin phosphor layer is used. We see that the rate of rise $(A \cdot i + B)$ increases linearly with the current density but also depends on the decay constant. In practice it has been found that the rate of rise is more rapid than that predicted from this theoretical relation. This discrepancy remains even when the correction for the presence of an α process is made. At present the reason for this lack of correlation between theory and experiment is not known. However, no theoretical treatment of the initial rise of luminescence has yet been attempted which includes the effect of nonradiative processes on the rate of rise of emission with time.

c. Phosphorescence of Long Duration. Figure 8 shows that the decay of phosphorescence after excitation to a given brightness is much more rapid for electron excitation than for ultraviolet excitation. We can give a satisfactory explanation of this difference in terms of electron traps, which were discussed in Section III, I. We shall refer to traps which can store electrons for long times as deep and those which only hold electrons for a short time as shallow. When excitation is due to electrons the energy is absorbed in a relatively thin layer of phosphor and the excitation density is high. Thus the available deep traps in the volume

excited are saturated and a large proportion of shallow traps are filled. We find a saturation of the afterglow at long decay times due to the trap saturation and a rapid decay at shorter times because the electrons are released rapidly from shallow traps after excitation ceases. For ultraviolet excitation the volume of phosphor excited is much greater and the excitation density in it is much smaller than for electron excitation. Thus many more deep traps are available and less shallow traps are filled in proportion by the excited electrons. In practice the relatively low efficiency of electron excitation for producing phosphorescence has been overcome by the use of cascade screens.[3] The incident electrons excite a phosphor layer which emits short wavelength visible radiation. The latter excites the long afterglow phosphor layer, underneath the "exciter" phosphor layer, nearest to the glass wall of the tube and so use is made of the more efficient optical excitation of phosphorescence. In a recent paper Leverenz[48] has attempted an explanation of the difference between electron and ultraviolet excitation of phosphorescence different from the above. Since the bulk of the absorbed energy is lost in non-radiative transitions he assumes that the energy liberated thermally by this process causes local heating of the phosphor and so empties the electron traps more rapidly. His explanation is not easily proved by experiment and does not include an explanation of the different light sums and saturation effects for the two types of excitation. We prefer the more reasonable explanation in terms of relative excitation densities given above.

5. Deterioration Processes in Phosphors

Very few theoretical studies have been made of the basic mechanisms of electron burn of phosphors. There is no doubt that both high and low voltage burn is accompanied by a formation of colloidal metal particles in the phosphor crystals. However, the colloidal deposit cannot account for the fall in efficiency by its filter action since visible darkening is not seen until a considerable decrease in efficiency has occurred. Reversible effects observed in low voltage burn[32] show that in its early stages the burn involves a metastable process. This has a parallel in the silver halide photographic process where the visible image can be removed by long wavelength irradiation. In phosphors heating produces a partial restoration of efficiency in some cases. The general nature of darkening in phosphors of the zinc sulfide types has been discussed by Gordon, Seitz and Quinlan[49] but their observations are not particularly directed towards electron burn. One of the chief points of theoretical significance with respect to low voltage electron burn is that the most susceptible phosphors belong to the Mn class. Within this

class it is also significant that the greatest susceptibility is shown by phosphors with the longest exponential decays (fluorides). This would be explained if the interaction between electrons and luminescence centers necessary to produce deterioration can only occur when the center is excited. The long life time of a center would then increase the probability of burn.

IV. Present Trends of Practical and Fundamental Research

1. Developmental Studies of Practical Interest

In this section we discuss the trend of researches which have as their aim the provision of better phosphors for use in cathode ray tubes.

a. Systematized Studies. Quite recently a series of three papers has been published which shows the trend of empirical studies of phosphors.[50] Before the late war the available phosphors for television screens were relatively limited and development was somewhat hampered by the inability to provide time and man power for extensive studies of particular phosphor systems. There was also the difficulty of reproducing new types of phosphors. With the growing realization that "hit and miss" methods of phosphor study were becoming less and less productive, and with the advent of laboratories specially designed for phosphor research, there has been an increase in systematic investigations. The three papers quoted above provide an indication of the value of this approach. In this particular case the phosphor class chosen for study was that comprising the zinc-cadmium sulfide-selenide phosphors with various activators. The effects of making small but carefully controlled changes in phosphor constitution have been studied by measurements of the changes in luminescence characteristics. By this means the whole range of possible phosphors and their properties is investigated and enables the optimum specimen to be selected for any particular cathodoluminescence application. The experimental results are presented in the form of contour diagrams showing the relation of a particular luminescence characteristic to the phosphor composition. We give as an example Fig. 17 which shows how the peak wavelength of the cathodoluminescence spectrum varies with phosphor constitution; the system under investigation is that designated as Zn:Cd:S:Se:Ag. These diagrams form compact sources of relevant information. Their production involves the preparation of a large number of phosphors and is necessarily a long term study. The form of such contour diagrams will obviously be of considerable value in theoretical interpretation of phosphor processes.

b. High Brightness, High Power Cathode Ray Tubes. It is not possible to discuss in detail the present relation of phosphor research to the

demand for large scale projection television. In the cathode ray tubes used for this purpose the electron energies range up to 80 kv and current densities are correspondingly high. Hence the operating conditions will limit the choice of suitable phosphors to those having sufficient stability and luminescence efficiency under these conditions. Because of this limitation, at present there is no possibility of simultaneously making a choice of spectral emission characteristics and decay constants. At such

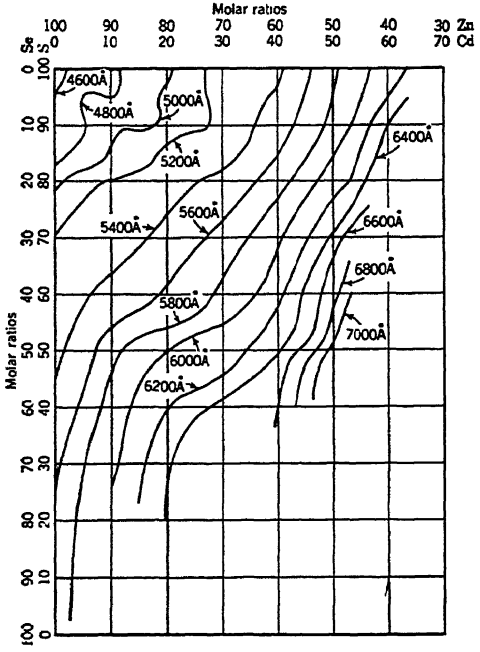


FIG. 17.—Contours of peak wavelengths of cathodoluminescence (6 kv electrons) current density 1 μ a/cm.^{2}) for the phosphor series Zn, Cd:S, Se:Ag (Schrader, Lasof, and Leverenz).

temperatures as are reached at the screen of a projection cathode ray tube the temperature quenching of fluorescence of the phosphor becomes important. We might mention here that the necessary studies of the practical behavior of phosphors at these temperatures has provided information of fundamental interest. It has been found that the variation of fluorescence of phosphors under cathode ray excitation is different from that under ultraviolet excitation.[51] We illustrate this by Fig. 18. This shows the fluorescence vs. temperature curves for a zinc sulfide phosphor in the temperature region where thermal quenching is considerable. It will be seen that the luminescence persists to higher

temperatures when cathode ray excitation is used. This difference is very similar to that found between the curves due to different wavelengths of ultraviolet excitation for photoconducting phosphors inclusive of zinc sulfides.[52] Experiments on thermal quenching under cathode ray and ultraviolet excitation can furnish important information about the nature of the nonradiative centers in phosphors which limit the efficiency of cathodoluminescence.

c. Research on Phosphors for Radar Applications. It is not possible, for obvious reasons, to obtain an up to date view of the development of phosphors for radar cathode ray tubes. The properties of the ideal

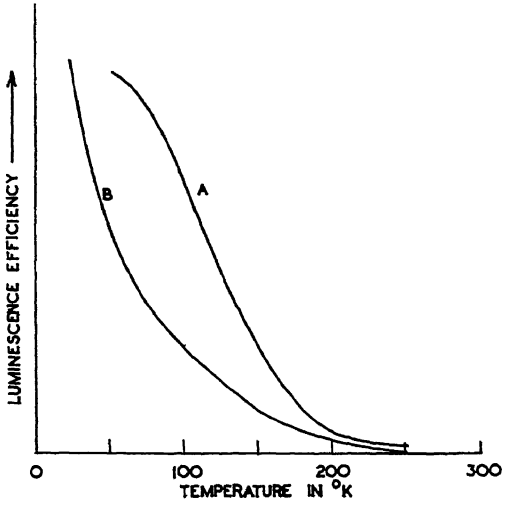


FIG. 18.—Temperature dependence of phosphor luminescence for different modes of excitation (Schrader, Lasof, and Leverenz). A—Electron excitation, B—Ultraviolet excitation.

phosphor have long been known for the various radar indicator systems but the present materials fall far short of the desirable phosphor. The nearest approach to an ideal afterglow characteristic is an exponential decay, whose half life is comparable with the radar scan period. For some applications this can be realized but in general phosphors with hyperbolic decay of afterglow have to be used. The advent of cascade screens alleviated this position somewhat. We would suggest that useful study might be expended on methods of controlling the afterglow of suitable phosphors so that the afterglow intensity remains constant during each scan but decays rapidly in the short interval before the beginning of the following scan. A certain degree of success could be achieved by using infrared radiation applied from outside the cathode ray tube to

enhance the phosphorescence. Undesirable phosphorescence at the end of each scan could be removed by application of quenching radiation to the screen. Among other present improvements suggested there is the use of phosphor screens as "dark trace" indicators.[53] The phosphor is excited by ultraviolet light, being maintained at a temperature just below the region of thermal quenching. The electron beam then produces a decrease in luminescence instead of excitation. It would be possible to control the "afterglow" of such screens more readily than in conventional phosphor screen usage and the general brightness levels obtainable are very much higher. This improves the contrast of the picture for given input signal strengths.

2. A New Approach to Fundamental Research

The interest of the pure physicist lies in the mechanisms of cathodoluminescence and its related phenomena. In this respect, as we have already observed, the past and present experimental studies do not lend themselves to fundamental interpretation because the phosphors used have been in the form of screens of complex structure. For this reason, and because a better knowledge of fundamental processes often leads to practical improvements, there is a need for a new approach to basic research. We would suggest that there should now be a departure from the use of conventional phosphor screens in spite of their high luminescence efficiency. Instead use should be made of single crystals of large size of insulating solids which show luminescence under cathode rays, although such emission is relatively feeble. Experimental difficulties of measurement are easily overcome by the use of modern photomultipliers. Such crystals, for example, the alkali halides, can be cut, polished, and mounted with ease, and if used would remove the complex surface conditions from the experiment. More use should also be made of evaporated layers of phosphors of controlled thicknesses. Some success has already been reported in evaporating fluoride phosphors.[14] This reversion to materials of simple form proved fruitful in the study of electronic processes in crystals, e.g., photoconductivity.[37] We would also suggest that more attention might be paid to the secondary emission characteristics of insulators and the mechanism of electron penetration. What has been accomplished for metals by the use of thin foils may be achieved for insulators by the use of thin lacquer films and evaporated layers.

V. Conclusion

In this review we have attempted to present the core of the knowledge of cathodoluminescence and the materials of interest with respect to it. It has not been found possible to deal adequately with the practical

aspects of the subject but it was thought necessary to stimulate interest in the basic research which lies behind the considerable advance made in phosphor quality for cathode ray tube use. It is our opinion that this aspect will become increasingly important in the future.

ACKNOWLEDGEMENTS

Thanks are due to the various authors and publishers for permission to reproduce some of the above diagrams, in particular to H. W. Leverenz for Figs. 1, 2, and 4, to J. H. Schulman for Fig. 3, to J. W. Strange and S. T. Henderson for Figs. 5, 6, and 7, to J. H. Nelson for Fig. 11, and to The Physical Society, London, The American Institute of Physics, and The Editors, *Zeitschrift für Physik*. Figures 17 and 18 are reproduced by permission from "Preparation and Characteristics of Luminescent Materials," edited by G. R. Fonda and F. Seitz, published by John Wiley and Sons, Inc., 1948.

REFERENCES

1. Symposium on Luminescence. *Trans. Faraday Soc.*, **35**, 2–238, 1939.
2. Randall, J. T., and Wilkins, M. H. F. *Proc. Roy. Soc.*, **A184**, 366–407 (1945).
3. Leverenz, H. W. R.C.A. Final Radar Screen Report, PB-25481, Washington, 1945.
4. Rothschild, S. *Trans. Faraday Soc.*, **42**, 635–642 (1946).
5. Kröger, F. A. *J. Electrochem. Soc.*, **93**, 156, 171 (1948).
6. Garlick, G. F. J., Wells, A. F., and Wilkins, M. H. F. Phosphor constitution and electron traps. *J. Chem. Phys.*, **17**, 399 (1949).
7. Kröger, F. A., and Gisolf, J. H. *Physica*, **6**, 84–96, 369–379, 779–784 (1939); **8**, 805–809 (1941).
8. Radar Convention Proceedings *J. Instn. Elect. Engrs.*, **93**, IIIA, part I (1947).
9. Schulman, J. H. *J. Appl. Phys.*, **17**, 902–908 (1946).
10. Kröger, F. A. *Physica*, **6**, 764–778 (1939).
11. Fonda, G. R. *J. Phys. Chem.*, **43**, 561–577 (1939); **44**, 851–861 (1940); **45**, 281–288 (1941); **46**, 1–8 (1942).
12. Hill, C. G. A. *Trans. Faraday Soc.*, **42**, 685–689 (1946).
13. Fonda, G. R. *J. Phys. Chem.*, **48**, 303–307 (1944).
14. Williams, F. E. *J. Opt. Soc. Amer.*, **37**, 302–307 (1947).
15. Strange, J. W., and Henderson, S. T. *Proc. Phys. Soc. Lond.*, **58**, 369–401 (1946).
16. Lenard, P., and Saeland, S. *Ann. Phys. Lpz.*, **28**, 476–502 (1909).
17. Debye, P. (Editor) Interference of Electrons. Blackie, London, 1931.
18. Piore, E. R., and Morton, G. A. *J. Appl. Phys.*, **11**, 153–157 (1940).
19. Martin, S. T., and Headrick, L. B. *J. Appl. Phys.*, **10**, 116–127 (1939).
20. Brown, T. B. *J. Opt. Soc. Amer.*, **27**, 186–192 (1937).
21. Nottingham, W. B. *J. Appl. Phys.*, **8**, 762–778 (1937); **10**, 73–83 (1939).
22. Nelson, R. B., Johnson, R. P., and Nottingham, W. B. *J. Appl. Phys.*, **10**, 335–342 (1939).
23. Fonda, G. R. *J. Appl. Phys.*, **10**, 408–420 (1939).
24. Garlick, G. F. J. Luminescent Materials. Oxford Univ. Press, Oxford, 1949.
25. Kröger, F. A., Hoogenstraaten, W., Bottema, M., and Botden, Th. P. J. *Physica*, **14**, 81–96 (1948).
26. Klasens, H. A. *Trans. Faraday Soc.*, **42**, 666–668 (1946)
27. Nelson, J. H. *J. Appl. Phys.*, **9**, 592–599 (1938).

28. Geyer, K. H. *Ann. Phys. Lpz.*, **41**, 117–143 (1942).
29. Knoll, M., Hachenberg, O., and Randmer, J. *Z. Phys.*, **122**, 137–162 (1944).
30. Bruining, H. Secondary Electron Emission in Solids. Berlin, 1942.
31. Epstein, D. W., and Pensak, L. *RCA Rev.*, **7**, 5–11 (1946).
32. Grotheer, W. *Z. Phys.*, **112**, 541–559 (1939).
33. Leverenz, H. W. *RCA Rev.*, **5**, 131–175 (1940).
34. Garlick, G. F. J., and Wilkins, M. H. F. *Proc. Roy. Soc.*, **A184**, 408–433 (1945).
35. Johnson, R. P. *J. Opt. Soc. Amer.*, **29**, 387–391 (1939).
36. Bloch, O. *Z. Phys.*, **52**, 555–600 (1928).
37. Mott, N. F., and Gurney, R. W. Electronic Processes in Ionic Crystals. Oxford Univ. Press, Oxford, 1940.
38. Garlick, G. F. J., and Gibson, A. F. *Proc. Phys. Soc., Lond.*, **60**, 574–590 (1948).
39. Kröger, F. A. Luminescence in Solids containing Manganese. Campen, Amsterdam, 1940.
40. Hoch, A. *Ann. Phys. Lpz.*, **38**, 486–495 (1940).
41. Garlick, G. F. J., and Gibson, A. F. *Nature, Lond.*, **161**, 359 (1948).
42. Bethe, H. *Ann. Phys. Lpz.*, **5**, 325–401 (1930).
43. Terrill, M. *Phys. Rev.*, **22**, 101–108 (1923); **24**, 616–621 (1924).
44. Leverenz, H. W. Solid Luminescent Materials; paper No. 8, Wiley, New York, 1948.
45. Fröhlich, H. *Proc. Roy. Soc.*, **A160**, 230–241 (1937).
46. Fano, U. *Phys. Rev.*, **58**, 544–553 (1940).
47. Frenkel, J. *Phys. Rev.*, **37**, 17–44 (1931).
48. Leverenz, H. W. *Phys. Rev.*, **69**, 686 (1948).
49. Gordon, N. T., Seitz, F., and Quinlan, F. *J. Chem. Phys.*, **7**, 4–7 (1939).
50. Lasof, S., Schrader, R. E., and Leverenz, H. W. Solid Luminescent Materials; papers No. 11, 12 and 13. Wiley, New York, 1948, pp. 205–257.
51. Lasof, S., Schrader, R. E., and Leverenz, H. W. Solid Luminescent Materials; paper No. 12. Wiley, New York, 1948, p. 219.
52. Garlick, G. F. J. Solid Luminescent Materials. Wiley, New York, 1948, p. 203.
53. Leverenz, H. W. *RCA Rev.*, **7**, 199–239 (1946).

Intrinsic Dielectric Breakdown in Solids

H. FRÖHLICH AND J. H. SIMPSON

University of Liverpool, Liverpool, England, and National Research Council of Canada, Ottawa, Canada

CONTENTS

I. INTRODUCTION

The importance of solid dielectrics in the generation, distribution, and application of electricity has, for many years, provided a stimulus for the experimental and theoretical investigation of the reasons for failure of these materials under electrical stress. Such failure may be due to one or more of several different causes depending on the chemical purity of the dielectric itself, its actual physical structure (existence of cavities etc.), and the way in which the electric stress is applied to it. As a result, it is often found in industrial or engineering practice that, under what appear to be similar conditions, electric strengths differing by a large factor may occur—and that these values of electric strength are, in general, considerably lower than those obtained under carefully controlled conditions as regards purity of materials and uniformity of applied field. In order to obtain consistent results in the experimental investigation of dielectric strength it is obviously necessary to resort to these special conditions and to eliminate those factors which, in practical

applications, produce inconsistent results. Thus, concentration of electric stress at the edges of electrodes, electric discharges in cavities within the dielectric, and breakdown in the adjacent insulating medium must be obviated by the use of carefully constructed electrodes and test specimens, and by the choice of a medium having the proper electrical characteristics. (Details of the necessary experimental precautions are given in a forthcoming book by Whitehead.[1])

Even when all precautions have been taken it is sometimes found that the dielectric breakdown strength depends on the size and shape of the specimen. Usually, when this is the case, the temperatures are fairly high and the conductivity of the material relatively large so that failure may be due to "breakdown through thermal instability." This type of breakdown occurs if the rate of dissipation of energy in the material due to the current produced in it by the field (rate of production of Joule heat) exceeds the rate at which heat can be carried away by thermal conduction through the specimen to the surrounding medium. Since the rate at which heat is conducted away depends upon the size and shape of the specimen the field strength for this type of breakdown depends on these factors also, and this "thermal" type of breakdown cannot be regarded as an inherent property of the material. This type of breakdown has been treated rather fully in the literature and will not be considered further here.[1,2]

Normally the value of dielectric strength obtained under carefully controlled conditions does not depend upon the size and shape of the specimen, except for very small thicknesses. It may then be regarded as a characteristic constant of the material depending only on parameters which determine the state of the material (e.g., temperature), the microscopic structure, and the chemical composition. In this case we have the "intrinsic dielectric strength," which is the highest dielectric strength obtainable under ideal conditions and represents an upper limit to the value attainable in practice. It may differ from breakdown values obtained in present day engineering practice by a factor of order 10.

Since the intrinsic dielectric strength is a constant of the material it is possible to calculate its value theoretically from our knowledge of the microscopic (atomic) structure of the specimen and the physical conditions under which it exists. The resulting theory of dielectric breakdown provides a method of increasing our knowledge of the subject which is complementary to the experimental method and a form of mutual assistance between theory and experiment occurs. Thus the absolute value of the breakdown field strength calculated theoretically helps experimentalists to set an upper limit to the values which they may expect to attain in the laboratory, and provides a method of ascertaining

that there are no serious errors in technique. Moreover, the theory makes predictions which lead to new experimental investigations and the results of these in turn suggest modifications in the theory. Such modifications are bound to be required as our knowledge of the breakdown process improves—largely because the whole breakdown phenomenon is very complex and any theory of it must necessarily contain some approximations. Thus, while we should expect the theory to be essentially correct it may not, especially in its early stages, describe accurately all details of the breakdown process. Insight into the mechanism of "intrinsic" behavior gained in this way should ultimately become of practical importance by suggesting ways in which commercial dielectrics may be improved.

Various theories have been proposed to explain the intrinsic breakdown process. One of these was the theory of breakdown due to thermal instability which, as mentioned above, has been shown to apply only under special conditions. Another, the mechanical theory, assumes that breakdown is due to mechanical rupture of the crystal caused by the forces which the electrical field exerts on its ions. According to this theory the electrical strength, like the mechanical strength, should depend, to a large extent, on the number of cracks or other crystal imperfections—a result which has been disproved experimentally. A similar lack of success has been characteristic of several other theories which need not be described here.

Experimental evidence has indicated that intrinsic breakdown is electronic in nature. In particular the fact that breakdown takes place in a very short time (of the order of 10^{-8} seconds) would seem to preclude any other explanation. An important series of experiments by von Hippel[3,4] also pointed to the possibility of an electronic phenomenon and stimulated the development of the theory to be discussed in this article. Von Hippel investigated dielectric breakdown in ionic crystals, (which are of very simple structure) including the dependence of breakdown strength on concentration in the case of mixed crystals. One of his sets of experimental results is illustrated in Figure 15. Other results are similar to these and indicate that the admixture of small amounts of one ionic crystal to another always leads to an increase in the dielectric strength. This behavior is similar to that of the electrical resistance of mixed crystals of metals. It is known that the latter is due to additional scattering of electrons (changing of direction and velocity) produced by changes in the forces acting on them when foreign ions are added to the original (pure) crystal. It is natural to assume that the rise in dielectric strength is due to the same cause, and that breakdown is therefore an electronic phenomenon. This assumption led to a theoretical study of

the movement of electrons in dielectrics and ultimately to the development of a quantitative theory of dielectric breakdown in crystals at low temperatures.

The further development of this theory as outlined in Sections III and IV led naturally to the prediction of an increase of dielectric strength with temperature as well. This result was subsequently confirmed experimentally by workers in the laboratories of the E.R.A.* who were following up the consequences of the theory, and by von Hippel and his coworkers. Other predictions of this section of the theory, for example the order of magnitude of the breakdown voltage, have also been confirmed at low temperatures. However it was always found experimentally that the breakdown strength decreased with temperature after a certain temperature had been passed—in direct contradiction to the behavior at low temperatures. This was originally thought to be the start of the thermal breakdown range but experimentalists later maintained that in many cases this was not true. Attempts to account for this latter fact led to the development of the second section of the theory outlined under case (a) in Sections III and IV. Further consequences of this theory which are discussed in Section IV have been largely confirmed experimentally.

In outlining the development of the theory we shall consider this high temperature section first as the qualitative ideas of the breakdown process are more readily understood in this case. The low temperature case will be dealt with when some familiarity with the general ideas involved has thus been acquired. However, since both parts of the theory are based upon the behavior of electrons in insulating materials, we must first discuss the properties of such electrons. This will be done in the following section.

II. Electrons in Solids

1. Electrons in Equilibrium in a Perfect Lattice

In order to obtain a general picture of the characteristics of electrons in solids we shall first consider a perfect crystal as it is seen from the point of view of modern physics. Such a crystal is defined as one having no impurities or defects in lattice structure and ions or atoms that are fixed in position (that is, having no thermal motion). As a prerequisite to obtaining the properties of such crystals it is necessary to obtain some idea of the energy states of the electrons.

It is well known that electrons in free atoms (in a dilute gas for example) have a number of possible discrete energy levels. If atoms

* The British Electrical and Allied Industries Research Association.

are brought together to form a crystal lattice, then (as was first shown by F. Bloch),[5,6] because of the forces which these atoms exert on each other (interaction), these discrete levels are broadened into bands each consisting of a large number of energy levels. Roughly speaking these bands are narrow for tightly bound electrons forming the inner shells of atoms, but relatively broad for loosely bound electrons. For states of sufficiently high energy the bands will overlap. If there are N atoms forming a simple type of lattice, then for each level in the isolated atom there will be N levels in the corresponding band. The width of an energy band is independent of the number of atoms N (unless N is very small) so that the distance between the levels within a band decreases as N increases. For a crystal of ordinary macroscopic size the levels are thus very close together and in practice one may consider a band as a continuum of energy levels between two boundaries. This is illustrated in Fig. 1.

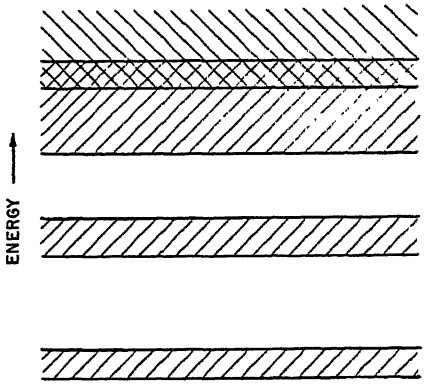


FIG. 1.—Energy bands of an electron in a perfect crystal. The shaded areas show allowed energy levels. Overlap of two high energy bands is illustrated.

When we consider the actual physical state of the electrons in a band we find that (for a perfect crystal in which each atom is fixed at its lattice point and has no thermal motion) an electron in a given level may be considered as moving through the crystal without being scattered, that is, without having its direction or energy changed by collision with the atoms of the lattice.

The number of the above closely spaced energy levels within a band is of great importance in deciding the behavior of the electrons. This is due to an important fundamental principle, known as the Pauli exclusion principle, which states that each completely defined energy level can be occupied by one electron only. The way in which these levels are defined cannot be discussed in detail here but we may observe that, in an energy band, for each state corresponding to a certain velocity of an electron, there is a state of the same energy but with opposite velocity. This will be true under all conditions, even, for instance, if an electric field exists in the crystal. Now each moving electron carries an electric current proportional to its velocity. Thus, if each state of a band is occupied by an electron the total current vanishes, regardless of the

magnitude of the applied field, because the contributions of pairs of electrons with opposite velocities just cancel. To produce a current would require changes in the velocities of some of the electrons. This would lead them into other levels of the band which, according to the above assumption, are already occupied. Hence such transitions are forbidden according to the Pauli principle. We thus reach the important conclusion that the electrons of a completely filled band cannot contribute to an electric current in the crystal. This is the case in perfect insulators. If, on the other hand, an energy band is only partially filled, a current may be produced by changing the way in which the levels are occupied. This is the case for metals, in which at least one band is not completely occupied, and their conductivity is thus accounted for.

The Pauli principle as applied to the "band theory" of solids thus accounts for the difference between conductors and perfect insulators. It should be remembered, however, that regardless of the state of the upper bands that contain electrons—whether fully occupied or not—at higher energies bands exist that would be empty in a perfect material. In an insulator, the upper full band occupied by the electrons in outer orbits (valency electrons) has a width of several *e*-volts* and the gap between the top of this band and the bottom of the first empty band is of the same order. If sufficient energy is provided (for example, by collision with a fast particle or by absorption of a light quantum) an electron may be lifted from this highest fully occupied band into the first empty band. Here it can take any one of a large number of states and so can contribute to an electric current. The empty band immediately above the highest fully occupied band of a perfect insulator is therefore called the conduction band. The distribution of bands in a metal and an insulator is shown in Fig. 2.

In the following, electrons in conduction bands will be treated as if they were completely free. The lowest level in the conduction band will thus be treated as the state in which a free electron is at rest. In a higher level, within the band, an electron will have a certain velocity v and energy $mv^2/2$ where m is the mass. This behavior of electrons in conduction bands follows as a reasonable approximation from theoretical investigations. These investigations indicate however that m should be replaced by an "effective" mass m^*, which may be different from m though usually of the same order.

Let us now consider a crystal which, at the absolute zero of temperature ($T = 0$) is a perfect insulator, that is, the highest completely filled

* The electron volt (abbreviated *e*-volt) is a unit of energy and is the amount of energy gained by an electron in passing through a potential of 1 volt. It is equal to 1.60×10^{-12} ergs.

band is followed by an empty conduction band separated from it by an energy gap $2V$. Then at a higher temperature a few of the electrons will be found in the conduction band, thus contributing to the conductivity. The number of these is proportional to $e^{-V/kT}$ and thus increases exponentially with temperature T (measured on the absolute scale). The crystal will, therefore, no longer be a perfect insulator but will have a finite conductivity that increases with temperature—in contrast to the conductivity of metals that decreases as the temperature rises. If the gap V is sufficiently wide the conductivity at normal temperatures is usually so small that the substance may, for all practical purposes, be

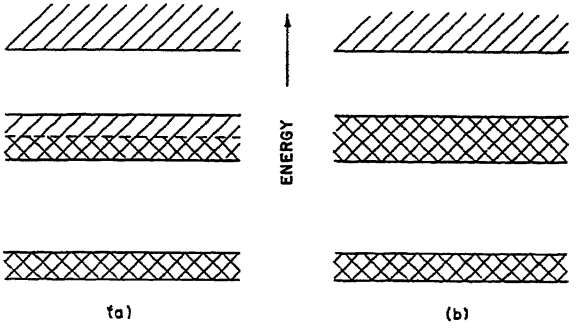


FIG. 2.—Energy bands of: (a) a metal (b) an insulator. Single shading indicates that levels are empty, double shading that they are filled. The presence of a partially filled band in the metal makes conductivity possible.

considered as an insulator—although strictly speaking it behaves as a semiconductor.

2. Effect of Lattice Defects and Impurities

The above section has been concerned with ideal or structurally perfect lattices. We now consider imperfections in structure which are important from the point of view of dielectric breakdown theory. In the first instance we shall be concerned with the simplest types of insulating solids. Later our discussion will be broadened to include more complicated materials.

All crystals, even if grown with great care, contain imperfections which include the following important types:[6]

(a) Lattice defects in a crystal containing no foreign atoms (impurities) and in which the constituents are in the correct chemical (stoichiometric) proportions. It has been shown by various authors that, in thermal equilibrium at any temperature above absolute zero, a certain number of such defects exists that increases exponentially with tempera-

ture. These defects may be vacant lattice points or atoms or ions in interstitial positions (between lattice points).

(b) Lattice defects, due to an excess of one constituent in a crystal containing no foreign atoms. The excess atoms or ions may go into interstitial positions or the lattice may be arranged so that there are vacancies at some lattice positions of one of the other constituents.

(c) Imperfections due to the presence of foreign atoms.

The importance of these imperfections from the point of view of dielectric breakdown theory lies in the fact that they change the distribution of charges in the lattice and act as trapping centers for free (conduction band) electrons. Thus it can be shown that an electron may be captured in the field of an interstitial ion or vacant lattice point and exist in discrete energy states similar to those of an isolated atom. The arrangement of energy levels in such a case is shown in Fig. 3. The situation is exactly analogous to that in a free atom, the discrete states (consisting of a ground level and higher excited levels) being at lower energies than the conduction continuum which corresponds to the continuum of energies above the ionization level of a free atom. However, due to the effect of the surrounding medium (ions or atoms, and electrons), the energy gaps between levels are much smaller than in a free atom. Now it will be recalled that the energy bands in a simple solid correspond roughly to those of a free atom. Hence we would expect the trapped electron levels to be much closer to the conduction band than the nearest full band. The excited levels of the trapped electron actually lie just below the conduction band as shown in Fig. 3. The ground level is separated from these by a gap of appreciable size but is well above the top of the highest full band.

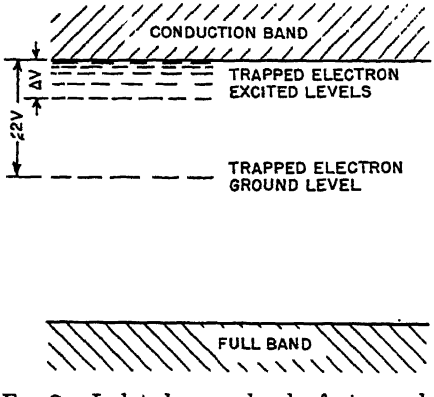


FIG. 3.—Isolated energy levels of a trapped electron.

It will be seen from this that even the purest crystal obtainable will have a number of captured electrons which may be taken up by trapping centers during the growth of the crystal or because of the effects of radiation, thermal fluctuations, etc. Such trapped electrons do not contribute to the conductivity of the crystal. However, at ordinary temperatures some of them will be ejected from the trapping centers into the conduction band and the probability of this process is quite high

because of the relatively small energy difference between the isolated states and the conduction band. Thus, in equilibrium, there will always be a certain number of electrons in the conduction band, although for very pure crystals the number of such electrons will be small.

Another type of defect which does not occur in carefully grown crystals, but which is important in many types of material, occurs when there is a drastic distortion of the lattice so that the crystal structure is regular in very small volumes only and, in general, the directions of the crystal axes in these small volumes vary from one to the next. There are thus large irregular sections between such volumes, and the distortion of the field of the lattice so produced results in the creation of a large number of trapping centers. An amorphous solid, such as glass, may be regarded as a limiting case of such a material, and it is thus apparent that we should expect a large number of trapping centers in this case. Plastic materials, built up of long chain molecules fitted together in a semi-random arrangement would also be expected to have many such centers. An appreciable number of these will no doubt be occupied by electrons which are trapped as the material cools from high temperatures during manufacture.

3. *Thermal Equilibrium of the Conduction Electrons*

It has been shown that, in insulators, there will always be a certain number of electrons in the conduction band and that these will behave similarly to free electrons of effective mass m,* which is of the order of m, the mass of an electron. The number of such electrons is very small compared to the number of possible states in the conduction band so that an electron may have practically any energy in the band. The limitation of the Pauli principle, which applies in the case of energy bands having an appreciable fraction of their states occupied by electrons, is not important in this case.

The way in which the energies of the conduction electrons are distributed in thermal equilibrium has an important bearing on the dielectric strength, and we now consider what this distribution is and how it is established. The equilibrium distribution is actually independent of the way in which energy is shared between electrons as long as such sharing is possible, and can be calculated by the methods of statistical mechanics. In this case the energy distribution is that given by Boltzmann and the number of particles having an energy E is proportional to $Ne^{-E/kt}$ where N is the total number of particles. This distribution will apply not only to the conduction electrons but to those in isolated levels as well.

The establishment of equilibrium is assumed to take place through collision processes. These will occur between isolated electrons and con-

duction electrons and between conduction electrons and lattice, as well as between the conduction electrons themselves. Two particles colliding will, in general, occupy different states before collision and after. In some cases a given particle will gain energy and in others lose it, and in equilibrium, if all particles are considered, the average number of processes in each direction will be equal. Furthermore if thermal equilibrium does not exist initially the collisions will tend to produce it since those particles having a large energy will tend to lose it to others and the inverse process will not occur so frequently. Hence the energy will be spread among all elements and the system will tend to a Boltzmann distribution.

Electrons in the conduction band thus exchange energy, by collisions, with each other and with the isolated electrons, as well as with the lattice. As has been mentioned previously the latter exchange would not occur in a perfect lattice having no thermal motion. However when thermal motion is taken into account it is found that in addition to scattering by defects in the lattice the electrons are scattered by the lattice itself, and this is connected with an exchange of energy. This is generally described as interaction with the lattice vibrations, and can be pictured as either absorption or emission of a quantum of energy $h\nu$ by the electrons if ν is the lattice vibration frequency. If lattice vibrations and electrons are in thermal equilibrium the total number of absorption and emission processes just cancel.

We now consider the behavior of the system when a weak electric field F is applied. A certain current density j will be set up in the crystal, given by

$$j = \sigma F \tag{1}$$

where the conductivity σ is independent of F. It follows from the general principles of electrodynamics that energy is transferred from the field to the electrons at a rate per unit volume per second

$$A = jF \tag{2}$$

In ordinary conduction this energy is transformed into heat energy (Joule heat) but eq. (2) does not necessarily imply this. It is a much more general relation which merely states that the energy A per second is transferred to the electrons and does not state what happens to it subsequently. In particular it makes no suggestion that the energy is transformed into heat energy of the crystal. In fact equation (2) holds for the case in which the electrons are continuously accelerated by the field and then A represents the rate of increase of kinetic energy of the electrons.

A simplified picture of the method of calculating the conductivity σ can be based on methods used in the kinetic theory of gases. We assume that an electron spends, on the average a time $2\tau_1$ seconds between collisions so that it will be accelerated for this period by the field, after which it will lose the excess velocity thus acquired. Hence at the end of the period $2\tau_1$ seconds the electron has a velocity $2eF\tau_1/m^*$ in the field direction in addition to its previous velocity. Since, at the beginning of the period (after the previous collision), this additional velocity was zero, the electron has, on the average, a velocity

$$\bar{v} = eF\tau_1/m^* \tag{3}$$

in excess of its average velocity without field. This latter quantity is of course zero. Each electron thus contributes an amount $e\bar{v}$ to the total current so that

$$\sigma = \frac{j}{F} = \frac{Ne\bar{v}}{F} = \frac{e^2\tau_1 N_1}{m^*} \tag{4}$$

since the conductivity must be proportional to N_1, the number of electrons in conduction levels. This derivation of σ is not strictly correct since the relaxation time τ_1 is not of the simple type described above and in calculating $\bar{v}$ an average over all energy states of the electrons should be made by the methods of statistical mechanics. However the result (4) is correct when τ_1 is an average quantity known as the mean relaxation time of the conduction electrons. Using this result we obtain

$$A = \frac{e^2\tau_1 N_1 F^2}{m^*} \tag{5}$$

The velocity $\bar{v}$ is, of course, the mean velocity of drift of the electrons under the applied field and is superimposed upon the ordinary thermal motion of the electron. A crude picture of the path of an electron of thermal energy with and without a field applied to the crystal is shown in Figs. 4b and 4a, respectively.

As indicated previously there are several types of collision (between conduction electrons and lattice, conduction electrons and isolated electrons, and among the conduction electron themselves) and the structure of the material will decide which type predominates. In very pure insulating crystals for instance there are few electrons in the conduction levels and a small number of trapped electrons. Consequently collisions between electrons will not be important and collisions with the lattice itself will decide the behavior of the electrons when a field is applied. In the case of an amorphous material however, the number of electrons in both conduction and isolated levels is much larger and collisions between electrons become important in deciding their mean velocity when a field

is applied. It is necessary to consider collisions between all electrons (conduction and isolated) in this case. Collisions between conduction electrons are of no importance for the destruction of the excess momentum acquired from the field because the total momentum is conserved in a collision between free electrons. This is not so, however, for collisions between conduction and trapped electrons, especially when the latter are in excited states where the distance between energy levels is small. Also, although the total number of interacting electrons is less than the total number of atoms or ions in the lattice, their effect in producing collisions is much greater, since the orbits of trapped electrons spread

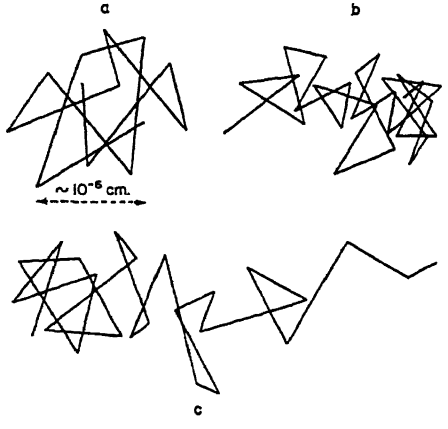


FIG. 4.—Crude picture of the motion of electrons:[2] (a) without external field; (b) in an external field when the electron has low energy ($E < E_c$): (c) in an external field when the electron has high energy ($E > E_c$). This is the runaway condition—cf. Section III.

over a large volume of the crystal. A conduction electron will traverse, on the average, many lattice distances in a crystal before it arrives in the neighborhood of an ion or atom which will scatter it appreciably, so that collision processes between electrons will predominate.

III. The Mechanism of Breakdown

In the previous section it was pointed out that when an electric field is applied to an insulator a current flows. For weak fields this current will be very small, though finite, and will increase approximately linearly with the field. In strong fields, which are still well below breakdown however, the current increases with field more rapidly than at this linear rate. The energy transferred from field to electrons is given by eq. (2). Now this equation does not stipulate how this energy may be

dissipated and it would certainly be wrong to assume that it appears as Joule heat under all conditions. However, if the crystal is in a stable condition this energy must be dissipated somehow, and this means that it must ultimately be transformed into heat and carried away by thermal conduction; since there are few free electrons, most of the heat must be conducted away by the lattice. The temperature of a solid to which an electric field is applied must therefore be higher than that of its surroundings. If the size and shape of the body are such that the conditions for breakdown are determined by this heat removal process we have the case of breakdown through thermal instability referred to in Section I. If heat dissipation is not an important factor intrinsic breakdown will occur when a sufficiently high field has been reached. We now investigate the conditions under which this may happen.

If a crystal is in equilibrium with a field applied to it, heat must be carried away by the lattice at the same rate as it is supplied by the field. Thus, in equilibrium we have

$$A = B \tag{6}$$

where B is the rate per unit volume at which energy is transferred from the electrons to the lattice. This transfer of energy from electrons to lattice is the main bottleneck in the overall transfer from field to lattice and is responsible for the instability connected with breakdown.

It will be clear from the discussion in Section II that in considering B we must distinguish between the following two cases.

(a) When the density of electrons in the conduction band and the excited levels is sufficiently high that collisions between electrons are much more frequent than between electrons and lattice vibrations.[7]

(b) When collisions between electrons are rare compared with those between electrons and lattice vibrations.[8]

In case (a), because of the larger number of collisions per unit time, the time taken for the electrons to reach an equilibrium distribution of energy among themselves will be much shorter than the time required for equilibrium to be reached in the exchange of energy between electrons and lattice. Hence the electrons may be considered as a system essentially in equilibrium all the time, and the exchange of energy between electrons and lattice is (relatively) so slow that this equilibrium among the electrons is not disturbed. The electrons therefore have an equilibrium temperature T and energy will only be transferred from electrons to lattice if T is greater than T_0, the lattice temperature. As the field applied to a crystal is increased the rate of energy transfer from field to electrons, A, increases, and since eq. (5) must be satisfied, B, the rate at which energy is transferred from electrons to lattice, must increase also.

This is only possible if the difference in temperature $(T - T_0)$ increases. Thus, as the field increases the electronic temperature rises relative to that of the lattice. It will be shown later (Section IV) that for fields greater than a certain critical value the rate of transfer of energy from electrons to lattice cannot keep pace with that from field to electrons. Hence no stable state is possible and the electronic temperature continues to rise (accompanied by an increase in the number of conduction electrons) until breakdown occurs.

In case (b) it is assumed that collisions between electrons can be neglected since collisions between electrons and lattice vibrations occur much more frequently. Under this condition, in the absence of a field, an electron with energy greater than the most probable energy at that temperature, will, on the average, lose energy to the lattice vibrations and tend to reach the most probable energy state. This rate of loss $b(E)$ is, roughly speaking, inversely proportional to the time between two collisions. This quantity increases with electronic energy, approximately $\propto E^{\frac{1}{2}}$; hence $b(E)$ decreases with energy,

$$b(E) \propto \frac{1}{E^{\frac{1}{2}}} \tag{7}$$

This rate of loss is modified by the field which, on an average over several collisions, transfers energy to an electron at a rate $a(E)$, so that $[b(E) - a(E)]$ is the total rate of loss of energy. Here

$$a(E) = \frac{e^2\tau(E)F^2}{m^*} \tag{8}$$

[cf. eq. (5) which follows from (8) by adding the contribution of all electrons]. The time of relaxation $\tau(E)$ is proportional to the time between two collisions and hence increases with energy E. Thus, since $b(E)$ decreases but $a(E)$ increases with E it is found that a critical energy E_c exists where $a = b$. For an electron which originally had an energy below this critical value, $E < E_c$, the rate of loss $b(E)$ is larger than the rate of gain $a(E)$ so that it still tends to reach lower energies though at a lower rate $(b - a)$ than without a field (b). If, however, the energy of the electron was larger than the critical energy, $E > E_c$, then the electron would not, on the average, lose energy at all but would gain energy.

It follows from (8) that the quantity $a(E)$ increases not only with energy, but also with the field F. Since $b(E)$ is independent of F, it follows that the critical energy E_c where $a = b$, also depends on F; it decreases as F increases. The existence of such a critical energy beyond which electrons (on the average) accelerate to higher and higher energies, means that, if no other processes were involved, equilibrium could never

be reached in any field, because every electron has a finite chance, due to fluctuations, of reaching an energy above the critical value. This can be shown exactly by calculation, and is illustrated in Fig. 4c. However another factor enters when an electron reaches an energy I above which excitation or ionization of the lattice ions or atoms is possible. When this energy has been reached a new type of collision is introduced (producing ionization or excitation of the lattice) and our original assumption, that the electron collides with lattice vibrations only, no longer holds. Furthermore, the average distance an electron goes before making a collision leading to excitation or ionization of the lattice is much smaller than the average distance between collisions with lattice vibrations. (The former is of the order of a few lattice distances only while the latter is of the order of one hundred lattice distances.) Hence for electronic energies greater than I collisions with the thermal vibrations of the lattice are rare compared with those leading to ionization (excitation) and we should expect that practically all electrons which reach the energy I will carry out collisions of the latter type.

There is now a chance that equilibrium may be established and whether or not this happens depends on the properties of electrons with energies approximately equal to I. An exact calculation of the condition under which equilibrium becomes possible is difficult and has not yet been carried out, but an approximate condition can be determined by considering the average behavior of an electron. As mentioned previously, if ionization is not a factor, an electron with energy $E > E_c$ will, on the average, gain more and more energy but if $E < E_c$ it loses energy. However, practically all electrons reaching energies greater than I will carry out collisions producing ionization or excitation of the lattice and then these considerations regarding the average behavior of electrons with energy less or greater than E_c no longer hold. If our calculation were to lead to values of $E_c > I$, then any electrons with energy less than I may, by a fluctuation, reach the energy I, although the probability of this process will be quite small. If, however, the field is sufficiently strong to make $E_c < I$ then all electrons which happen to have energy $E > E_c$ will gradually reach the energy I whereas electrons with energy $E < E_c$ will, on the average, lose energy. An electron, under these conditions, need only reach the energy $E_c < I$, because of a fluctuation, to have a good chance of carrying out an ionization (excitation) process. Since the probability of an electron reaching a certain energy, by a fluctuation, increases very greatly as this energy is decreased (i.e., approaches the most probable value), we should expect a large increase in the number of ionization processes in fields sufficiently strong to make $E_c < I$. We shall therefore assume that breakdown occurs for fields in

the neighborhood of that which makes $E_c = I$. As stated above, this condition cannot be considered as exact. However, it may be expected to yield the correct magnitude of the breakdown voltage—and general characteristics of the phenomenon—when collisions between electrons can be neglected.

This argument assumes that only a fraction of the electrons reaches the unstable "runaway" condition when the applied field is such as to produce breakdown. Seeger and Teller[9] have used the condition that all electrons should reach the unstable condition before breakdown occurs. This certainly gives an upper limit for the breakdown field, but in view of the above discussion it seems that breakdown should occur below this value.

IV. Results of the Theory

In this section we shall outline the development of the theory of breakdown for the two cases whose breakdown conditions have been discussed above.

1. High Temperature Theory[7]

In the first case (case (a) of the previous section) because of the high density of electrons in the conduction band and excited isolated levels the electrons form a system which is in equilibrium at all times. This condition will obtain at sufficiently high temperatures for practically all substances, since the number of electrons in the conduction band, and in the excited levels just below it, will increase with temperature. Conversely, at very low temperatures the number of conduction electrons (which are responsible for the exchange of energy among all the electrons) will be very small and the above equilibrium condition among the electrons cannot hold. Thus for any substance there will exist a certain critical temperature T_c at which the type of breakdown will change. Above this temperature the theory of case (a) will apply, while below it the theory of case (b) of the previous section will be more correct. The more impure or "amorphous" a substance is, the lower will its critical temperature be.

In case (a), a quantitative calculation of the breakdown field would require detailed knowledge of the behavior of slow electrons which is not available, but a useful qualitative picture of the way in which breakdown occurs can be given, and, on the basis of a special model, the way in which the breakdown field varies with temperature can be calculated. The important factor in such developments is the density of electrons in conduction and isolated levels; and this in turn depends on the temperature and on the number of possible trapping sites.

The simplified model used for an impure crystal or an amorphous solid is that shown in Fig. 3. It is assumed that the excited isolated levels exist in a closely packed group of width ΔV just below the conduction band. Electrons can make transitions from one level to another within this group and transfer energy to the lattice vibrations—although they remain trapped during these processes. The isolated ground level is assumed to be considerably below the excited levels (the energy gap being $2V$ where $V \gg \Delta V$).

The number of electrons in the conduction levels, N_1, is given by (cf. Section II, 1)

$$N_1 = C_1(T)e^{-V/kT} \tag{9}$$

and the number in the excited isolated levels, N_2, by

$$N_2 = C_2(T)e^{-V/kT+\Delta V/kT} \tag{10}$$

where $C_1(T)$ and $C_2(T)$ vary slowly with temperature and can be considered as constant. Hence

$$\frac{N_1}{N_2} = \frac{C_1(T)}{C_2(T)}e^{-\Delta V/kT} \tag{11}$$

and is constant at a given temperature. Also we assume that $\Delta V \gg kT$ so that it will be seen from eq. 11 that, for reasonable values of C_1 and C_2, $N_2 \gg N_1$.

Now, substituting for N_1 in eq. (5) we obtain for the rate of energy transfer per unit volume from field to electrons:

$$A = \frac{e^2\tau_1 C_1}{m^*}F^2e^{-V/kT} \tag{12}$$

The rate of energy transfer per unit volume from electrons to lattice vibrations is given by

$$B = \frac{h\nu}{\tau_2}N_2\left(\frac{e^{h\nu/kT_0-h\nu/kT}-1}{e^{h\nu/kT_0}-1}\right) \tag{13}$$

in which $h\nu$ is a quantum of energy (the smallest amount of energy which can be transferred from electrons to lattice or vice versa), τ_2 is an average time (a relaxation time) during which a quantum from one electron is transferred, and the quantity in brackets is a temperature factor which vanishes when the electron temperature T is equal to the temperature T_0 of the lattice—as we should expect since then the number of quanta of energy emitted by the electrons to the lattice is equal to the number absorbed from the lattice, and under these conditions obviously $B = 0$. This temperature factor has the form shown by the full line in Fig. 5.

It increases with T and reaches a constant value when $T \gg T_0$. As mentioned in Section III the equilibrium condition is $A = B$ and the electronic temperature T can be calculated from this condition. Thus if we equate the values of A and B in eq. (12) and (13) and divide through by all terms of B not depending on T we obtain

$$DF^2 e^{-\Delta V/kT} = e^{h\nu/kT_0 - h\nu/kT} - 1 \tag{14}$$

in which D depends slightly on T but can be treated as a constant.

From this equation we see that, for a given applied field, the electronic equilibrium temperature T can be calculated in terms of the lattice temperature T_0. We can solve for T by plotting the two sides of the equation separately on a temperature base and finding the point at which

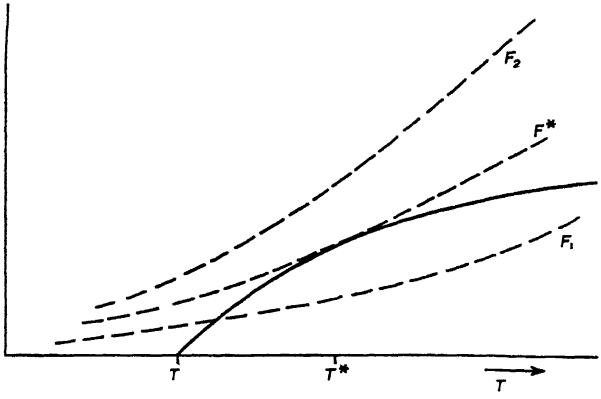


FIG. 5.—Solution of eq. (14) (c.f. reference 7). ——— right hand side of (14) proportional to energy loss per electron to the lattice. — — — left hand side of (14) proportional to energy gain per electron from the field F for three different values F_1, F^*, F_2 of F. F^* is the breakdown field.

the curves intersect. The right hand side of this equation is the same as the temperature factor in (4.5) except for a multiplying constant. The left hand side is essentially A/N_2, except for a constant; that is, it contains the term N_1/N_2, the fraction of electrons in the conduction band, which as shown by eq. (11) increases with the electronic temperature T. Also since it is an energy loss it is proportional to F^2. Now F^2 is an independent parameter and hence in plotting the left hand side of eq. (14) in Fig. 5 we obtain a set of curves for different values of F, which are shown dotted.

We now consider how T, the equilibrium temperature of the electrons, varies as F is increased. For F very small the intersection of the two curves occurs for values of T only slightly above T_0, the lattice temperature. As F is increased the temperature T rises to higher values, as

might be expected. However as Fig. 5 shows, if the field is increased beyond a critical value F^* the two curves do not intersect, and for $F > F^*$ an equilibrium temperature of the electrons cannot be reached. Thus the transfer of electric energy to the electrons causes a rise in electronic temperature leading to an increase in the fraction of conduction electrons and hence to a further increase in the transfer of electric energy and so on. Above F^* the rate of this process for any electronic temperature T is larger than the rate of loss of electric energy to the lattice vibrations. Hence stable conditions cannot be realized. A calculation of this critical field gives

$$F^* = Ce^{\Delta V/2kT_0} \tag{15}$$

where C may be considered constant.

2. *Low Temperature Theory*[8]

In case (b) we are dealing with a material at temperatures below the critical temperature. The number of electrons in the conduction levels is now very small and we are no longer dealing with an electronic system in equilibrium but must concern ourselves with the behavior of single electrons. In this case we denote the rates of transfer of energy from field to an electron and from electron to lattice by a and b respectively. Then, as pointed out in Section III, the relation $a = b$ holds for a critical energy E_c of an individual electron. (We must deal with electronic energies in this case; the concept of temperature implies an averaging over a large number of electrons.) Now, as stated in Section III, breakdown will occur when the field is such that the critical energy of the electron E_c is less than I, the ionization or excitation energy of the lattice. This breakdown field can be calculated from the condition $a(I) = b(I)$ where $a(I)$ is the rate of energy transfer from field to an electron of energy I and $b(I)$ is its rate of energy transfer to the lattice. Hence we have

$$a = \frac{e^2}{m^*}\tau(I)F^{*2} = b \tag{16}$$

where the relaxation time τ is that for electrons of energy I.

Hence

$$F^* = \left(\frac{m^*}{e^2}\frac{b}{\tau}\right)^{\frac{1}{2}} \tag{17}$$

In this expression the exponential temperature-dependent factor of case (a) is now missing and the main temperature-dependent factor is τ. A factor similar to τ does actually occur in the calculation of case (a) but since its variation with temperature is very small compared to

that of the exponential factor, it is included in the constant C of eq. (15). The quantity τ becomes smaller as the temperature increases and therefore, in this case, F^* should increase with temperature. This temperature dependence of F^* will not in general be as pronounced as in case (a), however, because τ does not vary as rapidly with temperature as the exponential factors in eq. (15). τ also depends upon the number of foreign atoms and defects in the lattice. In fact, $1/\tau$ behaves in the same way as the resistance of metals, increasing with temperature and with the addition of impurities, and for the same reason, namely the scattering of electrons by thermal motion and by imperfections in the lattice.

Thus

$$\frac{1}{\tau} = \frac{1}{\tau_T} + \frac{1}{\tau_f} \tag{18}$$

where τ_T is the part of τ depending on temperature and τ_f the part depending upon foreign atoms. τ_f is inversely proportional to the concentration of foreign atoms for small concentrations; that is, $1/\tau$ and hence F^* increases with increasing concentration of impurities.

3. Discussion

We have now reached the stage where we can give a qualitative description of the phenomenon of dielectric breakdown over the whole temperature range. The following predictions can be made: 1. There are two temperature ranges, separated by a critical temperature T_c, in which the effects of temperature on dielectric breakdown are different. For temperatures greater than T_c the dielectric strength decreases as the temperature increases, according to eq. (15). For temperatures less than T_c there is an increase of dielectric strength with temperature, which may, however, be so small that the dielectric strength is practically constant over a considerable temperature range. We thus expect the dielectric strength to be a maximum in the neighborhood of T_c. 2. An admixture of foreign atoms lowers the critical temperature T_c. This is due to the fact that adding foreign atoms increases the number of trapping centers and thus increases the number of electrons in the conduction band and isolated levels at a given temperature. Thus the temperature at which collisions between electrons become important (T_c) is lowered. Experimentally this would show as a lowering of the temperature at which the breakdown strength is a maximum, as the number of foreign atoms is increased. 3. An admixture of foreign atoms increases the breakdown strength F^* in the low temperature region. In the high temperature region however the addition of foreign atoms may, because of the

decrease of the critical temperature described in 2, cause a decrease in dielectric strength.

There are some additional results of the theory which are important since they also may be tested experimentally. Thus, in the high temperature region it is possible to predict the variation of conductivity of the material with field strength in a certain range.[7] As seen from eqs. (4) and (9), the conductivity at temperature T is proportional to the number of electrons in the conduction levels and hence to $e^{-V/kT}$. For very weak fields the electronic temperature will approach the lattice temperature T_0 and the conductivity will approach

$$\sigma_0 = C_3 e^{-V/kT_0} \tag{19}$$

where C_3 is a constant which varies slowly with temperature. Hence

$$\sigma = C_3 e^{-V/kT} = \sigma_0 e^{-V/kT+V/kT_0} \tag{20}$$

and since T is a function of the field strength, it follows that the conductivity σ also depends upon the field F. It actually turns out that

$$\log \frac{\sigma}{\sigma_0} = \frac{1}{2.718} \frac{F^2}{F'^2} \frac{V}{\Delta V} \tag{21}$$

if F is a strong field but still well below the breakdown value. If the temperature dependence of the conductivity σ_0 is known V can be obtained from eq. (19), and if the dependence of the breakdown voltage on lattice temperature is known ΔV can be obtained from eq. (15). Consequently the ratio of conductivities given by (21) can be determined as a function of F with no arbitrary constants involved. Since this function may also be obtained experimentally, a direct check of one of the fundamental parts of the theory is possible.

The low temperature breakdown theory gives an important result which cannot be obtained in the high temperature case, namely, the actual absolute value of the breakdown field. Assuming that $m^* = m$, the mass of a free electron, this is expressed in terms of properties of the material which may be measured directly, and is given by

$$F = 1.64 \times 10^5 \left(\frac{d}{G}\right)^{\frac{1}{2}} (\epsilon - \epsilon_0) \frac{\lambda_0}{\lambda^{\frac{3}{2}}} \left(1 + \frac{1}{e^{h\nu/kT} - 1}\right)^{\frac{1}{2}} \text{ volt/cm.} \tag{22}$$

in which d is the density of the material in g./cc., G is the molecular weight, ϵ is the static dielectric constant, $\sqrt{\epsilon_0}$ is the refractive index in the near infrared region, λ is the residual wavelength* in μ ($1\ \mu = 10^{-4}$

* The residual wavelength of a crystal is the wavelength at which pronounced reflection takes place. It is normally in the long wave infrared region and is connected with the natural vibration frequency of the atoms or ions of the crystal. In all but the simplest materials there will be more than one residual frequency.

cm.) and λ_0 is the wavelength of the first maximum of the ultraviolet absorption in Angström units (1 Å.u. = 10^{-8} cm.). This formula applies to the case of simple crystals of the NaCl type which may be considered to have one residual frequency only (see footnote). A more general formula can be derived but will not be given here.

It is thus possible to check the low temperature dielectric breakdown theory by comparing the experimentally determined breakdown field with that given by the above formula. This will be done in the following section.

The dependence of the dielectric strength of mixed crystals upon temperature and concentration of foreign atoms, in the low temperature range, may be calculated[10] using the general formula (16). This calculation has the advantage that the detailed properties of τ, other than that expressed by eq. (18) are not required, and in particular the assumption that $m^* = m$ need not be made. As stated previously

$$\frac{1}{\tau_f} \sim c \qquad \text{if } c \ll 1 \tag{23}$$

where c, the concentration, is the ratio of the number of foreign atoms to the total number of atoms (original plus foreign) in the crystal. Using this relation together with eqs. (16) and (18) we obtain the expression

$$F^2(c, T) - F^2(0, T) = cH \qquad \text{if } c \ll 1 \tag{24}$$

where F(c, T) is the breakdown strength at concentration c and temperature T, $F(0, T)$ is the breakdown strength of the original (pure) crystal at temperature T and H depends on neither c nor T. Since all quantities in this equation except H may be determined experimentally, it is possible to calculate H, and the extent to which this quantity remains constant over a range of concentrations and temperatures will determine to what extent the theory is valid.

Another interesting prediction of the low temperature part of the theory is that, for very thin layers, the breakdown field (in volts per centimeter for example) should increase as the thickness of the material is decreased. Thus for thicknesses less than several mean free paths (10^{-5} cm.) there should be an appreciable increase in dielectric strength.

V. Experimental Results—Comparison with Theory

There is now available a considerable number of experimental results dealing with various features of the intrinsic breakdown of dielectrics and in this section we shall compare these results with the theoretical predictions of Section IV.

1. General Features

We first consider the characteristics of intrinsic dielectric breakdown over the whole of the temperature range. It has been observed by Keller[11] that the general features predicted by the theory are well brought out by the measurements of von Hippel and Lee[12] on mixed NaCl-AgCl crystals. Figure 6 which is taken from the paper by von Hippel and Lee presents the data of these measurements. Attention is drawn to the following points:

1. There exists a critical temperature T_c below which the dielectric strength increases with increasing temperature and above which there is a decrease with increasing temperature. Thus a maximum of dielectric strength occurs in the neighborhood of T_c.

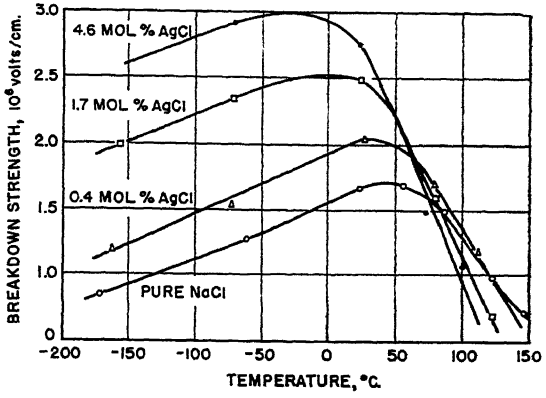


FIG. 6.—Dielectric strength of mixed crystals of NaCl and AgCl as a function of the temperature, measured by von Hippel and Lee.[12]

2. Adding AgCl to the NaCl crystal lowers the temperature of the maximum of dielectric strength, which corresponds roughly to the critical temperature T_c; and the larger this admixture of AgCl the lower does T_c become. This is in line with the second prediction of Section IV.

3. The dielectric strength of the NaCl crystal in the low temperature region is increased by admixture of AgCl, as expected because of the increased probability of scattering of the electrons as described in Section IV. Conversely, in the high temperature region, the dielectric strength of the mixed crystals decreases with increasing concentration of AgCl. This, again, is as predicted in Section IV, and is due to the increasing number of electrons in the conduction and isolated levels as the number of foreign (AgCl) atoms is increased.

Thus the qualitative conclusions regarding the dielectric strength over the whole temperature range and for increasing numbers of imperfections

in the lattice (produced by foreign atoms) have been confirmed by experiment. This must be regarded as important evidence regarding the assumptions on which the two sections of the theory are based.

2. *Existence of a Transition Temperature*

The existence of a transition between two types of breakdown in different temperature ranges is well brought out in a number of other cases. The results of von Hippel and Maurer[13] on soda-lime glass which are replotted in Fig. 7 indicate a pronounced transition at about 220°K. For temperatures below this the variation of dielectric strength with temperature is very small while at higher temperatures this variation is practically exponential as required by the theory. Data on the dielectric strength of mica (cf. Austen and Whitehead[14]) is plotted in Fig. 14. This material has a dielectric strength which changes slightly with temperature up to about 350°K. and then shows the sudden drop which is characteristic of the high temperature region.

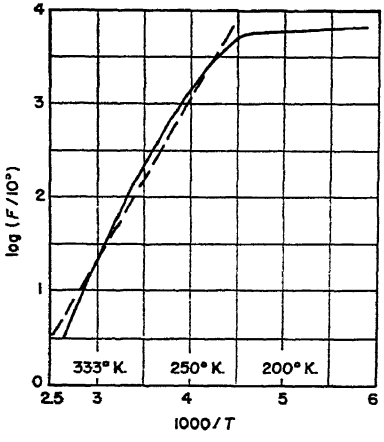


FIG. 7.—Temperature dependence of the dielectric strength of soda-lime glass. ———Experiments by von Hippel and Maurer.[13] — — —Theory according to eq. (15) with $\Delta V/k = 3400°$ (from reference 7).

The difference between the crystalline and amorphous forms of the same material (quartz) is well brought out by the results of von Hippel and Maurer in Fig. 8. Over the same temperature range breakdown characteristics showing the trends typical of the two sections of the theory are obtained. At low temperatures the amorphous form (fused quartz) has a higher dielectric strength than the crystalline form because of the increased scattering introduced by imperfections in the lattice (cf. also Austen and Whitehead[14]), but the transition to the high temperature range in which F^* has a large negative temperature characteristic starts at lower temperatures for fused than for crystalline quartz, so that at high temperatures the latter has the higher dielectric strength.

Results for long chain polymers also exhibit a marked transition temperature. The data of Austen and Pelzer[15] and Oakes on[16] polythene is plotted in Fig. 9 (taken from Oakes' paper). Polythene has the same composition and molecular structure as paraffin, shown in Fig. 10, but the chain length is very much greater. At low temperatures its mole-

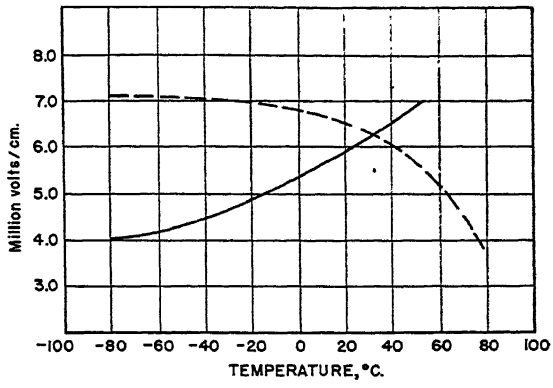


FIG. 8.—Temperature dependence of the breakdown field of crystalline (———) and of fused (— — —) quartz according to von Hippel and Maurer.[13]

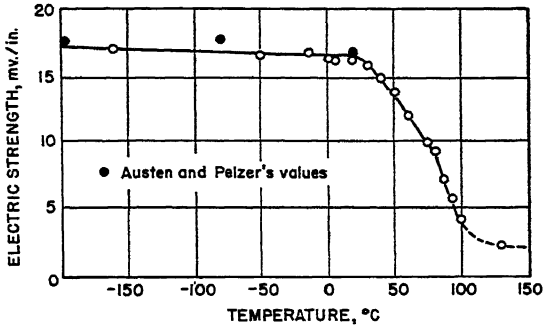


FIG. 9.—Temperature dependence of electric strength of polythene according to Austen and Pelzer,[15] and Oakes.[16]

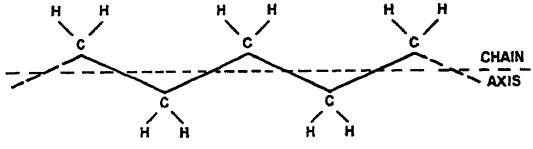


FIG. 10.—Schematic diagram of structure of paraffin or polythene. As indicated by the dotted lines at the ends, the chain length is, in general, greater than that shown in the diagram. It is very much greater in the case of polythene.

cules are highly ordered with respect to each other and it may be regarded as partly crystalline in nature.[17] At higher temperatures the degree of order decreases and the accompanying increase in the number of trapping centers produces the breakdown characteristic of a high temperature region. Oakes has also determined the effect of introducing "foreign" atoms by replacing some of the H atoms in Fig. 10 by Cl atoms—thus

obtaining chlorinated polythene.[18,19] These data are plotted in Fig. 11, which compares the dielectric strength characteristics of pure polythene and chlorinated polythene containing 8% Cl. It will be noted that "hot chlorination" produces a material whose breakdown characteristic is more of the amorphous type than the "cold chlorination" product (i.e., at low temperatures the dielectric strength of the former is higher than that of the latter material, but the transition to the decreasing high temperature characteristic occurs at a lower temperature). This is to be

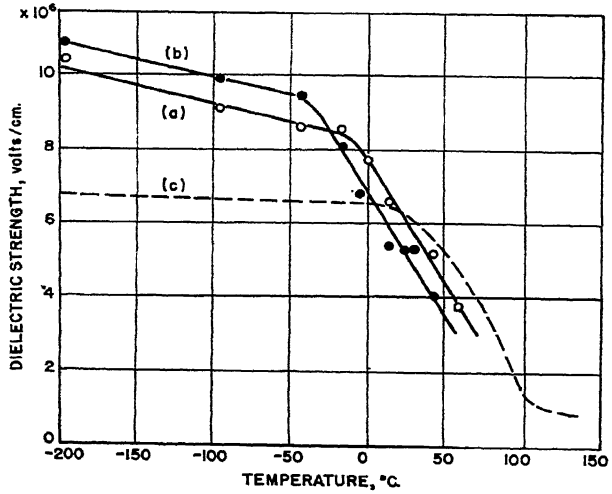


FIG. 11.—Temperature dependence of the dielectric strength of chlorinated polythene according to Oakes.[19]

(a) 8% Cl-cold chlorination
(b) 8% Cl-hot chlorination
(c) Polythene (from Fig. 9)

expected since the addition of chlorine at high temperatures where the long molecules can be forced into new positions more easily by the large chlorine atoms, should tend to produce a more disordered solid. Austen and Pelzer[15] have obtained similar data on copolymer polyvinyl chloride acetate in which alternate positions in one of the H rows in Fig. 10 are occupied by chlorine or acetate radicals, so that dipoles are introduced into the material in a more regular way than in chlorinated polythene. Their results which include similar tests on pure polythene are shown in Fig. 12. They apply to three temperatures only and show considerable scatter but it can be seen that addition of these different radicals makes the breakdown characteristic approach that for amorphous materials.

Another point which should be mentioned is that for these long chain chlorinated polymers there is a decrease in dielectric strength at temperatures below T_c rather than the increase predicted by the theory. In the case of the polymers there are definite structural changes near the transition point and the material softens.[20] The decrease in dielectric strength at lower temperatures may possibly be associated with the commencement of these changes of crystal structure. Data on physical changes in

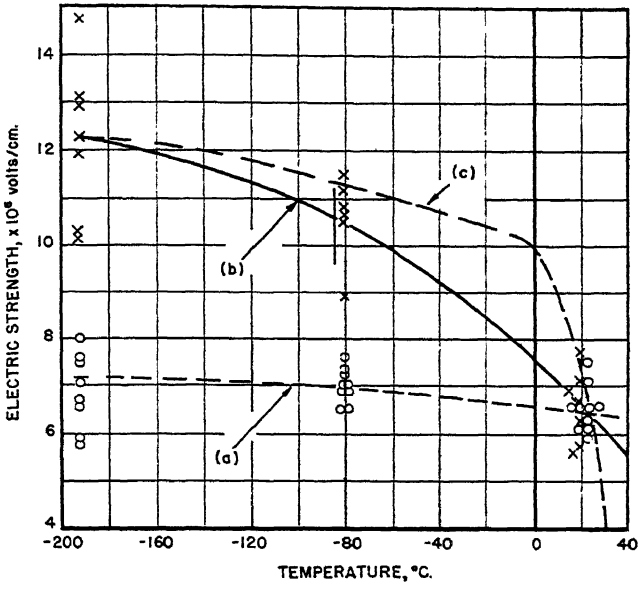


FIG. 12.—Variation of electric strength of polythene and polyvinyl chloride acetate with temperature.

(a) Polythene[15]
(b) Polyvinyl chloride acetate—curve of Austen and Pelzer[15]
(c) Polyvinyl chloride acetate—another curve fitting data equally well

the materials are not yet sufficiently complete to determine whether or not this is the case. Similar considerations apply to more completely amorphous substances in which changes of structure occur over considerable ranges of temperature.

Some recent work by Lindsay and Berberich[21] on ceramics has indicated that these materials also have a transition temperature at which the breakdown characteristic changes in a manner similar to that predicted here. However in view of the low absolute value of the dielectric strength it would be necessary to investigate the characteristics of these materials further, in particular (as suggested by Lindsay and Berberich)

to ascertain to what extent the breakdown is influenced by nonelectronic phenomena (motion of ions, etc).

3. *Low Temperature Region*

Experimental data on the low temperature region of breakdown are also available from various sources. Results of tests on potassium bromide obtained by Buehl and von Hippel,[3,13] Austen and Whitehead,[14] and Austen and Hackett,[22] are compared with the theoretical (absolute) value in Fig. 13. At room temperature this absolute value agrees

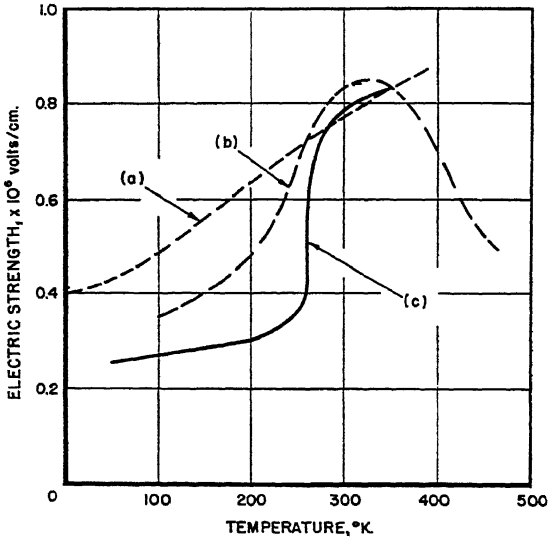


FIG. 13.—Temperature dependence of the electric strength of potassium bromide.

(a) Theory
(b) Measurements by Buehl and von Hippel[3,13]
(c) Measurements by Austen and Hackett[22]

numerically with that obtained experimentally. The variation of dielectric strength with temperature shows qualitative agreement with the theory which is as much as we can hope for considering the discrepancy between the two sets of experimental results. There is again a rise in dielectric strength with temperature, as expected, because of the increased scattering of the electrons. The Electrical Research Association's experimental results for mica[14,22] are plotted in Fig. 14. The absolute theoretical value for this material cannot be obtained due to a lack of the requisite optical data, but, assuming a value for the residual wavelength, the dependence of the dielectric strength on temperature can be calculated. The theoretical curves thus obtained are plotted,

for two different residual frequencies, in Fig. 14, using the experimental value at room temperature (300°K.) to fix the actual magnitudes of the breakdown values. The smaller slope of the low temperature curve for mica compared to KBr is accounted for theoretically by the fact that the residual ray wavelength (λ in eq. (22)) for KBr is much larger than that for mica.

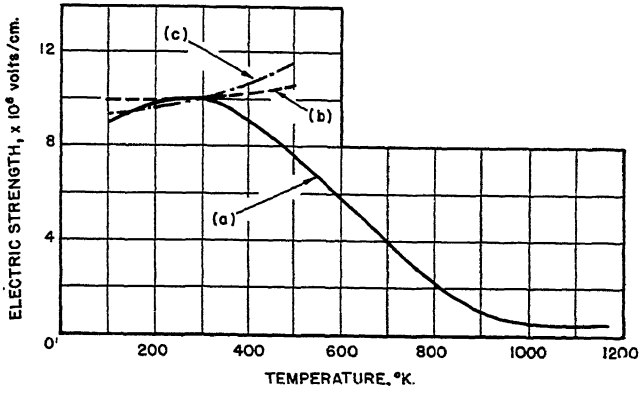


FIG. 14.—Temperature dependence of the electric strength of mica.[14]

(a) Approximate experimental curve drawn through data of Austen and Whitehead
(b) Theoretical curve for a residual ray wavelength of 10μ
(c) Theoretical curve for a residual ray wavelength of 20μ

There is considerable spread in the experimental data and by using maximum values only it is possible to draw curve (a) so that there is a much sharper transition at 400°C. approx. The theoretical curves are adjusted to fit the experimental data at $T = 300$°K. as explained in the text.

Data on the dielectric strength of mixed crystals, in addition to that discussed at the beginning of this section, are provided by the experiments of von Hippel[3,4] on mixtures of KCl-RbCl, and are shown in Fig. 15. This again shows that addition of foreign atoms raises the dielectric strength of a pure crystal.

A comparison of experimental (von Hippel) values of dielectric strength and theoretical absolute values calculated from equation (22), assuming $m^* = m$ is given in Table I. In this case there are no arbitrary

TABLE I. F_{100} in 10^5 v./cm. for $T = 300$°K.

	NaCl	NaBr	NaI	KCl	KBv	KI	RbCl	RbBr	RbI
Experiment.....	15	10	8	8	7	6	7	6	5
Theory.....	15	13	14	8.9	7.3	6.3	8.4	5.4	4.7

constants in the theoretical results and no adjustment could be made to give "average" agreement over the whole range of values.

Due to the fact that von Hippel's results are for electrical fields in the (100) direction in the crystal, while breakdown usually occurs in the (110) direction, the theoretical value obtained from eq. (21) is multiplied by $\sqrt{2}$. This procedure seems to be a reasonable one although it has not yet been fully justified. The assumption $m^* = m$ can be justified as far as order of magnitude is concerned but a more precise determination of m^* would require much more detailed theoretical work. The

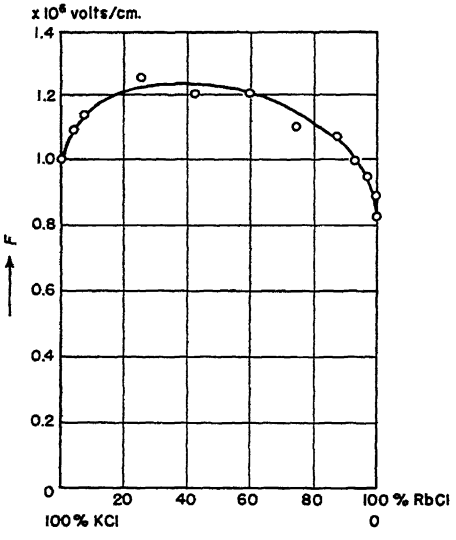


FIG. 15.—Electric strength of mixed crystals of potassium chloride (KCl) and rubidium chloride (RbCl) at constant temperature as determined by von Hippel.[4]

agreement between experimental and theoretical results is surprisingly good and indicates that the correct order of magnitude of the breakdown field strength is obtained from the theory. A more precise statement than this is hardly warranted until a more accurate determination of m^* is available.

The results of von Hippel and Lee[12] (Fig. 6) on mixed crystals of sodium chloride (NaCl) and small concentrations of silver chloride (AgCl) enable us to determine the quantity $(F^2(c, T) - F^2(0, T))/c$ of eq. (24) for various concentrations and temperatures. The results obtained[10] are given in Table II. In view of the fact that F^2 varies by a factor 10 and since, due to the shortage of experimental data, there is some error possible in locating the curves, we may consider that these results provide fair evidence for the correctness of eq. (24).

TABLE II. Variation of $(F^2(c, T) - F^2(0, T))/c$ with concentration and temperature.

T, °C.	$c = 0.004$	$c = 0.017$
−150	18	20
−100	25	21
− 50	28	24
0	30	23

4. *Associated Phenomena*

Data concerning the dependence of dielectric strength on thickness of the sample has been given by Austin and Hackett,[22] and Austin and

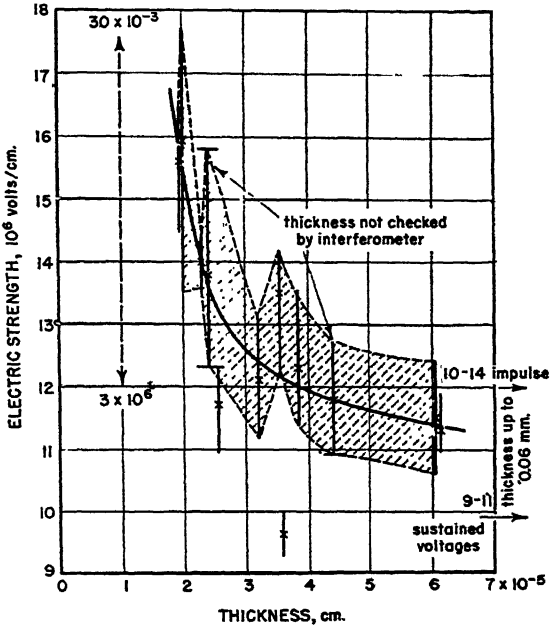


FIG. 16.—Electric strength of mica as a function of thickness (Austen and Whitehead[14]).

X indicates mean value and vertical lines indicate range of dispersion.

Whitehead,[14] for mica and by Plessner[23] for crystalline materials (CaF_2, NaF, KBr) and amorphous materials (silica and polystyrene). Results for the crystalline materials indicate that the dielectric strength increases when the thickness is decreased to values of the order of the mean free path. In the amorphous materials, for which the mean free path is much shorter, no dependence of dielectric strength on thickness was noticed

for the samples used. This was as expected since the minimum thickness of amorphous film used (limited by experimental technique) was considerably larger than the mean free path in such materials. Typical results for mica are shown in Fig. 16. In the work of Plessner an interesting phenomenon observed was the increase in dielectric strength of thin films of crystalline material when the field is applied in pulses of very short duration. This phenomenon is not explained by the theory in its present form and further work both experimental and theoretical would be required to account for it.

We finally mention some interesting work by Turner and Lewis[24] on the conductivity of soda-lime glass in strong fields. Their results satisfy the relation of conductivity to field strength expressed in eq. (21). However, confirmation that these currents are entirely electronic in nature should be obtained before we may consider that the theoretical relation has been completely verified.

In view of the agreement with experimental results indicated by the detailed discussion of this section, it seems safe to conclude that the basis of the theory whose development is outlined in the preceding sections is substantially correct, and that the theory accounts for the essential features of intrinsic dielectric breakdown phenomena in solids.

ACKNOWLEDGEMENTS

Thanks are expressed to the Physical Society, the Royal Society, the American Physical Society and the Institution of Electrical Engineers for permission to use figures contained in their publications.

REFERENCES

1. Whitehead, S. Dielectric Breakdown. Clarendon Press, Oxford, in press.
2. Fröhlich, H. *Rep. Phys. Soc. Progr. Phys.*, **VI**, 411–30 (1939).
3. Buehl, R. C., and von Hippel, A. *Phys. Rev.*, **56**, 941–47 (1939).
4. von Hippel, A. *Z. Phys.*, **88**, 358–65 (1934).
5. Bloch, F. *Z. Phys.*, **52**, 555–600 (1928).
6. Mott, N. F., and Gurney, R. W. Electronic Processes in Ionic Crystals. Clarendon Press, Oxford, 1940.
7. Fröhlich, H. *Proc. Roy. Soc.*, **A188**, 521–32 (1947), based on *Rep. Brit. Elect. Allied Indust. Res. Ass.*, **L/T153**.
8. Fröhlich, H. *Proc. Roy. Soc.*, **A160**, 230–41 (1937); *Proc. Roy. Soc.*, **A172**, 94–106 (1939), based on *Rep. Brit. Elect. Allied Industr. Res. Ass.*, **L/T105**; *Proc. Roy. Soc.*, **A178**, 493–98 (1941), based on *Rep. Brit. Elect. Allied Industr. Res. Ass.*, **L/T113**; *Proc. Roy. Soc.*, **A188**, 532–41 (1947).
9. Seeger, R. J., and Teller, E. *Phys. Rev.*, **54**, 515–19 (38).
10. Fröhlich, H. *Phys. Rev.*, **61**, 200–201 (1942).
11. Keller, K. J. *Physica*, **14**, 15 (1948).
12. von Hippel, A., and Lee, G. M. *Phys. Rev.*, **59**, 824–26 (1941).
13. von Hippel, A., and Maurer, R. J. *Phys. Rev.*, **59**, 820–23 (1941).

14. Austen, A. E. W., and Whitehead, S. *Proc. Roy. Soc.*, **A176**, 33–50 (1940), based on *Rep. Brit. Elect. Allied Industr. Res. Ass.*, L/T114.
15. Austen, A. E. W., and Pelzer, H. *J. Instn. Elect. Engrs.*, **93** (Part I), 525–32 (1946).
16. Oakes, W. G. *J. Instn. Elect. Engrs.*, **95** (Part I), 36–44 (1948).
17. Bunn, C. W., and Alcock, T. C. *Trans. Faraday Soc.*, **41**, 317–25 (1945).
18. Fröhlich, H. *Rep. Brit. Elect. Allied Industr. Res. Ass.*, L/T169.
19. Oakes, W. G. *J. Instn. Elect. Engrs.*, **96**, 37–43 (1949).
20. Richards, R. B. *Trans. Faraday Soc.*, **41**, 127–37 (1945).
21. Lindsay, E. W., and Berberich, L. J. Electrical Properties of Ceramics as influenced by Temperature. A.I.E.E. Technical Paper 48-175, scheduled for publication in *Trans. Amer. Inst. Elect. Engrs.*, **67** (1948).
22. Austen, A. E. W., and Hackett, W. *Nature, Lond.*, **143**, 637–39 (1939).
23. Plessner, K. W. *Proc. Phys. Soc., Lond.*, **60**, 243–56 (1948).
24. Turner, C. H. M., and Lewis, W. E. *Nature, Lond.*, **159**, 334–36 (1947).

The Microwave Magnetron

GUNNAR HOK

University of Michigan, Ann Arbor, Michigan

CONTENTS

I. Introduction

Although the name "Magnetron" was coined as far back as 1921, the high-power magnetron oscillator is entirely a product of World War II, and is one of its most remarkable developments. Characteristic of the status of the magnetron at the outbreak of the war is the fact that the Germans ruled it out as a prospective source of high microwave power and concentrated their efforts on the improvement of other oscillators. The magnetron proved to be particularly well suited for the pulse technique used in radar and contributed very much to the ultimate superiority of the Allies in radar search and fire-control equipment.

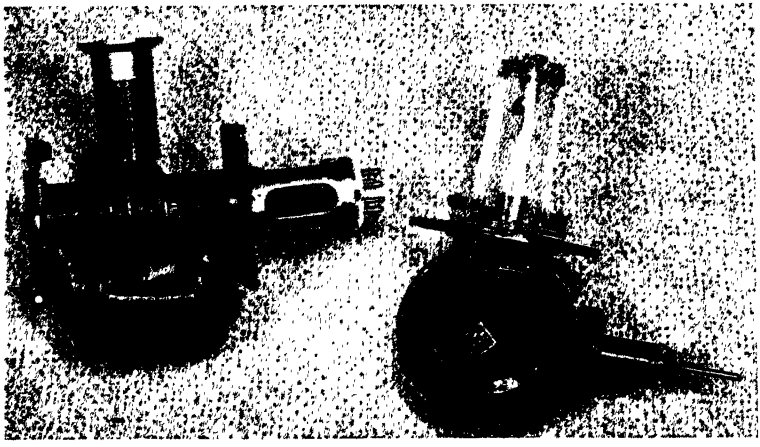


FIG. 1.—A QK-59 magnetron (left; 65 watts continuous operation, 3000 Mc/sec.) and a 2J34 magnetron (right; 300 kw pulse operation, 3000 Mc/sec.).

The magnetron oscillator was in the beginning developed on an experimental, more or less intuitive, basis. Through the later work of Hartree, Slater, Brillouin, and others a qualitative understanding of its principles of operation was obtained and a firm foundation laid for the expansion of its useful range in frequency and power according to the rules of similitude.[1]

In this article the author will try to sketch the magnetron art of today, including the present understanding of the mechanism of oscillation in a magnetron, the properties of a magnetron oscillator as compared to other microwave oscillators, and a review of recent contributions to the further development of the magnetron towards meeting the ever more exacting demands of modern navigation, ordnance, communication, industry, and basic research in physics.

1. *The Magnetron as a Nonlinear Oscillator*

The nonlinear theory of oscillation as developed by Van der Pol[2] and others for vacuum tube oscillators sheds some light on a number of properties and phenomena that the magnetron has in common with other oscillators, even though the approximate form of the nonlinear differential equations of the oscillation can not be derived as easily as for triode and pentode oscillators, since the magnetron as we know it today has no counterpart to the static characteristic of those tubes. It is easy to establish, however, that the oscillator can be considered as a combination of a linear multiresonant network and a nonlinear circuit element, i.e., a circuit element whose impedance is a function of the amplitude of the oscillation. Characteristic of such an oscillator are:

1. The existence of steady-state amplitudes of oscillation of different frequencies; by definition, of course, at least one of them should be larger than zero under normal operating conditions.

2. Interaction, due to the nonlinearity, between all oscillations in the system, even though they may in the linear resonator be orthogonal in time, space, or both.

3. Appearance of a spectrum of combination frequencies.

In the oscillators under discussion the nonlinearity is small, so that all oscillations are approximately sinusoidal.

The fact that the DC voltage sustains the oscillations is one of the symptoms of the nonlinearity and nonorthogonality in the system. As far as the principal oscillation is concerned this effect of the nonlinear properties of the oscillator can be represented in an AC equivalent network by an admittance with a negative conductance component. The value of this admittance will be a function of the DC voltage.

Any steady-state oscillation requires that the susceptance as well as the conductance of the whole network be zero. A change of the DC voltage will, in general, cause a change of both susceptance and conductance of the nonlinear element, so that a new steady-state frequency and amplitude automatically result. This frequency change, usually referred to as "frequency pushing" in many oscillators, including magnetrons, adds an undesirable frequency modulation when the oscillator is amplitude-modulated by means of varying the DC voltage.

Similarly, a variation in load impedance changes the conditions of equilibrium so that a corresponding variation in frequency and amplitude of oscillation is produced. The conventional name for this kind of frequency variation is "pulling."

The coupling that the nonlinearity introduces between oscillations in the various natural modes of the linear resonator leads to the phe-

nomena called *mode competition.* Initially, thermal fluctuations etc. excite small oscillations in all the natural modes. Oscillations have a tendency to build up at all of those frequencies for which the total conductance is negative. It is possible for several oscillations to build up to a steady state simultaneously, particularly if their frequencies are very different, but in general one oscillation prevails, suppressing all the others. To the author's knowledge this phenomenon has not been thoroughly investigated by means of the general nonlinear theory of oscillation, although Van der Pol's treatment of the coupled-circuit triode oscillator points the way.

The interaction between different oscillations in the magnetron is also of importance with regard to noise. If the total conductance for one mode of oscillation is very close to zero, noise may produce appreciable transients although no continuous oscillations can be maintained at that frequency. The noise may modulate a steady-state oscillation of another frequency. Even if no such regenerative amplification of the noise takes place, the effective noise band-width of the resonator is likely to be increased by the nonlinear coupling between the many natural modes of the resonator.

2. *Main Components of a Magnetron*

All magnetrons have the following main features in common:

1. A *cathode* supplying the necessary electrons by thermionic emission, by secondary emission, or by a combination of both; it is usually cylindrical in shape but may occasionally be a straight or helical filament.

2. A cylindrical *anode,* coaxial with the cathode, forming the outside boundary of the *interaction space* between anode and cathode. The anode is divided into segments by axially directed slots.

3. A *resonant system* or *tank circuit,* of which the anode or anode segments form part and serve as input terminals; it determines the frequency of oscillation and is provided with means of transferring power to the output terminals.

4. A *magnet* to maintain a constant and substantially uniform axial magnetic field in the interaction space.

All these parts with the exception of the magnet are usually inside the evacuated tube. The tank circuit and the electrode system are much more intimately related than in ordinary vacuum-tube oscillators; we can hardly describe the electron interaction without anticipating some facts about the resonant system. Neither can we analyze the operation of the resonant system without knowing the requirements imposed on it by the space-charge conversion process in the interaction space.

In the following we shall first discuss the interaction of the electrons with the electromagnetic field and the process of transferring power from the DC electric field to the radio-frequency field. A later section will analyze different types of tank circuits used in magnetrons and the part they play in the mechanism of oscillation.

II. The Magnetron Space-Charge Mechanism

The trajectories of the electrons in an oscillating magnetron are extremely complicated and intersect in such an intricate manner that no coherent flow pattern is established. For this reason all attempts to derive from the electromagnetic field equations and the laws of motion the relations between the amplitude of oscillation on the one hand and the DC voltage, the magnetic field, and the geometry on the other hand have been fruitless. Nonetheless, it can be said that the design of magnetrons now is established on a reasonably firm theoretical foundation. From a careful study of the system of differential equations for the motion of the electrons it has been possible to deduce a number of *scaling rules* by means of which a magnetron that has proved to operate with good efficiency can be reproduced at different voltage, frequency, or power with otherwise essentially the same operating properties. There has, furthermore, been established from theory a lower limit for the operating voltage, the *threshold voltage*, which makes it possible to calculate approximately the operating voltage of a magnetron for any value of the magnetic field. The value of this guidance from theory has been well confirmed by frequent use over a number of years.

We shall not attempt to give a complete picture of the magnetron theory. After a recapitulation of the motion of an electron in a combined magnetic and electric field we shall give a qualitative description of the space-charge mechanism in a magnetron under oscillatory as well as nonoscillatory conditions.

1. Electron Motion in Crossed Fields

An electron of a certain velocity $\mathbf{v}$ entering a uniform magnetic field of flux density $\mathfrak{B}$ experiences a force

$$\mathfrak{F} = e\mathfrak{B} \times \mathbf{v} \tag{1}$$

where e is the charge of the electron. All quantities in this equation and throughout this paper are given in rationalized mks units. If the velocity vector is in a plane perpendicular to the magnetic field, the electron will assume a circular path of radius r such that

$$\frac{mv^2}{r} = \mathfrak{F} = \mathfrak{B}e \times \mathbf{v} \tag{2}$$

$$\omega_c = \frac{v}{r} = \frac{e}{m}\mathfrak{B} \tag{3}$$

$$r = \frac{v}{\omega_c} \tag{4}$$

where m is the mass of the electron and $\omega_c/2\pi$ is the *cyclotron frequency*, the number of revolutions per second in the circular orbit.

It is significant that the angular velocity ω_c of the circular motion is determined by $\mathfrak{B}$ alone and independent of the initial velocity $\mathbf{v}$.

In a uniform electric field $\mathfrak{E}$ perpendicular to the magnetic field an electron can move in a straight line perpendicular to both fields, if its velocity is

$$v_c = \mathfrak{E}/\mathfrak{B} \tag{5}$$

The motion of electrons that enter such a region of "crossed fields" with a velocity different from v_c can be interpreted as a circular motion with the angular velocity ω_c superposed on a rectilinear drift with velocity v_c. The radius of the circular motion is such that the initial energy is conserved. The electron orbits are trochoids. If the electron starts at rest from a plane equipotential cathode it will describe a cycloid, periodically returning to the cathode and reaching a maximum distance from the cathode of

$$d = 2r = 2v_c/\omega_c = 2\frac{m}{e}\mathfrak{E}/\mathfrak{B}^2. \tag{6}$$

In a cylindrical interaction space with axial magnetic field and radial electric field analogous orbits will be obtained. The drift will no longer be rectilinear but circular, concentric with the cathode.

2. Nonoscillatory Space Charge. Hull's Critical Voltage

If the cathode is a thermionic emitter it will be immersed in a cloud of electrons, usually referred to as "space charge." As long as the potential difference between anode and cathode is constant and smaller than a certain critical voltage no electrons will reach the anode because of the deflecting force of the magnetic field. The presence of the space charge will radically change the potential distribution and the orbits of the electrons. The law of conservation of energy, however, makes it possible to derive for a coaxial cylindrical diode a relation between the potential E_s at the edge of the electron cloud where the radial velocity of the electrons is zero, the radius r_s of the edge of the cloud, and the flux density; a relation that is valid whatever the potential distribution inside the cloud and the orbits of the electrons may be. The important assump-

tions are that the kinetic energy of the electrons at any point is such as if they had left the cathode with zero velocity and that the potential and the magnetic field do not vary with time. The result, first obtained by A. W. Hull,[3] is the following formula:

$$E_a = \frac{e\mathfrak{B}^2 r_a^2}{8m}\left[1 - \left(\frac{r_c}{r_a}\right)^2\right]^2 \tag{7}$$

where r_c is the radius of the cathode.

When r_a is equal to the anode radius r_a E_a becomes the critical voltage E_c, i.e., the potential difference between anode and cathode for which current just begins to flow.

If all the electrons in the cloud move in concentric circles, i.e., the radial velocity is everywhere zero, the same relation holds between potential E and radial coordinate r at any point inside the cloud as at the edge of the cloud, and eq. (6) applies throughout the space charge, if E_a and r_a are replaced by E and r, respectively.

The question of the actual electron orbits in the space charge of a nonoscillating magnetron has been controversial. The concentric circular orbits give the *single-stream* or *Brillouin* solution of the space-charge equations, while paths of cycloidal character give the *double stream* solution, so called, of course, because simultaneously there are one stream of outgoing and one stream of returning electrons. The stability of the single-stream solution has been questioned; on the other hand, strong arguments have been presented that no double-stream solution can exist in any space-charge cloud of smaller radius than about 2.3 times the cathode radius.

The theory predicts that as the anode voltage is raised past the critical value, also called the "magnetic cut-off voltage," the current jumps from zero to a finite value I_c. This is most easily explained from the double-stream solution; when the critical voltage is reached, all the orbits touch the anode and the electrons are captured. To maintain the same voltage distribution, the number of electrons leaving the cathode doubles so as to make up for the disappearance of the previous stream of electrons returning to the cathode. Like the critical voltage this current I_c is one of the important parameters in magnetron operation.

Actually the rise of the current in the vicinity of the critical voltage is somewhat gradual, more so than can very well be explained by the initial velocity distribution of the electrons emitted from the cathode. The reasons for this are not completely understood.

3. Electron Resonances

It was explained above how an electron in crossed fields between cylindrical electrodes may have one of two different types of circular or

quasi-circular motion, or a combination of both. One motion takes place with the angular velocity ω_e (eq. 3), equal to 2π times the cyclotron frequency, and is in general not concentric with the electrodes. The other motion was characterized as a drift of the electron, or the center of the first-mentioned motion if present, perpendicularly to the electric field and consequently concentrically with the electrodes. Now if an electric field of radio frequency is present between the electrodes and at least at some point tangential to the orbit of an electron an exchange of energy may take place between this field and the electron because of the accelerating or retarding action of the field. This exchange will be of significant magnitude if the RF field has the same phase every time the electron passes this point, i.e., if a condition of resonance or synchronism exists between the electron motion and the RF field. The frequency of the RF field can be adjusted to resonate with either of the two types of electron motion, and these two types of electron resonances are utilized in two different categories of magnetron oscillators.

4. Cyclotron-Frequency Oscillators

The earliest magnetron oscillators operated at the cyclotron frequency. Their efficiency and stability were poor and this type of operation has been abandoned. There is no reason to discuss them here.

5. Traveling-Wave Magnetrons

In the traveling-wave magnetrons the drift of the electrons about the cathode is utilized. The anode voltage for normal operation is far below the critical voltage (Section II, 2), so that in the absence of any RF field no electrons would reach the anode. The balance between the outward force on the electron due to the electric field and the inward force due to the magnetic field would be such as to maintain the constant curvature of the circular path. An RF voltage between the anode segments produces fields that are equivalent to a number of electromagnetic waves traveling circumferentially along the anode surface. The tangential component of the electric field of one of these waves will cause the electrons at the edge of the space charge to slow down at some points and to speed up at others. This upsets the balance of forces that exists in the absence of RF fields, so that the retarded electrons will drift outwards beyond the Hull radius (r_s in eq. 6) towards the anode, and the accelerated electrons will drift inward towards the cathode, because of the change in the magnetic force acting on each electron. If the angular velocity of the electrons in their drift about the cathode is equal to that of the wave, i.e., if the electrons are synchronous with the wave, a continuous exchange of energy between the electric fields and the retarded

electrons may take place. As the electrons drift outwards they gain energy from the DC field, but since they simultaneously give up energy to the retarding RF field no new circular equilibrium orbit can be established. This process thus results in a continuous transformation of DC power to RF power. Some RF power is lost because some electrons are accelerated tangentially by the RF field and therefore return to the cathode with a surplus of energy, but this loss is relatively small, particularly if the electrons reach synchronous velocity fairly close to the cathode, where the RF field is weak. The tangential acceleration and retardation produces a bunching of the electrons, so that the electron density in the regions where acceleration takes place is very small if not zero. The electrons that are once captured by the revolving field remain in synchronism for the rest of their transit to the anode.

This is the principle of operation of the magnetron. Its space charge can be pictured as forming the hub and spokes of a wheel revolving about the cathode and just touching the anode. The power that keeps the wheel spinning is derived from the DC voltage, and the spokes induce RF currents in the tank circuit as they sweep past the anode segments. The RF voltage between the segments makes the oscillation self-sustaining by maintaining the shape of the space-charge wheel.

6. Self-Consistent-Field Computations

The exact shape of the wheel is not very well known. A general solution of the field equations and equations of motion is prevented by the fact that the electron trajectories intersect, so that the electron velocity is not a single-valued function of the space coordinates. By laborious numerical calculations according to the self-consistent-field method of successive approximations, electron trajectories and potential distributions have been found for a few operating conditions.

Figure 2 is a free-hand sketch of the result of a self-consistent-field calculation. The frame of reference is revolving with the angular velocity of the sinusoidal RF field along the anode, so that all field quantities appear stationary. Since the magnetic force on the electrons is proportional to their velocity the laws of electron motion will not hold in the new reference system. This discrepancy can be removed, however, by the appropriate reduction of the radial electric field at each point to an equivalent electric field in the new system. The solid lines indicate the equivalent equipotential surfaces and the dotted lines some representative electron trajectories. The shaded regions are inaccessible to the electrons, because the equivalent potential is negative. The arrowheads on the equipotentials indicate the direction of the drift of the electrons.

The sketch shows how a "spoke" forms where these lines point from the region of the cathode towards the anode.

The results of these calculations have given valuable information about the approximate trajectories of the electrons and confirm the idea of the wheel-shaped space charge. The method is too laborious, however, to supply any complete picture of the magnetron operation, of the build-up process of the oscillation, etc.

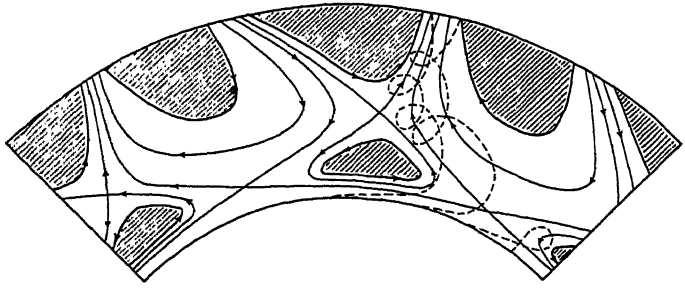


FIG. 2.—Sketch of equipotentials and electron trajectories in an oscillating magnetron. The shaded regions have negative equivalent potential and are inaccessible to the electrons.

7. *Threshold Voltage*

The *threshold voltage* or *Hartree voltage* E_t, below which no oscillations can take place, has been derived from theory and checks within 20% with experimental observations. It is

$$E_t = \frac{2\pi \cdot 10^8}{n\lambda_0}(r_a^2 - r_c^2)\left\{\mathfrak{B} - \frac{0.106}{n\lambda_0\left[1 - \left(\frac{r_c}{r_a}\right)^2\right]}\right\}. \qquad (8)$$

Here n is the "mode number," i.e., the number of complete RF voltage cycles along the inside of the anode circumference; λ_0 is the resonance wavelength of the tank circuit.

In a diagram showing the anode voltage vs. magnetic field the line representing the threshold voltage is a tangent to the Hull cut-off parabola (Fig. 3). Different lines are obtained for different values of the mode number n. The region in which oscillations can be obtained is the wedge-shaped area above the point of tangency between the cut-off parabola and the Hartree line. Curves of constant anode current run parallel to and above the Hartree line, the higher up the larger the current. The efficiency of the electronic process of converting DC power to RF power increases from left to right along these curves. An estimate of this electronic efficiency η_e can be made by a comparison of the energy

of an electron that hits the anode with synchronous velocity and the energy of an electron that traverses the interelectrode space without losing any energy to the RF field. The former energy represents the loss, the latter the total input. Consequently

$$\eta_e \cong 1 - \frac{mc^2}{2eE_b}\left(\frac{2\pi r_a}{n\lambda_0}\right)^2 \qquad (9)$$

where E_b is the anode voltage.

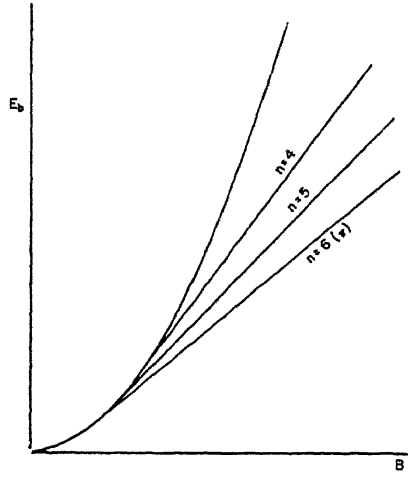


FIG. 3.—Hartree diagram.

8. *Scaling*

Guidance in design of magnetrons is given by a number of *scaling rules*, which make it possible to reproduce a successful design at different voltage, wavelength, or power level, so that the new tube has the same efficiency and similar characteristic curves as the model tube. The hypothesis is that if the variables involved in the operation are expressed as certain dimensionless ratios E/E_0, $\mathfrak{B}/\mathfrak{B}_0$, I/I_0 the characteristics of all magnetrons will be alike. Even if this is only approximately true, this idea has removed most of the guesswork and been immensely helpful in the development of magnetrons for new applications.

The constants E_0, $\mathfrak{B}_0$ and I_0 were chosen by Slater so that they were relatively easy to derive but still could be expected to be significant for the operation of an oscillating magnetron. E_0 is the DC anode voltage that gives an electron a velocity synchronous with the RF field at anode radius in the absence of any RF voltage. $\mathfrak{B}_0$ is the magnetic field calculated from the Hull formula (eq. 6) when the plate voltage is E_0 (i.e., $E_s = E_0$, $r_s = r_a$). I_0 is the current that just begins to flow when the plate voltage is E_0, the magnetic field $\mathfrak{B}_0$, and the RF voltage zero.

The constants are

$$E_0 = 253{,}000\left(\frac{2 - \cdot_a}{\lambda}\right)^2 \qquad (10)$$

$$\mathfrak{B}_0 = \frac{0.021}{n\lambda_c[1 - (r_c/r_a)^2]} \qquad (11)$$

$$I_0 = \frac{8440a_1}{[1 - (r_c/r_a)^2]^2\left\{\frac{r_a}{r_c} + 1\right\}}\left(\frac{2\pi r_a}{n\lambda_0}\right)\cdot\frac{h}{r_a} \qquad (12)$$

The quantity a is a function of r_c/r_a, which in all practical cases differs very little from unity. h is the axial length of the electrodes. Other symbols have the same meaning as before.

Suppose, for instance, that a certain 10-cm. magnetron works with good efficiency at 10 kv. anode voltage and 50 amps. anode current. It is desired to design a 3-cm. magnetron with the same efficiency. This can be accomplished by reducing all the geometric dimensions of the 10-cm. design in the proportion 10 to 3. The constants E_0 and I_0 in eqs. (10–12) remain unchanged while $\mathfrak{B}_0$ is increased in the proportion 3 to 10. The new magnetron will operate at 3 cm. at a higher magnetic field but at the same voltage, current, power, and efficiency as the original tube at 10 cm. This procedure is not the only one possible, however, since another independent variable, the mode number n, is available.

The process described illustrates *wavelength scaling;* in *voltage scaling* the original design is reproduced at a different anode voltage by changing the anode radius r_a, the mode number n, or both, in such a way that E_0 is changed in the desired proportion. In order to maintain constant power output it is also necessary to change the ratio h/r_a in such a way that I_0 changes in the inverse proportion to E_0. For reproduction of the original design at a different power level but at the same anode voltage and wavelength an analogous procedure of *current scaling* is followed.

The reason that the scaling rules and the threshold voltage do not agree exactly with experiments is of course that the theory is based on idealized conditions. It is assumed that no field quantities vary in the axial direction; actually the electron system is relatively short and the end effects are considerable. Furthermore, only one of the revolving field components produced by the RF voltage between the anode segments is taken into account. These components will be discussed in the next section.

As far as the scaling constants are concerned, some doubts have been expressed as to the significance of the diode current as a measure of the current in an oscillating magnetron. However, to the author's knowledge no superior alternative has been offered.

III. The Resonant System

In microwave oscillators resonators with distributed constants rather than lumped constants are used. In other words, resonance is characterized by a certain standing-wave formation rather than by the equalization of an inductive and a capacitive concentrated reactance. This means that the electrodes and the interaction space are immersed in the resonant system without any clear-cut demarcation surfaces. The

resonant system can be said to have three functions: to establish the RF field configuration in an interaction space required for the conversion process, to determine the frequency of operation, and to transfer the RF power from the interaction space to the output terminals.

1. The RF Field in Interaction Space

In the previous discussion of the interaction between the electrons and the RF field we have assumed that the latter constitutes a sinusoidal wave traveling around the interaction space with a sufficiently low phase velocity so as to enable the electrons to synchronize with it. We shall now examine the basis of this assumption.

Since the anode segments are located at different points in the resonant system, a voltage of different phase and magnitude will exist between

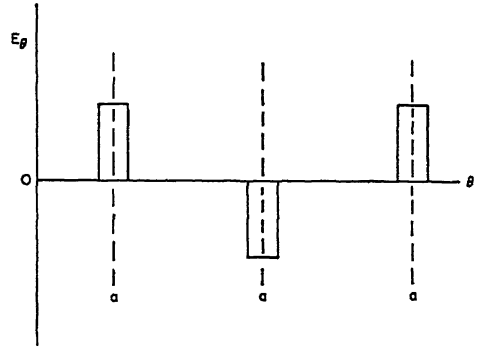


FIG. 4.—Variation of the tangential electric field in the π-mode at the anode surface. The dotted lines at (a) mark the center of the slots.

them. The most efficient use is made of space and geometrical complexity if the voltage at consecutive gaps along the anode is 180° or π radians out of phase. The magnetron is then said to operate in the π-mode.

Figure 4 shows the variation of the tangential field intensity along a cylindrical surface with the anode radius for π-mode operation. Such a field distribution is equivalent to the sum of an infinite series of components with sinusoidal space distribution. For an arbitrary mode number n this sum can be written

$$\varepsilon_\theta = \sum_{m=0,\pm 1,\pm 2}^{\pm \infty} A_m \exp\left\{j\left[\frac{2Y}{\lambda_p} + 2m\right]n\theta\right\} \tag{13}$$

where Y is the distance between centers of adjacent slots, λ_p the phase wavelength measured along the anode circumference, θ the angular

coordinate in a cylindrical coordinate system z, r, θ, coaxial with the interaction space, and n the mode number, i.e., the number of wavelengths along the circumference of the anode. The mirror image of $\mathcal{E}_\theta$, obtained by replacing θ by $-\theta$, is also a satisfactory expansion.

For a given mode number, $A_0 \exp(jn\theta 2Y/\lambda_p)$ is the fundamental component traveling towards negative θ, while other values of the number m in eq. (13) give shorter waves traveling in both directions. An exception is the π-mode; there $m = -1$ gives a wave equal to the fundamental in amplitude and wavelength but traveling in the opposite direction. Also the overtones appear for the π-mode in identical pairs of opposite direction of rotation.

This analysis derives its interest from the fact that the space charge may occasionally synchronize with one of the overtones instead of the fundamental. In π-mode operation, however, this is very unlikely to happen.

Equation (13) applies to magnetrons with multicavity resonators. Some of the other arrangements described below give a somewhat different expansion, because the mode number n must be given a different meaning.

2. *Classification of Magnetrons*

Magnetrons are divided into the following classes largely because of the fundamental differences in the arrangement of their resonant systems.

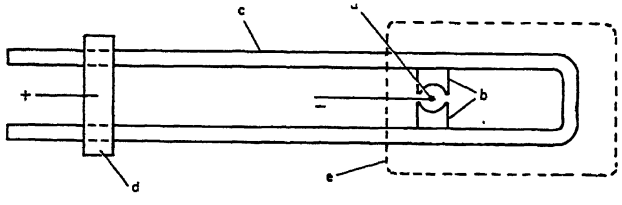


FIG. 5.—Split-anode magnetron; (a) cathode, (b) anode, (c) resonant line, (d) tuning bar, (e) glass envelope.

The *split-anode magnetron* represents an early design with glass envelope, filamentary cathode, and a parallel-line tank circuit, which is usually in part outside the vacuum (Fig. 5). It still has some applications where wide tuning range is important. An interesting modification is the *neutrode*, where a large section of the anode is connected to a neutral part of the tank circuit (Fig. 6).

The *multicavity magnetron* is at present the prevalent type for most applications in navigation, communication and physical research. Figure 7 shows some different designs schematically.

The *interdigital magnetron*[4] (Fig. 8) illustrates another approach to the resonant system with a multisegment anode.

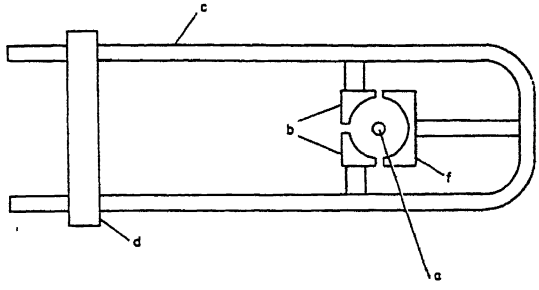


FIG. 6.—Magnetron with neutral electrode, neutrode; (a) cathode, (b) anode, (c) resonant line, (d) tuning bar, (f) neutrode.

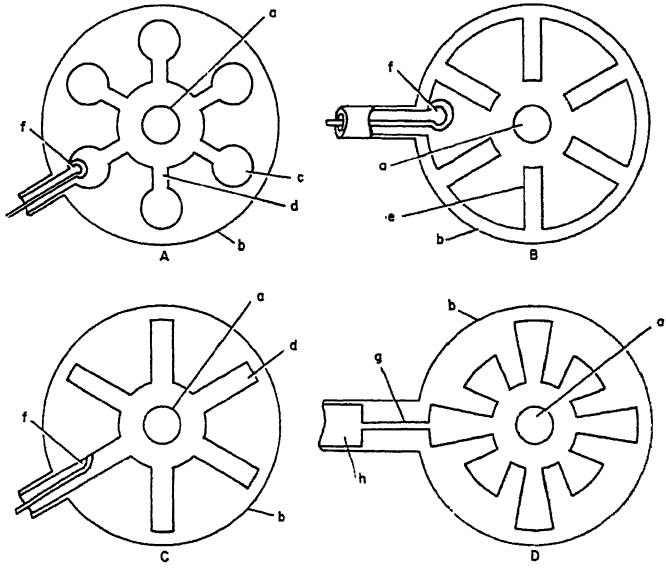


FIG. 7.—Multicavity magnetrons; (A) hole-and-slot magnetron, (B) vane magnetron, (C) slot magnetron, (D) rising-sun magnetron; (a) cathode, (b) anode block, (c) hole, (d) slot, (e) vane, (f) coupling loop, (g) output transformer, (h) wave guide.

3. General Discussion of the Resonant System

The phase difference between consecutive anode segments, preferably 180°, can be provided in fundamentally two different ways. These are demonstrated in Fig. 9 by showing the segments connected to transmission lines, the idea being of course that each line is connected back to

itself in such a way that the phase variation along the loop is an integer multiple of 2π radians. In Fig. 9b zero phase variation is also acceptable.

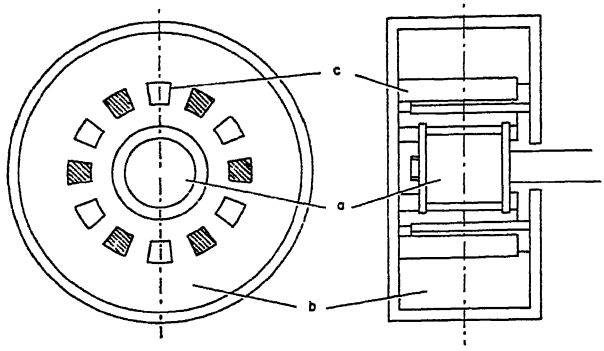


FIG. 8.—Interdigital magnetron; (a) cathode, (b) cavity resonator, (c) fingers.

The first arrangement is the fundamental multicavity anode system, while the second leads to the straight interdigital anode. Because both have their advantages as well as disadvantages most magnetrons are designed according to a combination of the two systems. The most common way of realizing such a combination is to connect alternate anode segments or vanes in a multicavity magnetron by *straps*. A conventional form is the *double-ring strapping* shown in Fig. 10.

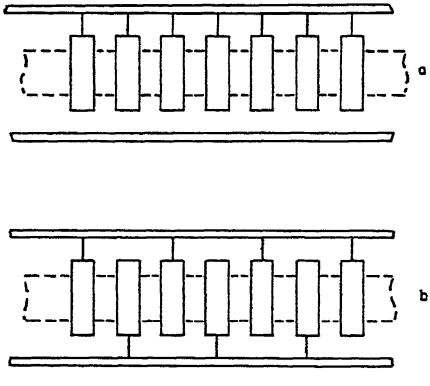
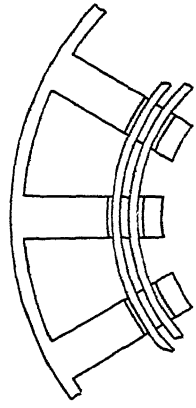


FIG. 9.—Interconnections between anode segments; (a) multicavity system, (b) interdigital system.

It is evident that there is a very close family relation between a ringstrapped magnetron and an interdigital magnetron. Consider the problem of providing an array of anode segments and ring straps with such structural support that the RF voltage between the two systems of opposite phase is not shortcircuited. If this support is given by quarter-wave slots in a solid block, one slot corresponding to each gap between the anode segments, the result is a strapped multicavity system. If, on the other hand, a quarter-wave resonator supports the whole length of the straps, we have an interdigital magnetron. In the former case the straps are supported

by the anode segments, or directly by the individual supports for the anode segments; in the latter the anode segments are supported by the straps, which now form the resonant cavity.

Historically the multicavity system was developed first, and the strapping was added as an afterthought. It is hard to escape the observation that this early tradition has had an unduly restraining influence on the development of the art. It is possible that an unprejudiced re-evaluation of the problem can give us designs superior to those in use at present. The interest shown in interdigital magnetrons in the last few years may at least in part come from a feeling that such a re-evaluation holds some promise. It would seem a more logical method of attack first to interconnect the anode segments in the desired manner and then to design a resonator around the result, rather than vice versa.



The main advantage of the multicavity resonator is the orthogonality of the resonator fields and transverse electro-magnetic (TEM) waves on the coaxial line formed by the anode block and the cathode and its supporting structure. Since the operating frequency always is in the cut-off region of all other wave types on this line, no RF power leaks out there.

The most objectionable feature of the multicavity structure is the insufficient mode separation. In order to get 180° phase difference between adjacent segments it is necessary to operate very close to the resonance frequency of the individual resonators. Consequently the phase velocity of the wave propagation along the resonator varies rapidly with the frequency, so that the difference between the frequencies for which the wave completes one turn around the anode in n and in $n - 1$ cycles, respectively, is small.

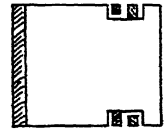


FIG. 10.—Double-ring strapping.

An ingenious solution of this difficulty is offered by the rising-sun system where alternate resonators have two different resonance frequencies.[5] The frequency of the π-mode is approximately halfway between these two frequencies, where the phase shift at neither of the two kinds of resonators varies very fast with the frequency (curved in Fig. 13). The reduced symmetry, however, produces an electric field component that is independent of the angular coordinate and under certain conditions reduces the electronic efficiency of the tube.

The chief merit of the interdigital magnetron is ample mode separation. The true π-mode or "zero-order mode," characterized by infinite

phase velocity along the "straps" is far removed from the higher order modes.

The most serious disadvantage of the π-mode interdigital system is its nonorthogonality to TEM waves on the cathode structure, which necessitates a choke arrangement to prevent loss of RF power via the cathode terminals. This is particularly obnoxious in broad-band tunable tubes.

The higher-order modes restore the orthogonality but they appear in degenerate pairs; i.e., two modes of identical frequency but different space distribution, and their expansion in revolving field components is not so favorable; it is therefore questionable if as high efficiency can be obtained as in the zero-order mode.

The interdigital magnetrons designed so far have not been readily adaptable for efficient cooling of the anode segments, and this has perhaps unjustly put the whole class into the category of low-power tubes. In general the interdigital magnetron has been approached with only a limited objective in mind, and its possibilities have not by far been exhausted. Furthermore, the broad field between the straightforward multicavity resonator and the straight-forward interdigital resonator has been explored only with the former type as a starting point. The opposite approach is likely to yield new and possibly more satisfactory solutions to the magnetron resonator problem.

As for the load impedance that the tank circuit offers to the conversion process in the interaction space, the pure multicavity resonator and the pure interdigital resonator are approximately equivalent. The strapped resonator and all other hybrids, on the other hand, have an inherently higher tank circuit capacitance and consequently lower slot impedance.

4. Equivalent Network of Multicavity Resonator

In order to calculate the properties of a resonator we should solve Maxwell's equations for each geometrically simple part of the resonator and match the field components at the boundary surfaces between the different parts. This is the correct procedure to obtain quantitatively correct results.[6] However, for the qualitative understanding of the functioning of the resonator it is helpful to find a lumped-constant analog or equivalent network of the resonator.

It has already been indicated that the resonator is analogous to a re-entrant artificial line. Since in general the phase wavelength in interaction space is short compared to the free-space wavelength, interaction space is predominantly capacitive, and we would expect the equivalent network to look like Fig. 11a. This is also correct, if the

axial length of the system is very large, or the holes or slots are closed at the ends.

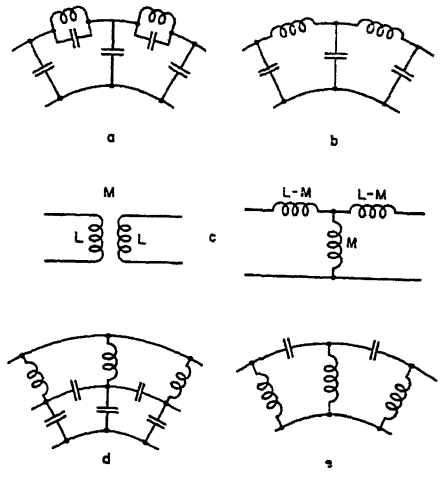


FIG. 11.—Equivalent networks of multicavity resonators; (a) a resonator with long or closed cavities; (b) the same network as (a) simplified; (c) equivalent T of coupled inductances; (d) equivalent network of resonator with short, open cavities; (e) the same network as (d) simplified.

The artificial line then propagates only waves of frequencies below resonance of the individual resonators, and is for such frequencies analogous to a low-pass filter (Fig. 11b).

If the axial length is not very long and the holes and slots open into end cavities the magnetic field lines from one slot will return through the adjacent slots, so that a magnetic coupling exists between the resonators. Figure 11a can easily be modified to account for this by introducing the equivalent T-network of two mutually coupled inductances (Fig. 11c). The most interesting result is for unity coupling (axial length considerably shorter than half a wavelength). The equivalent network is changed to a high-pass filter (Fig. 11d and e).

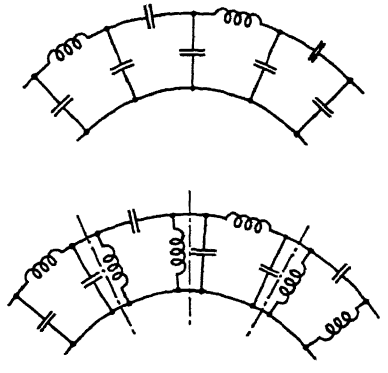


FIG. 12.—Equivalent network of a rising-sun magnetron.

The equivalent network of the rising-sun resonator can in a similar fashion be reduced to the form shown in Fig. 12.

The variation of the frequency of resonance with mode number is shown in Fig. 13 for a long unstrapped magnetron, for a short or heavily strapped magnetron, and for a rising-sun magnetron.

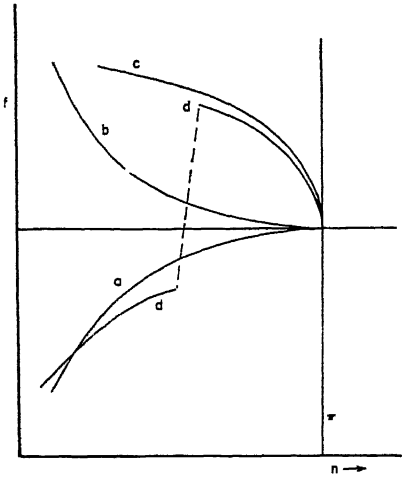


FIG. 13.—Variation of resonance frequency with mode number for (a) an unstrapped resonator with long or closed cavities, (b) an unstrapped resonator with short, open cavities, (c) a heavily strapped resonator, (d) a rising-sun resonator.

5. *Single-Mode Equivalent Network*

For operation in one mode only the magnetron oscillator can be represented by an equivalent network as shown in Fig. 14. The resonator mode under discussion is indicated as a simple resonant circuit LCG_0, the space charge as an admittance Y_s. The output coupling device is represented by an ideal transformer and a certain length of line l_c. Beyond the output terminals of the tube the transmission line or wave guide is terminated in a load admittance Y_L.

6. *Resonator Characteristics*

The important characteristics of the tank circuit as a whole are easily defined in terms of the constants of this network. The Q is the ratio of the stored energy to the energy dissipated per radian, or in terms of the equivalent network

$$Q = 1/\omega LG \tag{14}$$

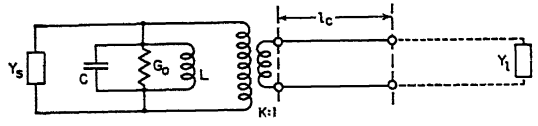


FIG. 14.—Single-mode equivalent network of a magnetron.

where L and G are the total inductance and conductance of the system. In order to separate losses from output power it is customary to use the terms internal Q, Q_0, and external Q, Q_E, the latter referred to a load equal to the characteristic impedance of the output line

$$Q_0 = 1/\omega LG_0 \tag{15}$$
$$Q_E = Z_0/k^2\omega L \tag{16}$$

The efficiency of the tank circuit, called the circuit efficiency of the magnetron is

$$\eta_c = \frac{1}{1 + Z_0 G_0 / k^2} = \frac{Q_0}{Q_0 + Q_E} \tag{17}$$

The *slot conductance*, i.e., the resonance admittance seen from the space-charge generator in interaction space, is

$$G_s = \frac{k^2}{Z_0} + \frac{1}{Q_0}\sqrt{\frac{C}{L}} = \frac{k^2}{\eta_c Z_0} \tag{18}$$

7. *Effect of Strapping*

The influence of strapping is usually explained with reference to such an equivalent circuit for each mode. The straps are connecting anode segments that in the π-mode have equal RF potential while the different sets of straps are 180° out of phase. The total effect of the straps is to increase the equivalent capacitance of the π-mode and reduce its resonance frequency. When the system is oscillating in any other mode, the straps will connect points of different phase position, so that considerable current flows through the straps. There will be a corresponding reduction in the voltage between straps of different groups. The equivalent inductance of this mode will be reduced because the straps by-pass part of the mode inductance, and the capacitance will be increased less than that of the π-mode. The effect is consequently to spread the resonance frequencies of the various modes farther apart from that of the π-mode; the change will be larger the larger the capacitance between the straps and the smaller the inductance of the straps.

8. *Output Coupling*

The transducer that extracts power from the resonator and delivers it at the output terminals is an important part of the tank circuit. Its design determines the constants k and l_c in the equivalent network (Fig. 14). The choice of k affects the Q, the circuit efficiency, and the slot conductance; it is a compromise between frequency stability and efficiency.

The design of the transducer is also affected by the output line (wave guide or coaxial line) and the frequency range covered by the magnetron. It is a pure circuit problem and will not be further discussed here.

IV. Cathodes

In a magnetron the cathode operates under conditions that differ considerably from those in other vacuum tubes. The main reason for this is the *back bombardment* or *back heating* produced by electrons that are accelerated instead of retarded by the RF field and consequently

return to the cathode with some surplus energy. This phenomenon has both advantages and disadvantages; the chief advantage being the increased emission obtained because of the secondary electrons produced by the bombardment; the draw-back, on the other hand, is the heating of the cathode by the impinging electrons.

In most radar applications the magnetron operates in short pulses, so that the *duty ratio*, the fractional operating time, is small, usually of the order 1:1000. Under such conditions the heating effect is unimportant, but the extremely high current densities obtainable from oxide-coated cathodes in short pulses is of great significance. This increased emission is probably not entirely due to secondary electrons; the process involved is not yet fully understood.

In continuous operation the opposite is true; the back heating presents a difficult problem and is a limiting factor in cathode design, so that the higher emission capability can hardly be utilized.

Some moderately successful attempts have been made to use cold cathodes functioning entirely by secondary emission, in conjunction with small thermionic emitters to initiate the oscillations. This additional complication in design, however, will probably be justified only for very high power at relatively low frequencies.

Since an article in the first volume of this publication has been devoted to the subject of oxide-coated cathodes very little need be included here. Before the war oxide-coated cathodes were used for a current density of the order of 0.1 amp./cm.^{2} In pulsed magnetrons the emission density is anything from 10 to 100 amps./cm.^{2}, and in continuous-wave magnetrons up to 1 amp./cm.^{2} The limiting factor in pulse operation is the sparking that occurs when the voltage gradient at the cathode surface becomes appreciable. It has been found that the purity of the nickel base is important to reduce sparking. Spot-welding a nickel screen of fine mesh to the nickel sleeve before application of the coating has also been found beneficial.

Considerable difficulty has been encountered in obtaining satisfactory life from oxide-coated cathodes in wide-range tunable magnetrons for continuous duty. Inefficient emitters, like tungsten or thoriated tungsten have been used for such applications. Because of the large heating power required, back heating is here less critical.

Sintered thoria cathodes have shown considerable promise for both pulse and continuous operation. Emission densities up to 50 amps./cm.^{2} pulsed and 1.5 amps./cm.^{2} continuously with satisfactory life have been reported.

The cathode problem becomes progressively more difficult as the frequency is increased, because the geometrical dimensions become

smaller and the current density at the cathode higher for reasonable output power.

V. Magnetron Characteristics

1. Performance Chart

The operation of a magnetron for varying anode voltage and magnetic field is conventionally presented in its performance chart (Fig. 15). The coordinates in this diagram are anode voltage and current; curves are drawn for constant magnetic field, constant output power, constant efficiency, and occasionally also constant frequency. Notable features are: 1. the rapid increase of current with voltage at constant magnetic field, and 2. the relatively uniform efficiency along the upper half of the constant-field curves.

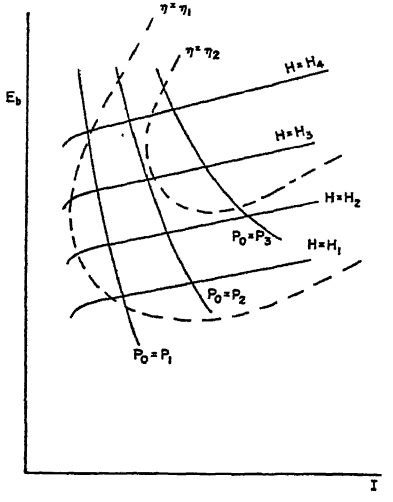


Fig. 15.—A magnetron performance chart.

2. Rieke Diagram

The reaction of a magnetron to variations of the load impedance are shown in the Rieke diagram (Fig. 16). The presentation is adapted to the microwave method of measuring impedance by standing-wave observations on a transmission line. The polar coordinate plane in this diagram represents the complex reflection coefficient of the load. All observations are taken for the same anode current, and curves are shown for constant power and constant frequency, respectively. The center of the diagram corresponds to matched load (zero reflection), and it is desirable to have the region of maximum output (the "power center") close to this point. The constant-frequency curves indicate the amount of frequency pulling obtained with variations in load impedance. There is of course a simple relation between the frequency pulling and the external Q of the resonator; a lightly loaded resonator will give a relatively high frequency stability.

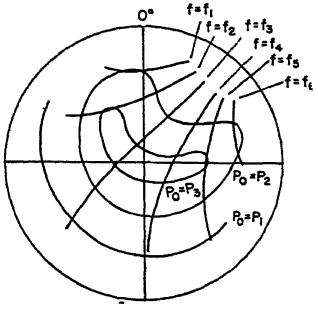


Fig. 16.—Rieke diagram.

3. *Mode Selection*

One of the most troublesome problems in the development of multicavity magnetrons has been the control of mode selection and the achievement of mode stability. Factors that influence the behavior of a magnetron in these respects are the frequency separation between the various mode resonances, the existence and amplitude of low-velocity revolving field components of these modes, and the properties of the power supply. In a diagram of anode voltage versus anode current there is for each mode a finite region where oscillation can take place (Fig. 17). These regions may or may not overlap. If the anode voltage rises slowly enough, the lower mode may be excited, and stable oscillations in that mode may continue, if the power supply has constant-current charac-

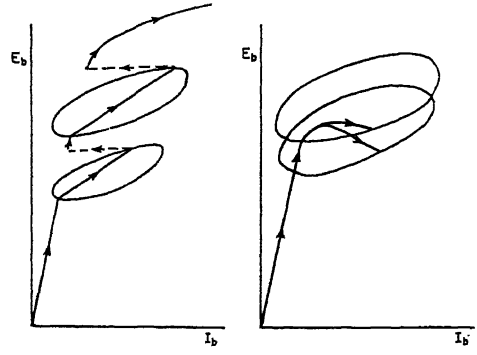


FIG. 17.—Oscillatory regions of two modes of a magnetron.

teristics. A fast rise and a constant-voltage supply may give stable oscillations in the upper mode without ever exciting the lower mode. Intermediate conditions may cause instantaneous excitation of the lower mode during rise and fall of the anode voltage.

The size and shape of the oscillatory regions depend on the load impedance, so that a change in load may cause a magnetron to shift to another mode. A pulsed magnetron may sometimes alternate in a random fashion between two different modes.

If the voltage rises too far, so that the limits of the uppermost mode are exceeded, oscillations will cease and only leakage current will flow, unless the critical voltage is reached, so that diode current can flow along the whole axial length of the electrode system.

4. *Noise*

The noise properties of magnetrons have attracted considerable attention in recent years.

The *preoscillation noise* plays an important part in the initiation of oscillations, particularly in pulsed oscillators, where short starting time is paramount. The preoscillation noise is, even below the threshold voltage, several orders of magnitude larger than predicted from calculation of thermal noise and shot noise. Unless another primary source of noise is found, it must be concluded that the space charge amplifies the noise considerably. It is not too difficult to imagine a process that will lead to this result. The field due to the primary noise accelerates some electrons, decelerates others. The latter will drift outside the Hull space charge. Under continued influence of the noise some of them will continue outwards while others are turned back. A few will ultimately reach the anode. A sparse, nonuniform space charge will thus exist outside the Hull cloud; whirling past the anode segments it will induce random currents of a different order of magnitude from the noise that initiated its drift. Since the procedure takes place below the threshold voltage and without any phase coherence, it is not a regenerative effect. It is more akin to the drift-space amplification in a klystron than to the operation of an oscillating magnetron.

The *noise modulation* of an oscillating magnetron has been traced to some definite sources but the problem is hardly yet completely solved. It is known that mode instability occasionally causes excessive noise. Evidence has been found that an oxide cathode is surrounded by evaporated barium and strontium of a pressure of 10^{-2} mm. mercury.[7] The back bombardment of the cathode causes ionization of this vapor, which is presumably one appreciable source of noise. It has been found possible to reduce the noise considerably by such a design of the cathode that the back bombardment does not hit any oxide-coated surface of the cathode.

VI. Modulation

The application of the magnetron for various communication purposes is contingent on the possibility of modulating its output power. It has already been pointed out that the magnetron is admirably suited for the pulse operation used in most radar applications. This is of course equally true in relation to all other pulse-modulation schemes with low duty ratio.

1. Amplitude Modulation

Conventional amplitude modulation can be applied to the magnetron by varying its anode voltage. There are, however, several drawbacks to this method. The modulation is not linear, and it is accompanied by considerable frequency modulation, since not only the interaction-space

conductance but also its susceptance is a function of the anode voltage and the amplitude of oscillation.

The frequency stability can be improved and the incidental frequency modulation reduced at the cost of lower output power and efficiency by loosening the coupling between the resonator and the load, i.e., by increasing the external Q of the resonator, or by increasing the stored energy in the system by means of an external resonator coupled to the magnetron tank circuit. The frequency modulation is still of serious proportions, however.

Amplitude modulation requires a high degree of mode stability, so that there is no risk that modulation peaks will produce mode shift or interruption of the oscillations.

Another method of amplitude modulation will be mentioned below in connection with a related method of frequency modulation.

2. Frequency Modulation

Frequency modulation or electronic tuning of magnetrons has been the subject of extensive work both during and since the war. The common idea of all the methods used is that, because of their motion, the presence of electrons in some part of a resonant system introduces a reactance that can be controlled either by the density or the geometry of the electron beam or cloud.

3. Electron-Beam Modulator[8]

If a dense beam of electrons is sent through a region in a resonator where the electric field intensity is high (Fig. 18) the electrons will

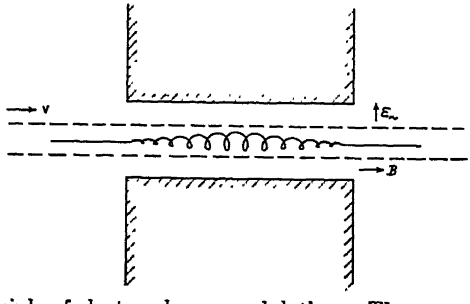


FIG. 18.—The principle of electron-beam modulation. The expanding and contracting helix represents the trajectory of an electron.

oscillate with the frequency of the field and with an amplitude that depends on the intensity and frequency of the field. The effect of these oscillations can be considered as a reduction of the dielectric constant of the space traversed by the beam.

The effective dielectric constant is

$$k = 1 - \frac{Ne^2}{\epsilon_0 m\omega^2} \tag{19}$$

where N is the number of electrons per cubic meter, e and m the charge and mass of the electron, and ϵ_0 the dielectric constant of free space in the mks system.

The change of the resonance frequency of the resonator depends on this constant and on the ratio of the energy stored within the volume of the beam to the total stored energy in the whole resonator.

At microwave frequencies the change is very small. One way to increase the frequency shift is to add a constant magnetic field parallel to the beam. This helps to focus the beam, and what is more important, it introduces a resonance phenomenon at the cyclotron frequency. If the magnetic field is adjusted so that the resonator frequency and the cyclotron frequency differ by only 10–30%, a considerably increased frequency shift is obtained. Figure 18 illustrates how an electron follows a helical path, the radius of which increases, decreases, increases again, etc., as the oscillations produced by the electric field beat with the natural cyclotron oscillation. In order to reduce the incidental amplitude modulation to a minimum, the transit time of the electrons should be adjusted so that the electrons leave the resonator with zero amplitude of oscillation. Either an increase or a decrease of the resonance frequency can be obtained, depending on whether operation takes place on one side of the cyclotron frequency or the other. A grid in the electron gun varies the electron density of the beam and controls the frequency modulation of the oscillator driving the resonator.

The resonator tuned by the beam may be the resonator of the magnetron. A number of beams may be used, each passing through one slot in the multicavity structure. In that case the magnetic field is supplied by the same magnet that maintains the magnetic field in the interaction space. As an illustration the following figures may be quoted for a magnetron operating at 800 Mc/sec. with 500 watts output. A frequency modulation of 4 Mc/sec. was obtained by means of five beams of 300 volts and a total current of 500 ma.

For very high frequencies the dimensions of the magnetron resonator becomes very small, and it is easier to apply the frequency modulation to an external cavity coupled to the magnetron resonator. Using this arrangement a 400 watt, 2700 Mc/sec. magnetron has been frequency modulated 15 Mc/sec. by a 500 volt beam of 150 ma.

If the cyclotron frequency is made equal to the operating frequency, no frequency modulation results, but the radius of the helical path keeps

increasing steadily as an electron proceeds through the resonator. An appreciable loss of RF power results, which can be electronically controlled by varying the density of the electron beam. Successful experiments using this scheme to amplitude-modulate a magnetron have been reported.[9] Frequency modulation caused by the varying load on the magnetron can presumably be compensated by a slight change of the cyclotron frequency. Reasonably good linearity is claimed for up to 85% to 90% modulation of a 1 kw tube at 850 Mc/sec. with a maximum frequency variation of ± 15 kc/sec.

4. Magnetron-Diode Modulator

In the electron-beam modulation described in the previous section the limiting factor is the relatively small electron density that can be realized in the beam. An appreciably higher density exists in a non-oscillating magnetron operated below the critical voltage, and considerable attention has been paid to the possibility of using such a magnetron as a variable reactance. In this case the electron density can not very well be controlled by a grid, but the radius of the electron cloud can be varied by means of the anode voltage. Since no appreciable anode current flows, the required modulator power is not excessive. The weakness of this modulator is that the RF losses cannot be minimized as neatly as in the electron beam modulator. Since the space charge is continuous from the cathode surface some electrons will strike the cathode and waste RF energy however small the amplitude of oscillation is. A certain amount of incidental amplitude modulation is therefore inherent in this scheme.

The cyclotron resonance can also here be expected to enhance the shift in resonance frequency.

A number of different designs have been suggested and a few of them tried. The simplest structure would be a diode with coaxial cylindrical electrodes; its variable reactance would tune a resonator coupled to the magnetron to be modulated. An ordinary magnetron operated below its oscillating voltage can also be used. A small-signal theory has been worked out for this case.[10] Like the electron-beam modulator the magnetron diode can be applied either directly to the resonator of the magnetron oscillator to be modulated or to an external resonator coupled to it. In the former case the space along the axis of the multicavity resonator has been divided between the two cathodes, the oscillator cathode and the modulator cathode, placed end-to-end, insulated from each other. Four megacycle per second modulation of a 3000 Mc/sec. magnetron with an RF loss of 20% has been reported. The input power of the modulator was 18 watts for an RF output of 35 watts.

With a 3000 Mc/sec. magnetron modulated by an external reactance tube the following results are quoted. The frequency modulation was found to have a broad maximum at $\omega_e/\omega = 2$ with less than 10% RF loss. The loss was greatest at $\omega_e/\omega = 1$, as could be expected. The frequency modulation increased rapidly with increased cathode radius. When the cathode radius of the modulator was 85% of the anode radius a frequency modulation of 20 Mc/sec. was obtained. The modulator power was large, 50 watts for 15 watts output, and the modulation nonlinear.

This method allows sufficient space-charge densities to make the effective dielectric constant negative. This means that the velocity of propagation of electromagnetic waves becomes imaginary, and the space-charge cloud acts as a perfect reflector. A coaxial cylindrical condenser with such a space-charge cloud as inner "conductor" has a capacitance that can be varied by means of the DC voltage applied. An 8% frequency variation with less than 10% change in output power is claimed for a 300 watt split-anode magnetron operating at 850 Mc/sec.[11]

The development of the synchronous wheel of space charge as the plate voltage approaches the threshold voltage could also be expected to create what could be called a synchronous susceptance. Presumably the anode voltage variations affect the phase difference between the spokes of the wheel and the RF voltage. Such an effect is the frequency pushing that is very much in evidence in the oscillating magnetron. Below the threshold voltage, however, it seems to be so small as to be completely masked by the condenser action of the expanding cloud.

Thus of the two ways of frequency-modulating magnetrons that have been proposed, one, the electron-beam method, has reached a certain maturity, while the other, the magnetron diode method, has been struggling under the handicap that no large-signal mathematical analysis of the magnetron space charge exists. In order to substitute empirical data for theoretical insight impedance measurements on the space charge below the threshold voltage are being made over wide ranges of the variables involved. However, considerable success has been obtained with a modulator working in the region of negative effective dielectric constant of the space charge.

VII. The Magnetron Amplifier

It is within the scope of a critical review of the present state of the magnetron art to point out important gaps. The failure of the art to include a power amplifier constitutes such a gap. The magnetrons are at the present time the only high-power high-efficiency generators of frequencies above 1000 Mc/sec. Their weakness is their low frequency

stability. An improvement by several orders of magnitude could be expected if master-oscillator-power-amplifier transmitters were developed. High-stability low-power oscillators for such transmitters are in existence, but not high-efficiency power amplifiers. The high efficiency of the conversion process in magnetron oscillators suggests that the same or a similar process be used for power amplification. The complicated continuous interaction process and the reentrant nature of the transmission line along the anode are magnetron features that have to be modified in order to make stable amplification possible. A comparison with the traveling-wave amplifier indicates that it can be done. A magnetron amplifier, however, would be superior to the traveling-wave amplifier as far as efficiency is concerned, because in a magnetron synchronization of an electron takes place at a low energy level, and the electron continues throughout its course to extract more energy from the constant electric field and to give it up to the alternating field, while in DC traveling-wave tube the electron enters the helix with maximum kinetic energy and can give up only a fraction of it to the RF field.

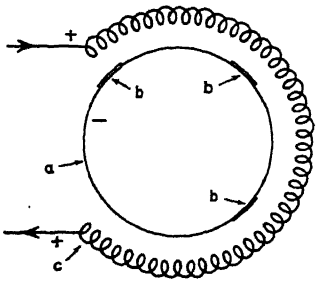


FIG. 19.—Warnecke's magnetron amplifier; (a) cathode, (b) thermionic emitter, (c) helix acting as anode and wave guide.

A magnetron amplifier designed and built in France has been described by R. Warnecke during a visit to the United States (Fig. 19). The electrons are emitted from a very small part of the cathode surface and interact with the RF field external to a flat helix which serves as anode. A second emitting surface marks the beginning of another stage of amplification. Performance figures quoted for a tube of this kind are 15 db gain and 10% efficiency at 200 Mc/sec. The efficiency can be expected to be lower than for a conventional magnetron oscillator, because the impedance of a nonresonant tank circuit is inherently several orders of magnitude lower than that of a resonant one. Since the optimum slot conductance can be achieved in a magnetron oscillator it is understandable that it is difficult even to approach it in a magnetron amplifier of this kind.

VIII. TRENDS IN MAGNETRON DEVELOPMENT

A certain stabilization and standardization in magnetron design has begun to take shape, particularly for the oldest applications, radar and navigational aids. Some important improvements, particularly as regards cathodes, are still in the research stage and may bring changes in

the conventional design. The problem of mode separation and mode stability is also a subject of active research and new solutions may produce a corresponding revision of the resonator design. Various strapping systems are being developed as well as interdigital resonators of different kinds.

The limits of power and frequency are still being pushed upwards. High power for continuous duty is required for various industrial and communication applications, while radar and particle accelerators in nuclear physics demand steadily higher pulsed power. In the highest frequency range the microwave spectroscopy is a relatively new field of application of magnetrons in the study of the basic laws of physics.

The magnetron theory still offers unsolved problems, inviting analytical as well as experimental study. The part played by random fluctuations in the operation of a magnetron is another subject of considerable interest.

Perfection of magnetron modulation for television and other communication continues to be an active field of research and development.

REFERENCES

No references are given to British and American wartime reports. The subject matter contained there has in general been adequately covered in comprehensive books and papers published since the war. The first reference below lists a number of those in addition to a translation of an early Russian paper. The magnetron volume of the Radiation Laboratory series is by far the most detailed and complete treatise of the subject published so far. No attempt has been made in this paper to document each fact presented by reference to chapter and page of these sources.

1. Alekseev, N. F., and Malairov, D. D. *Proc. Inst. Radio Engrs., N.Y.*, **32**, 136 (1944). Generation of High-Power Oscillations with a Magnetron in the Cm. Band.
Boot, H. A. H., and Randall, J. T. *J. Instn. Elect. Engrs.*, Part IIIA, **93**, 928 (1946). The Cavity Magnetron.
Megaw, E. C. S. *J. Instn. Elect. Engrs.*, Part IIIA, **93**, 977 (1946). The High-Power Pulsed Magnetron. A Review of Early Developments.
Willshaw, W. E., Rushford, L., Stainsby, A. G., Latham, R., Balls, A. W., and King, A. H. *J. Instn. Elect. Engrs.*, Part IIIA, **93**, 985 (1946). The High-Power Pulsed Magnetron, Development and Design for Radar Applications.
Fisk, J. B., Hagstrum, H. D., and Hartman, P. L. *Bell Syst. Tech. J.*, **25**, 167 (1946). The Magnetron as a Generator of Centimeter Waves.
Gorham, J. E. *Proc. Inst. Radio Engrs., N.Y. Wav. Electrons*, **35**, 295 (1947). Electron Tubes in World War II.
Harvard Univ. Radio Research Laboratory Staff. Very High Frequency Techniques. McGraw-Hill, New York, 1947, Vol. 1, Chapters 20–22.
Collins, G. B., editor. Microwave Magnetrons. (Radiation Laboratory Series) McGraw-Hill, New York, 1948.
2. Van der Pol, B. *Proc. Inst. Radio Engrs., N.Y.*, **22**, 1051 (1934). The Non-Linear Theory of Oscillation.

3. Hull, A. W. *Phys. Rev.*, **18,** 31 (1921); *Phys. Rev.*, **23,** 112 (1924).
4. Crawford, F. H., and Hare, M. D. *Proc. Inst. Radio Engrs., N.Y. Wav. Electrons*, **35,** 361 (1947). A Tunable Squirrel-Cage Magnetron.
5. Hollenberg, A. V., Kroll, N. M., Millman, S. *J. Appl. Phys.*, **19,** 625 (1948). Rising Sun Magnetrons with Large Number of Cavities.
6. Okress, E. C. *J. Appl. Phys.*, **18,** 1098 (1947). A Magnetron Resonator System.
7. Sproull, R. L. *J. Appl. Phys.*, **18,** 314 (1947). Excess Noise in Cavity Magnetrons.
8. Smith, L. P., and Shulman, C. I. *Proc. Inst. Radio Engrs., N.Y. Wav. Electrons*, **35,** 644 (1947). Frequency Modulation and Control by Electron Beams.
Shulman, C. I., and Kurshan, J. *Proc. Inst. Radio Engrs., N.Y. Wav. Electrons*, **35,** 657 (1947). A Frequency-Modulated Magnetron for Super-High Frequencies.
Donal, J. S., Jr., Bush, R. R., Cuccia, C. L., and Hegbar, H. R. *Proc. Inst. Radio Engrs., N.Y. Wav. Electrons*, **35,** 664 (1947). A 1-Kilowatt Frequency-Modulated Magnetron for 900 Megacycles.
9. Donal, J. S., Jr., and Bush, R. R. Convention paper, Inst. Radio Engrs., March 1948. A Spiral-Beam Method for the Amplitude Modulation of Magnetrons.
10. Lamb, W. E., and Phillips, M. *J. Appl. Phys.*, **18,** 230 (1947). Space-Charge Frequency Dependence of a Magnetron Cavity.
11. Peters, P. H. Convention paper, Inst. Radio Engrs., March 1948. A New Magnetron Frequency-Modulation Method.

Ferromagnetic Phenomena at Microwave Frequencies

GEORGE T. RADO

Naval Research Laboratory, Washington, D.C.

CONTENTS

The Gaussian system of units is used in this article. Each symbol denotes the same quantity throughout the text (e.g., c is used exclusively for the velocity of light in free space), but the letter μ, as specified in Section III, denotes a permeability whose significance is generalized whenever necessary. All vector quantities are printed in bold face type. The term "microwave frequencies," as used in the present article, is regarded as characterizing a particular frequency range of the radio frequency ("RF") spectrum; this range is taken to include frequencies from 10^2 to 10^5 megacycles per second ("Mc"), corresponding to wavelengths between 3 meters and 3 millimeters. Papers published since October 1948 are briefly mentioned in the "Addendum" at the end of this article.

I. Scope of the Present Article

Studies of ferromagnetism at microwave frequencies reveal a number of interesting phenomena related to the dispersion of the permeability, resonance absorption, and hysteresis effects. The present article is a survey of what seem to be the principal results and most significant considerations in this field. Several contributions are omitted as no attempt was made to prepare a comprehensive review or a historical presentation of the entire subject. Instead, particular emphasis is placed on a discussion of the more recent developments and their relation to the general problems of ferromagnetism.

Due to space limitations it is not feasible to include details of microwave techniques and electromagnetism; these topics are treated in a number of books[1] and articles.[2] It seems appropriate, however, to review briefly a number of results pertaining to ferromagnetic domains.

In Section II of the present article some of the important stages in the development of microwave ferromagnetism are considered. This summary proceeds largely along historical lines but chronology is occasionally sacrified in favor of an integrated presentation of related subjects. Since an attempt is made in this section to give an over-all account of the field before discussing specific problems, the material is somewhat condensed and definitions as well as fuller explanations are relegated to Sections III, IV, and V.

Section III contains a fairly extensive discussion of the concept of ferromagnetic permeability. A relatively large amount of space is devoted to definitions and physical interpretations of the various radio frequency permeabilities in order to facilitate the presentation of the material in the remaining sections.

Section IV takes up pertinent investigations prior to 1945 and includes a brief discussion of the results in terms of present knowledge. In Section V advances made since the beginning of post-war research are considered; this section is arranged according to experimental considerations rather than systematically. Due to the role of the skin effect, permeability measurements in metals (Section V, 1) present many problems different from those in semiconductors (Section V, 2), and for a number of reasons it is convenient to discuss ferromagnetic resonance (Section V, 3) under a separate heading.

In spite of several decades of research, microwave ferromagnetism is still not adequately understood and any survey of the subject must reflect the gaps in present knowledge. It is hoped that the information which follows will provide a general introduction to this challenging field.

II. Summary of Principal Developments

Ferromagnetic phenomena at microwave frequencies have been studied ever since Hertz' discovery of ultra-short electromagnetic waves. Even in the early investigations attempts were made to determine whether iron and other ferromagnetic metals could be magnetized in accordance with the rapid alternations of the magnetic field intensity.

Toward the turn of the century it was fairly well established that the initial permeability of iron and nickel at wavelengths of a few meters is of the same order of magnitude as at very low frequencies or in static fields. In 1903, moreover, Hagen and Rubens[3] reported on some infrared

experiments of fundamental significance to ferromagnetism as well as to the electric conductivity of metals. They measured the reflection characteristics of steel and nickel at wavelengths ranging from about 1 to 14 microns (corresponding to frequencies of the order of 10^7 Mc) and compared their data with the predictions of Maxwell's equations; both metals behaved as if their permeability were unity and the conductivity equaled the static value.

As a result of these investigations the dispersion of the permeability was known to occur in the spectral region containing the far infrared and the microwave range of frequencies. Arkadiew[4] subsequently demonstrated, in a series of experiments started in 1912, that the permeability of metals decreases from the static value to unity within the relatively narrow region of centimeter and decimeter wavelengths. Although Arkadiew's original theory of "magnetic viscosity" is in some respects inconsistent with modern concepts of ferromagnetism, he appears to have been the first to realize on the basis of dispersion data that the permeability at very high frequencies must be characterized by a phase as well as an amplitude. He also showed that the resulting representation of the permeability by means of a complex number is completely equivalent to a representation in terms of the two types of apparent permeability determined in his experiments.

Arkadiew's researches were extended by several investigators[5] during the years between the two World Wars and the general trend of his results was confirmed. It must be emphasized, however, that a survey of the literature reveals a striking lack of agreement between the various data on the frequency dependence of the apparent, initial, permeabilities. This is no doubt due to the widespread use of poorly defined magnetic samples, and to a number of experimental techniques which later proved to be faulty. Occasionally, moreover, the limited validity of certain theoretical results was apparently not realized. In this connection it should be mentioned that several investigators have found sharp peaks in their permeability versus frequency curves which were later shown to be due to experimental inaccuracies.

One of the most significant contributions in this period is due to Kreielsheimer[6] who studied the dependence of RF permeability on the amplitude of the high frequency field. His results indicate that the microwave permeability is largely independent of the field intensity, and that irreversible processes of magnetization are damped out at a much lower frequency than the reversible ones.

No generally accepted theory of ferromagnetic behavior at very high frequencies has yet been published. Some features of the experimental results, however, may be qualitatively understood by the Landau-

Lifshitz[7] resonance theory of domain spin rotations and by Becker's[8] theory of eddy current damping of domain wall displacements. Both theories have stimulated a number of advances in microwave ferromagnetism and their basic considerations are likely to be developed further. The most recent theory of dispersion is that of Kittel,[9] who ascribes both the decrease of the weak-field permeability at microwave frequencies, and the existence of an out-of-phase component of magnetization, to the incomplete penetration of the domains near the surface of a metal by the applied high frequency field. Although Kittel's theoretical model describes adequately the general trend of the dispersion in iron, recent experimental results indicate that the applicability of this theory may be limited to particular situations.

It appears probable that the characteristics of ferromagnetic dispersion are due to several fundamental phenomena (even in weak RF fields); a comprehensive theory is not likely to be formulated until the effects of the various processes of magnetization are separated and the domain structure of actual materials is clarified.

The rapid development of microwave techniques during World War II has given a new impetus to ferromagnetic research at very high frequencies. As may be expected, a great variety of accurate experiments may now be performed with the use of the highly perfected ultra-short wave triodes, cavity oscillators, wave guides, and coaxial lines. It is indeed a tribute to the competence of the early investigators that many significant results originated from experiments employing such relatively obsolete apparatus as spark-gap oscillators, Hertzian wire-gratings, and Lecher wire systems.

High frequency ferromagnetic measurements in metals are difficult because the conductivity severely limits the penetration depth of electromagnetic waves ("skin effect"). Although the attenuation introduced into a conducting system by the permeability of one of its walls is readily determined, a reliable measurement of the change in phase velocity due to ferromagnetism requires an exacting technique. The method of Johnson and Rado[10,11] involves both types of measurement and thus permits a simultaneous determination of both components of the complex permeability. By using a polarizing magnetic field (which is parallel to the high frequency field) as well as the frequency as independent variables, and comparing the RF and static permeabilities measured on the same sample, the relation between the magnetic properties of the surface layer and the interior of the metal could be established. The results of this work include a distinction between several reversible processes of magnetization characterized by different relaxation times, and an estimate of an upper limit for the size of certain

ferromagnetic domains. A particular type of hysteresis effect was also investigated.

Another post-war investigation is Birks'[12,13] experimental work on the microwave permeability of ferromagnetic semiconductors. In these ferrite-type materials and oxides the influence of the skin effect is negligible; but the preparation of the samples and problems related to their chemical composition introduce new difficulties that had to be surmounted. Birks' data include values of the complex permeability for the wavelength range from about 3 cm. (where the permeability is nearly unity) to several meters, and their general characteristics are those of a damped resonance. In other experiments Birks used a superposed static field and his results indicate the action of a strong-field effect in the high frequency magnetization. It is interesting to note that this effect is similar to the behavior of metals[10] previously studied.

With the exception of the Landau-Lifshitz theory the investigations mentioned up to this point did not consider explicitly the gyromagnetic effects resulting from the spin of the electrons responsible for ferromagnetism. The theories of Becker and Kittel, as well as the present interpretation of most of the experiments mentioned above, involve no assumption with regard to the elementary magnetic moments whose interaction causes the existence of regions of spontaneous magnetization. In the recent ferromagnetic resonance experiments, however, the principal effect is due to resonant absorption of radio frequency energy by the "gyroscopic" magnetic moment of the electrons. Expressed in classical terms, the resonance phenomenon occurs when the Larmor precession frequency of the electronic spins in a static magnetic field is equal to the frequency of an applied RF field. The *internal* fields in a ferromagnetic substance due to the crystalline anisotropy are known to be of the order of 1000 oersteds and hence the corresponding Larmor frequencies are in the microwave range. Nevertheless, the first successful ferromagnetic resonance experiments, those of Griffiths,[14] employed an *external* static field directed perpendicular to the RF magnetic lines of force. At the high field strengths used the material was practically saturated, so that the absence of complicated anisotropy effects permitted a determination of the gyromagnetic ratio, and hence of the Landé splitting factor, g, from the frequency and the static field measured at resonance. Griffiths' interpretation of his results was tentative, however, for the experimental values of g were considerably larger than the value $g = 2$ appropriate for electron spin. This difficulty was shown by Kittel[15,16] to be due to a certain demagnetizing field which contributes (in most experimental situations) an additional term to the classical equation of motion. When evaluated in terms of Kittel's theory,

Griffiths' data led to g values only slightly larger than two ($g \approx 2.2$), with one exception, other experiments[17–21] furnished similar results, but no satisfactory explanation of the residual deviation has yet been published. The fundamental processes responsible for the finite width of the resonance curve are not well understood, and present treatments of relaxation effects are largely phenomenological.

Beginning with the work of Jordan,[22] a number of interesting ferromagnetic phenomena have been observed in the frequency range between a few cycles and several megacycles per second. These include some of the after-effect and relaxation phenomena contributing to the so-called "residual loss" in ferromagnetic cores. Although these "low-frequency" effects are not within the scope of the present survey, it may be noted that many important developments in this field are due to Jordan,[22] Arkadiew,[4] Becker,[23] and Snoek.[24] The relation of these effects to ferromagnetic processes at microwave frequencies is by no means clear yet. There are indications, however, that the generalized concept of permeability (cf. Section III) will suffice for formal descriptions of the experimental facts. Future research, it is hoped, will integrate the various facts on time dependent ferromagnetic phenomena.

III. On the Concept of Ferromagnetic Permeability

The magnetic properties of matter may be characterized by specifying the relation between the vectors **B** and **H**, or **M** and **H**, of an electromagnetic field in the medium under consideration; here **B** is the induction or flux density, **M** the intensity of magnetization, and **H** the magnetic field intensity in the medium. Although the present discussion will be confined to *macroscopic* aspects of this problem,[6,25,26,27] the most general relations between the magnetic vectors are so complicated (since **B**, **M** and **H** are functions of position and time and are not necessarily parallel) as to be useless for most purposes. It is therefore appropriate to start with several simplifying assumptions with regard to the magnetic body to be described. The number of assumptions will be reduced only if additional basic properties of a medium (rather than incidental phenomena due to particular experimental configurations) may be characterized by such a generalization. Accordingly, only *homogeneous* and *isotropic* media will be considered throughout the major part of this section. For the present, moreover, it is assumed that the medium is free of permanent magnetization and that it is of infinite extent. Under these conditions the three magnetic vectors are uniform throughout the medium and **M** may properly be referred to as the magnetic moment per unit volume, or simply "magnetization." More important, however, is the fact that **B**, **M** and **H** are now parallel so that it is sufficient to consider

the magnitudes B, M and H. The equations

$$\mu = B/H \tag{1}$$

and

$$\kappa = M/H \tag{2}$$

may then be regarded as defining the permeability, μ, and the susceptibility, κ. These equations do not, in general, imply a linear relation (or a simple functional dependence) between B and H, or M, and H. Since $B = H + 4\pi M$, one obtains the equation

$$\mu = 1 + 4\pi\kappa \tag{3}$$

In the case of fields varying sinusoidally with time, the instantaneous values of B, M, and H, are described by the corresponding amplitudes and the frequency. Other types of time dependence, which occur less frequently in practice, can usually be represented in terms of harmonic components. As shown below, both μ and κ are generally complex quantities in all but static fields.

Under two specific conditions the foregoing simple description of magnetic properties may be retained for media of *finite* extent. 1) In the case of a closed magnetic circuit (e.g., toroidal core; cylindrical shell magnetized along the circumference) there is no demagnetization in the direction of closure so that in this particular direction the field H in the specimen may be identified with the externally applied field H_e. The physical meaning of μ may then be readily visualized: placing a sample into a magnetic field results in an increase, by a factor μ, of the flux density in the region now occupied by the sample. 2) In the case of an ellipsoidal specimen *uniformly magnetized* along a principal axis the solution of the appropriate boundary value problem leads to the simple relation

$$H = H_e - NM \tag{4}$$

where the demagnetizing factor, N, depends exclusively on the shape of the specimen. Finally, it should be noted that permanently magnetized media may also be included in the above mentioned description of magnetic properties provided those situations only are considered where H and the permanent component of M are parallel.

The typically ferromagnetic* properties of certain materials make it convenient to use the concept of permeability[6,25,26] in a somewhat

* A practical and accurate definition of the "cooperative phenomenon"[25] known as ferromagnetism will not be attempted here. It is useful, however, to recall a number of characteristic properties possessed by most ferromagnetic materials in static fields: (a) the permeabilities are substantially larger than unity; (b) saturation

generalized sense to indicate particular experimental conditions. With every point of the "normal magnetization curve" and of the "virgin curve," there is associated a differential permeability, $\mu_{\text{diff}} = dB/dH$, as well as an "ordinary" permeability, $\mu = B/H$. The former is the slope of the magnetization curve and may be regarded as the sum of a reversible permeability, μ_{rev}, which characterizes reversible phenomena, and an irreversible permeability, μ_{irrev}, which is caused by Barkhausen discontinuities. Thus

$$\mu_{\text{diff}} = \mu_{\text{rev}} + \mu_{\text{irrev}} \tag{5}$$

In the demagnetized state, $\mu_{\text{irrev}} = 0$, and $\mu = \mu_{\text{diff}} = \mu_{\text{rev}}$ for sufficiently *small* fields; each of these three quantities is then known as the *initial permeability*. Another important generalization of the concept of permeability is obtained as follows. A specimen whose magnetic state corresponds to a given point (B, H) on some particular magnetization curve is subjected to a *superposed* incremental field, H_{inc}, whose maximum value, H_{inc}', is arbitrary. If this field is altered in a *cyclic* manner then each pair of $(B_{\text{inc}}, H_{\text{inc}})$ values defines a point on a "minor hysteresis loop." With every point of the magnetization curve, therefore, there are associated several values of the *incremental permeability*, $\mu_{\text{inc}} = B_{\text{inc}}'/H_{\text{inc}}'$, since μ_{inc} is a function of H_{inc}'. *In the important limiting case of sufficiently small H_{inc}', however, the minor hysteresis loop degenerates into a straight line whose slope is independent of H_{inc}' and equal to the reversible permeability.*

At microwave frequencies it is useful to distinguish between two cases. 1) If there is no static magnetic field normal to the RF field then the experimental permeabilities are incremental permeabilities or special cases thereof; thus they may be represented by $\mu = B/H$, where **B** and **H** are now understood to represent the RF components ($\mathbf{B}_{\text{RF}}$ and $\mathbf{H}_{\text{RF}}$) of the fields. Kreielsheimer[6] measured the dependence of μ_{inc} on the RF field intensity and others[10] determined μ_{rev} as a function of a static field parallel to the RF field. Usually, however, the incremental permeability investigated was an initial permeability. 2) If, on the other hand, there is a static field normal to the RF field (as in the resonance experiments) then the nonequivalence of directions other than that of $\mathbf{H}_{\text{RF}}$ indicates that $\mathbf{B}_{\text{RF}}$ and $\mathbf{H}_{\text{RF}}$ are not, in general, parallel. Under these conditions it is customary to use the term permeability (or incremental

effects occur in experimentally attainable fields, and the permeability varies considerably with H; (c) hysteresis effects and permanent magnetization may be present; (d) the magnetic properties may show a marked dependence on the thermal, mechanical, and magnetic "history" of the sample, as well as on small impurities. All the typically ferromagnetic phenomena disappear at temperatures above the "Curie point."

permeability) to denote the ratio of the component of $\mathbf{B}_{RF}$ to the component of $\mathbf{H}_{RF}$ in some specified direction. However, a complete description of the magnetic properties of the medium requires the use of a *permeability tensor* relating $\mathbf{B}_{RF}$ and $\mathbf{H}_{RF}$. Polder[25] has derived this tensor for the special case of a completely saturated ferromagnetic substance. Although the present article is concerned primarily with the dependence of permeability on frequency and field intensity, it should be kept in mind that permeability is also a function of temperature, strain, and several other variables.

Permeabilities in static fields may be measured directly and the result of a given experiment can be represented by a single real number. (It is still assumed, as in eq. (1), that the material is macroscopically isotropic and homogeneous.) At high frequencies, however, μ cannot be measured directly, and it will be seen that two fundamentally different determinations of the electromagnetic properties of a substance are *necessary* to characterize the permeability. The result may be represented by two real numbers or by a complex number.

As mentioned in Section IV, 1, the RF permeability *at any given frequency* is determined by measuring the impedance, $Z_{meas} = R_{meas} + iX_{meas}$, of a circuit element containing the ferromagnetic material to be investigated. Using Maxwell's equations, this impedance may be calculated in closed form, $Z_{calc}(\mu) = R_{calc}(\mu) + iX_{calc}(\mu)$, for many geometrical configurations encountered in practice. Following the formulation adopted by Kittel,[9] the apparent permeability, μ_R, is then defined to be that value of μ which makes

$$R_{calc}(\mu) = R_{meas} \tag{6}$$

Similarly, another apparent permeability, μ_L, is defined to be that value of μ which makes

$$X_{calc}(\mu) = X_{meas} \tag{7}$$

Equations (6) and (7) define the *real* numbers μ_R and μ_L, respectively.

It is important to examine the relation between the "experimental" definitions (6) and (7) and the more fundamental definition (1). If one assumes the permeability to be a *real* number, it may indeed be defined by the amplitude ratio B/H; this requires, however, that the permeabilities determined by eqs. (6) and (7) be equal. A great number of experiments[5,6] prior to 1945 showed that μ_R and μ_L determined under practically identical conditions are, in fact, unequal, and that $\mu_R > \mu_L$. Post-war measurements of the complex permeability,[10] moreover, are equivalent to a simultaneous determination of μ_R and μ_L, and their results establish conclusively the inequality of the two apparent permeabilities.

Any adequate interpretation of the above mentioned facts requires a generalization of the concept of permeability. The most satisfactory procedure appears to consist in *assuming* that the permeability is a complex number. In defining the *complex* μ we may use

either $$Z_{\text{calc}}(\mu) = Z_{\text{exp}}, \tag{8}$$
or $$\mu = B/H \tag{9}$$

where B and H are now characterized by a *phase* as well as an amplitude. The definitions (8) and (9) are completely equivalent and appear to be sufficiently general to represent experimental results if a tensor representation is not required.

The concept of a complex permeability merits further consideration. If it is assumed that the existence of a permeability $\mu \neq 1$ gives rise to an energy loss in addition to the ohmic and hysteretic losses, then μ is represented by a complex number of the form

$$\mu = \mu_1 - i\mu_2 \tag{10}$$

In certain cases (e.g., ferromagnetic resonance), however, the permeability may be in the third quadrant of the complex plane.

Formally, a complex representation of μ is equivalent to the statement that the time dependent parts of B and H may be out of phase. Equation (10) may be written

$$\mu = |\mu| e^{-i\theta_\mu} \tag{11}$$

where $|\mu| = (\mu_1^2 + \mu_2^2)^{\frac{1}{2}}$, and $\theta_\mu = \tan^{-1}(\mu_2/\mu_1)$. Using the definition (9) and taking

$$H = H' \cos \omega t, \tag{12}$$

one therefore obtains

$$B = B' \cos(\omega t - \theta_\mu),$$

showing that B lags on H by the angle θ_μ. The primed quantities are amplitudes ($B' = |\mu| H'$), t denotes time, and ω is the circular frequency. It is interesting to observe that a graph of B vs. H is an ellipse whose significance is in some respects similar to the closed curve representing a minor hysteresis loop. In both cases the area of the loop is proportional to a magnetic energy loss per cycle, but the origin of the loss is different in the two cases. Corresponding to the complex μ there is a complex κ in accordance with eq. (2) and (3). The complex susceptibility is defined by

$$\kappa = M/H = \kappa_1 - i\kappa_2 = |\kappa| e^{-i\theta_\kappa}, \tag{13}$$

where $|\kappa| = (\kappa_1^2 + \kappa_2^2)^{\frac{1}{2}}$, $\theta_\kappa = \tan^{-1}(\kappa_2/\kappa_1)$, and $\kappa_2/\kappa_1 = \mu_2/(\mu_1 - 1)$.

Physically, a complex μ means that there may exist a magnetic loss, in formal analogy with the dielectric loss of an insulator characterized by a complex dielectric constant. The energy loss *per cycle* due to the out of phase component of magnetization is given by

$$\oint H\, dM = -\oint M\, dH = \pi\kappa_2 (H')^2, \tag{14}$$

and represents the energy absorbed from the RF field by a unit volume of the magnetic specimen. Equation (14) suggests that κ_2 (and hence μ_2) may be measured calorimetrically (cf. Section IV, 1) at any given frequency.

For want of a better designation, the magnetic loss connected with the out of phase component of magnetization in a ferromagnetic substance will be referred to as *ferromagnetic absorption*. As implied previously, ferromagnetic absorption may occur irrespective of ohmic losses (e.g., those due to eddy currents) and hysteresis losses. It should also be noted that the explanation of ferromagnetic absorption does not require postulation of a "magnetic current" analogous to the electric conduction current in certain lossy dielectrics. Instead, ferromagnetic absorption is ascribed to damping phenomena associated with the motion of the elementary magnetic dipoles in a ferromagnetic substance.

Magnetization processes in ferromagnetic materials are known to be of two kinds (cf. Section IV, 3): domain rotations and domain wall displacements. Both processes ultimately involve the reorientation, or "lining up," of the elementary gyroscopic dipoles, i.e., electron spins. It is clear that some sort of damping force (or forces) must act in these processes. If this were not so, numerous difficulties would arise: for example, a specimen could not be magnetized in the ordinary sense because the spins (or spin systems) would precess about the appropriate local field without lining up. A change in the applied field would merely alter the axis and speed of precession.

The detailed interpretation of the damping phenomena is a problem of atomic theory. But any damping force, whatever its origin, gives rise to a time delay in the magnetization. In static or low frequency fields this delay may not be detectable if the "relaxation time" associated with the damping is sufficiently small. At microwave frequencies, however, the ability of the dipoles to follow changes of the instantaneous applied field is severely limited by damping. This results in a phase delay in the magnetization, and causes a reduction of μ toward unity as the frequency is increased.

The well known behavior of polar molecules[29] at high frequencies is in some respects similar to the situation in ferromagnetism. In both cases

the polarization involves the orientation of dipoles, a process whose effectiveness is limited by damping. But the viscous forces in Debye's theory are due to completely different mechanisms than the damping forces in ferromagnetism. In addition, polar molecules are nongyroscopic, and generally they do not form domains.

Returning to the more practical aspects of high frequency permeability, it is now obvious that experimental results may be expressed either in terms of the two apparent permeabilities μ_R and μ_L, or by means of the complex permeability μ. This is already clear from the work of Arkadiew[4] who was the first to introduce these concepts. Although there is no simple and general relation between the complex μ and the apparent permeabilities, the following equations, derived by Arkadiew[4] and others,[10,16] are valid whenever there is a "strong" skin effect. This *restriction* means that the penetration depth of an electromagnetic wave is small compared to the thickness of the sample; the propagation constant is then proportional to $\mu^{\frac{1}{2}}$, and one obtains

$$\mu_1 = (\mu_R \mu_L)^{\frac{1}{2}} \tag{15}$$

and

$$\mu_2 = (\tfrac{1}{2})(\mu_R - \mu_L). \tag{16}$$

Thus

$$\mu_R = |\mu| + \mu_2 \tag{17}$$

and

$$\mu_L = |\mu| - \mu_2 \tag{18}$$

so that $\mu_R = \mu_L$ if the permeability is a real number ($\mu_2 = 0$).

It seems useful to list some of the symbols used in the literature to denote the high frequency permeabilities.

Symbol	Alternative symbols
μ_1	μ' μ μ μ_R
μ_2	μ'' q' ρ' μ_I
μ_R	μ_r μ_k
μ_L	μ_i μ_n

Throughout the present article, and in several recent papers, the symbols given in the first column are used. In the literature, μ_R is occasionally called the "outer permeability" or "consumptive permeability," and μ_L is sometimes referred to as the "inner permeability" or "conservative permeability." None of these four terms appears to have any fundamental significance.

Up to this point it has been assumed that the medium considered is macroscopically isotropic and homogeneous. In many important cases,

however, this assumption cannot be made. Single crystals of ferromagnetic substances, for example, are macroscopically anisotropic. Consequently the magnetization curve depends on the direction of **H** with respect to the crystalline axes, and the magnetization has, in general, a component normal to **H**. Corresponding to the two components of **M** one must therefore distinguish between a "parallel permeability" and a "normal permeability," both of which depend on the direction as well as magnitude of **H**. These permeabilities are ratios of particular field components. To obtain a general description of the magnetic properties of the medium it is necessary to represent the relation between **B** and **H** by a permeability tensor. Although single crystals are ideally suited for studies of fundamental magnetic phenomena, much information can be derived from measurements on polycrystalline specimens. The latter are *usually* isotropic *if* the proper mechanical and thermal processes are used in their preparation.

In general, a given ferromagnetic substance cannot be classified as macroscopically homogeneous or inhomogeneous. The classification depends on the type of experimental conditions considered. This is primarily due to the fact that ferromagnetic materials are characterized by a "semi-microscopic" magnetic structure. The elements of this structure are "small" regions called domains; they will be discussed briefly in Section IV, 3. For the present purpose it is sufficient to note that a given ferromagnetic substance may appear macroscopically homogeneous in most static measurements; at microwave frequencies, however, the same substance may seem to be macroscopically inhomogeneous. There are two reasons for this unusual behavior. First, the microwave skin depth in a ferromagnetic metal is often comparable to domain dimensions. Second, the microwave field in metals (and in some semiconductors) is nonuniform due to the skin effect.

Before discussing these basic reasons it should be pointed out that a number of accidental factors may differentiate the physical conditions near the surface of a sample from those existing in the interior. Some of these factors are stresses, impurities, and mechanical imperfections. Since the results of microwave measurements on metals refer to the "skin layer" only, it is extremely important to assess the effects of such factors even though their influence on static measurements may be negligible. Further comment on this point will be offered in Sections IV, 1 and V, 1, *b*.

The inhomogeneity due to domain structure has led Kittel[9] to use the term *effective permeability* for all measured permeabilities at microwave frequencies. This designation applies to the complex μ as well as to the apparent permeabilities. Since the implications of the term "effec-

tive" are rather important, it seems appropriate to discuss its physical significance.

The definitions of the three permeabilities, eqs. (6) to (9), were based on the assumption that B and H (and hence μ_R, μ_L, and the complex μ) are uniform throughout the sample. In the case of a skin effect, of course, the assumption of uniform fields is incorrect. Consequently the definitions can be retained, and used to define a uniform μ, only if the magnetic domains in the material satisfy a specific condition: they must be sufficiently small so that the field vectors do not change appreciably over the extent of several domains. Due to the small skin depths encountered at microwave frequencies, the dimensions of ferromagnetic domains may prove to be too large to fulfill this requirement; accordingly, the three permeabilities may be point functions of position. The use of the term "effective" may be regarded as a device for retaining the definitions and conclusions of this section in spite of the macroscopic inhomogeneity due to domains. Essentially, the new term amounts to introducing a hypothetical medium which is homogeneous (even in the presence of a skin effect) and cannot be distinguished from the actual medium by means of macroscopic electromagnetic measurements. To be more precise, eqs. (6) to (9) define the *uniform* permeability of a hypothetical sample whose surface impedance (or propagation constant) at a given frequency is the same as that of the actual sample. The definitions (8) and (9) are, of course, still equivalent, and the effective fields B and H are those used in Maxwell's equations.

For simplicity, the qualification "effective" will usually be omitted in the present article. Nevertheless, this important term must always be considered (in the case of time dependent fields) when interpreting experimental results by means of simplified theoretical models.

IV. Investigations Prior to 1945

1. Methods of Measurement

A comprehensive review of the determinations of apparent permeability prior to 1945 was published by Allanson.[5] His paper, combined with those of Arkadiew,[4] Kittel,[9] and Epelboim,[30] contains complete references on methods of measurement as well as on results. In the present section, only the general principles of the various methods are taken up, and certain fundamental inadequacies common to most experiments are pointed out.

Ferromagnetic permeabilities in static fields are measured in two intrinsically different ways. The ballistic methods are based on the electromagnetic induction law, whereas the magnetometer methods and

the inhomogeneous field methods depend essentially on the ponderomotive forces between magnets.

In alternating fields, on the other hand, the permeability has been derived from measurements of the resistance or the reactance of a circuit element containing the ferromagnetic sample. These measurements yield μ_R and μ_L, respectively. As explained in Section III, however, both quantities have to be measured to determine the (complex) permeability.

The term "circuit element," as used here, is to be understood in a general sense. At frequencies up to several megacycles it is usually possible to characterize AC circuits by lumped parameters. Thus the ferromagnetic specimen to be studied was usually the core of a coil (e.g., toroid), and its resistance or inductance was measured by means of an impedance bridge. Alternatively, the coil and a variable condenser were combined to form a circuit whose resonant properties depend on the resistance and inductance of the coil, and hence on μ_R and μ_L. One method of determining μ_L, for example, involved a measurement of the change of resonant frequency due to the insertion of the ferromagnetic core into the coil.

The representation of circuits in terms of lumped parameters is no longer valid at ultra-high and microwave frequencies. Accordingly, suitable "distributed parameter systems" were used for the permeability determinations. The apparatus employed included transmission lines, cavity resonators, and wire gratings. In all these cases the ferromagnetic metal to be investigated was made a part of the conducting system; ferromagnetic semiconductors were not studied in the pre-1945 period. Arkadiew[4] determined the permeability (both μ_R and μ_L) by measurements of the reflection of electromagnetic waves from gratings of fine wires, and several investigators,[5] notably Hoag and collaborators, obtained μ_R or μ_L from experiments on parallel (Lecher) wires. In more recent work, Zottu[31] and Glathart[32] used coaxial resonant cavities and coaxial lines, respectively; the ferromagnetic metal was the inner conductor in both experiments. Zottu measured the "Q" value of a cavity and obtained μ_R from this resistance determination. Glathart, on the other hand, made an inductance determination by measuring the wavelength (i.e., essentially the phase velocity) in a line, and derived μ_L from his data. The closed systems used by these investigators, as well as the waveguide techniques mentioned in Section V, minimize radiation losses and are superior in many respects to the open wire systems employed previously.

In the terminology of lumped circuits, the measurements of high frequency permeability (including the recent ones discussed in Section V) are essentially determinations of either resistance or reactance, or both.

The following table lists some of the related quantities that may be measured in practice.

Resistance	Reactance
Attenuation	Phase velocity, or wavelength
Absorption	Reflection
Resonator "Q"	Resonant frequency
Standing wave ratio	Position of voltage minimum

Measurements of the ferromagnetic or resistive properties of *metals* at microwave frequencies are complicated by the skin effect, which is of paramount importance whenever the conductivity, σ, is large. In particular, the effect of "imperfect metallic walls" ($\mu \neq 1$, $\sigma \neq \infty$) on the wave propagation in a conducting system is relatively small. It should be recalled, however, that a finite (> 0) attenuation, or a finite ($< \infty$) Q, is not too difficult to measure accurately. But a reliable determination of the *change* of phase velocity, or *change* of resonant frequency, requires special techniques. The fractional change $\Delta v/v$ of the phase velocity measured by a substitution method may, in part, be due to a difference in resistivity between the two metals that are being compared. Irrespective of any such difference, moreover, the contribution of ferromagnetism to the observed change is smaller than the quantity $(c - v)/c$, which is quite small itself. In addition, the effect on Δv of "unavoidable" geometrical changes, or imperfections, is much more important than their effect on the attenuation. The latter may be measured directly, but for practical reasons (unknown losses) it is often preferable to use a substitution method; in either case the value of σ must be known.

It is now well established that the conductivity of metals at microwave frequencies is a *real* number. Thus the difficulties due to the smallness of the reactance changes need not be surmounted if only a knowledge of σ is desired; this quantity may be obtained from resistance determinations *or* from measurements of reactance changes. The knowledge of ferromagnetic properties, on the other hand, requires *both* types of measurement. Since the permeability is usually complex, both μ_R and μ_L must be determined, and σ must be known from an independent experiment.

In view of these facts it is not surprising that the published measurements of μ_L are incomparably less accurate than those of μ_R. Although the experiments of Glathart[32] are an exception to this statement, certain results of Johnson and Rado[10] (cf. Section V, 1, *b*) indicate the necessity for more elaborate measurements. These involve[10] a (simultaneous) determination of μ_R and μ_L, an independent measurement of the RF value

of σ, and the establishment of the relation between the magnetic properties near the surface and in the interior of the metal.

Certain inaccuracies and inadequacies are common to many experiments in the period prior to 1945. For example, the conductivity at microwave frequencies was generally assumed to be equal to its static value. This assumption was justified by the results of the Hagen-Rubens experiments, or by (insufficiently accurate) measurements on nonmagnetic metals. Recent work by E. Maxwell[33] shows, however, that for wavelengths near 1 cm. σ is not necessarily equal to the static σ. The reasons for this deviation are not well understood, but the effect appears to depend on the condition of the metallic surface. At microwave frequencies it is undoubtedly necessary to determine[10] the high-frequency σ of the actual metallic specimen whose ferromagnetic permeability is being measured.

In evaluating the permeability (μ_R, μ_L, or μ) from impedance measurements (and separating out resistivity effects) one has to solve certain skin-effect equations which are derived from Maxwell's equations. These solutions are generally difficult to carry out even for the simplest geometries. In the case of a coaxial line, for example, one obtains Bessel functions of a complex argument. Certain approximations appropriate for a very strong skin effect have found widespread use, but it appears that the conditions for their validity have not always been realized. Inaccuracies of this kind, combined with spurious nonuniformities of the magnetic properties of wires, are probably responsible for the results of those investigators who reported a dependence[5] of the permeabilities on wire radius. However, one should not rule out the possibility that the RF permeability (or even the static μ) of a very thin wire may truly differ from that of a large sample of the same material; "genuine" effects of this kind may be expected if the domain structure in fine wires should ever prove to be nonrepresentative of the structure in large samples.

Since the skin effect makes it impossible to expose the interior of a metal to a high frequency field, the measured microwave permeabilities are properties of a thin surface layer only. One of the most serious deficiencies of the pre-1945 work is the absence of correlations[10] between static permeability determinations, which refer to the whole sample, and microwave measurements. Recent studies indicate that surface effects are sometimes at least as important in their influence on the measured microwave permeability as certain heat treatments and minor changes in chemical composition. It is a regrettable fact, moreover, that most investigators failed to state the mechanical and thermal history, and the chemical composition, of the samples used in their experiments. These

variables affect not only the microwave permeability but the static permeability also.

The facts mentioned above suggest that many of the pre-1945 results on the microwave permeability of metals should be treated with reserve. Particular caution is called for if the data are used to determine the relation between μ_R and μ_L. Quantitative comparisons of this kind may involve large errors of unknown magnitude because μ_R and μ_L were not measured on the same sample.

Calorimetric measurements of magnetic absorption in paramagnetic salts were performed successfully by Gorter[34] and others. In these experiments eq. (14) was used to determine the susceptibility κ_2. Similar measurements on ferromagnetic metals do not seem to have been made; this is probably due to the relatively large ohmic losses in metals. It is possible, however, that a calorimetric method may yet be applied to ferromagnetic semiconductors, although generally the electromagnetic methods appear to be more advantageous. A few investigators have used thermo-junction methods[5] involving a heater wire of the ferromagnetic metal to be studied. The high frequency resistance of the wire was determined by measuring its temperature. On the basis of these results, the value of μ_R *at the operating temperature* was calculated, but the real and imaginary parts of the permeability obviously could not be separated.

2. Results on the Initial Permeability of Metals

Most of the measurements on the microwave permeability of metals in the period prior to 1945 were concerned with the initial permeability. As the permeabilities at microwave frequencies are largely independent of field strength, and since polarizing fields were not used, the measured values of μ_R and μ_L represent initial permeabilities. The data on the frequency dependence of these quantities were summarized by Allanson[5] and Kittel.[9] Figures 1 and 2, taken from Kittel's paper, contain most of the reliable results on "iron" and "nickel" for frequencies above 100 Mc. Some useful data by Zottu[31] and Potapenko (quoted by Kittel[9]) were accidentally omitted in preparing the figures; a number of early results on magnetic spectra were not plotted since they are now discredited.

It is apparent that μ_R at a given frequency is larger than μ_L. This is in accordance with the concept of a complex permeability (cf. eq. (17) and (18)), and represents one of the few general conclusions that seem justified by the data. In the case of nickel it is particularly obvious that the frequency dependence of the permeabilities is hardly known. Even in iron, however, the discrepancies are rather serious. For example, E. Maxwell's value of μ_R on cold rolled steel is seen to agree with the

determinations of other investigators, but his equally careful measurement on electrolytic iron disagrees with other iron results. The inconsistencies in Figs. 1 and 2 are probably due to the factors discussed in Section IV, 1 rather than to ordinary experimental errors. In view of those factors it appears to be necessary to regard the data as *representative* only. The general trend of the results may be significant, but the individual data reflect particular surface conditions, uncertainties in

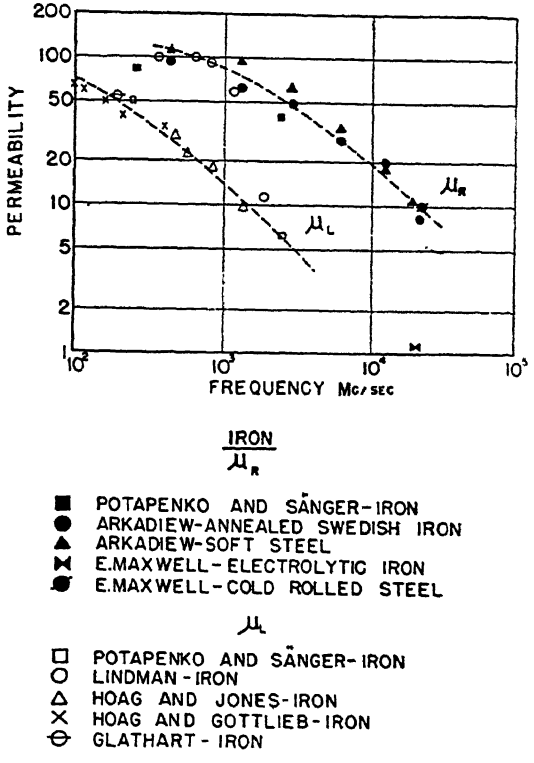


FIG. 1.—Compilation of experimental results on the frequency dependence of the apparent, initial, permeabilities of iron. (After Kittel.[9])

resistivity, and variations in chemical composition and in heat treatment. In the case of cobalt and ferromagnetic alloys the few available data, summarized by Allanson,[5] are even less certain.

No dependence of the microwave permeability on the high-frequency field intensity has yet been observed. Since large amounts of microwave power have never been available, however, the field strengths used were usually rather small (< 1 oersted). Consequently, the applicability of the negative results on the field dependence may be limited to low fields.

The careful experiments of Kreielsheimer,[6] which included frequencies up to about 6 Mc, show directly that for fields smaller than about 2 oersteds both μ_R and μ_L are independent of field strength. According to the trend of his data, and in agreement with some observations by Arkadiew,[4] this field-independence of the permeabilities applies most probably to the microwave region also; for small fields, therefore, no hysteresis effects are to be expected. In strong fields, on the other hand, Kreielsheimer found the apparent permeabilities to depend on the RF

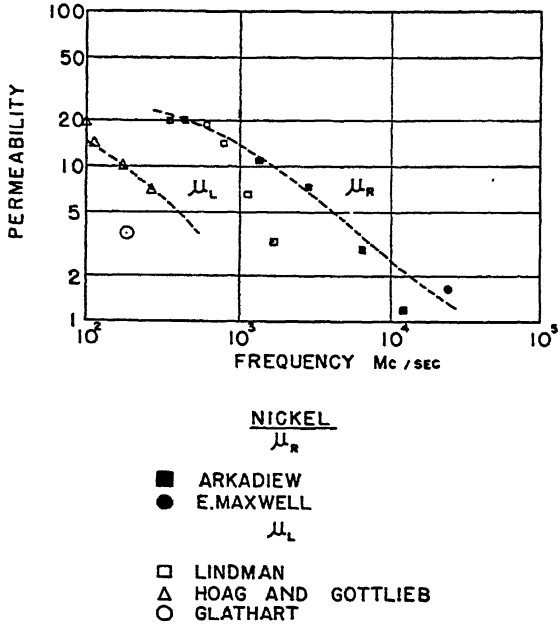


FIG. 2.—Compilation of experimental results on the frequency dependence of the apparent, initial, permeabilities of nickel. (After Kittel.[9])

field intensity. He observed, moreover, that the strong-field permeabilities begin to decrease (with increasing frequency) at frequencies below 1 Mc. Although this indicates that the irreversible phenomena (cf. eq. (5)) are damped out "earlier" than the reversible ones, effects due to irreversible processes could conceivably be detected at low microwave frequencies if sufficiently strong fields became available.

Prior to 1945, only a few investigations were concerned with the effect of variables other than frequency and field intensity on the apparent permeabilities at microwave frequencies. Glathart[32] measured the temperature dependence of μ_L for nickel at 200 Mc. He found that raising the temperature causes μ_L to increase gradually, to pass through a

maximum near the Curie point, and then to decrease abruptly. Sänger and Fejér,[5] as well as Glathart,[32] described some studies of the effect of tension on the microwave permeabilities, and Glathart[32] observed that μ_L for iron at 200 Mc is independent of small and medium static fields (< 100 oersteds) applied *normal* to the RF field.

3. *Theoretical Considerations*

According to the Weiss "molecular field theory"[35] of ferromagnetism, and its quantum-mechanical interpretation by Heisenberg,[35] a ferromagnetic substance should always be magnetically saturated. Ferromagnetic domains were originally postulated in order to reconcile this particular prediction of Weiss' remarkably successful theory with an unquestionable experimental fact: materials known to be ferromagnetic require a magnetic field to saturate them.

The outstanding weakness of the present domain concept is its unsatisfactory theoretical basis; an excellent discussion of some of the relevant problems was given by W. F. Brown, Jr.[36] If domains are assumed to exist, however, they may be used to explain a large variety of phenomena. The various experimental manifestations of domain structure include: the Barkhausen effect,[6] the Bitter patterns,[6,25] the Sixtus-Tonks reversals,[6] the depolarization of neutrons by ferromagnetic materials,[37] and certain electron diffraction experiments.[38] Néel's[39] recent theoretical work, and Williams'[40] observations of magnetic patterns on *unstrained* surfaces of single crystals, have substantially strengthened the foundations of the domain concept; but it should be emphasized that the various methods of domain observation do *not* necessarily indicate the same type of domain.

A considerable amount of evidence indicates, therefore, that a ferromagnetic substance contains discrete magnetic regions. Each of these regions, known as a domain, is spontaneously magnetized to a temperature-dependent saturation value. This means that the electron spins in a domain are nearly parallel; the alignment of the spins is due to the powerful "exchange force" described in books on quantum mechanics.[41] Most single crystals contain many domains, and in polycrystalline specimens the domain dimensions are *usually* smaller than the size of the crystallites. The various domains are magnetized in different directions unless the sample is saturated, but the change of the spin orientation between neighboring domains is gradual rather than abrupt. These changes occur in the relatively small transition regions known as "domain boundaries," "Bloch walls," or "domain walls."

In a demagnetized sample the directions of magnetization of the domains are distributed at random among certain preferred directions

called "easy directions of magnetization." Thus the net magnetization is zero, irrespective of whether the demagnetized sample is a single crystal or a polycrystal. The ferromagnetic anisotropy referred to above is usually attributable to crystalline fields; these fields are anisotropic due to "spin-orbit interaction."[42] It is possible, however, to create preferred directions of magnetization by applying stresses to a specimen.

The customary description of the magnetization process (cf. the books of Becker,[6] Bitter,[25] and Stoner[26]) is somewhat as follows. In weak fields the domain walls are displaced *reversibly* in such a way that the domains magnetized along easy directions close to the applied field become larger at the expense of unfavorably oriented domains. This mechanism determines the initial permeability, and accounts for the relatively small coercive force of many materials whose crystalline anisotropy is large. In medium fields *irreversible* wall displacements cause most of the domains in a crystallite to assemble in the easy direction nearest to the applied field. The magnetization curve in this range of fields is rather steep, and the irreversible displacements are detected as "Barkhausen discontinuities." In strong fields the magnetization of whole domains is turned toward the field direction. This process of "spin rotation" is usually reversible, and determines the approach to saturation.

The mechanism of both types of wall displacements is not well understood, but a theory originated by Becker,[43] which involves (undetermined) internal stresses, has led to a highly useful correlation of many experimental facts. Many unsolved problems remain, however, although some of the difficulties connected with the coercive force and hysteresis were removed by the recent "impurity theory" of Kersten.[44] Domain rotations, on the other hand, have been described successfully by the phenomenological theory originated by Akulov and developed by Gans.[45] According to this theory, the magnetization of a crystal is determined by the equilibrium between the torque due to the applied field, which tends to line up the domains, and the torque due to the anisotropy of the crystalline field, which tends to keep the spins in an easy direction of magnetization. With the latest modifications introduced by Lawton and Stewart,[46] this theory accounts satisfactorily for the upper part of single-crystal magnetization curves. In the case of polycrystals certain averaging processes were used (cf. Gans[45]), but it may be necessary to apply further considerations[10] to explain certain data.

The vague terms "weak," "medium," and "strong," chosen to characterize the fields in the description of the magnetization processes, were used purposely. Not only do conditions vary from one material to the other, but even in a given material wall displacements and domain rotations may occur in the same field[10] (cf. Section V, 1, *b*), so that the

standard description of magnetization is probably oversimplified. In fields of the order of the anisotropy fields ($\approx$ 1000 oersteds), however, the domain rotations predominate.

There is no reason to expect that the various magnetization processes will respond to high frequency fields in the same way. This statement is supported by the results of Kreielsheimer[6] (cf. Section IV, 2), who separated the dispersion due to reversible and irreversible wall displacements, and the investigations of Johnson and Rado[10] (cf. Section V, 1, *b*), who distinguished between some types of reversible wall displacements and rotations. The theories describing wall displacements and rotations at high frequencies are due to Becker,[8] and Landau and Lifshitz,[7] respectively.

In Becker's theory[8] the *reversible* wall displacements are determined by the instantaneous equilibrium between two types of force acting on a domain wall. The first is the applied sinusoidal magnetic field, and the second is a linear restoring force (proportional to the magnitude of the wall displacement) augmented by a damping force proportional to the velocity of the domain wall. These hypotheses lead to an equation for the displacement, and thus to an expression for the magnetization or the susceptibility. Becker's result may be written in the form

$$\kappa = \kappa_{stat} \frac{1}{1 + i\omega/\omega_c} = \kappa_{stat} \frac{1 - i\omega/\omega_c}{1 + (\omega/\omega_c)^2}, \tag{19}$$

where κ_{stat} is the static susceptibility, and ω_c is a critical circular frequency that determines a relaxation time $\tau = 2\pi/\omega_c$. The quantity κ_{stat} is introduced in the course of expressing the constant of the linear binding force by measurable quantities. To determine the damping constant, and hence ω_c or τ, Becker assumes that the damping forces acting on domain walls are the result of *microscopic eddy currents* induced by the changes in magnetization accompanying the wall displacements. Using an admittedly simplified model of domain structure, he then calculates ω_c in terms of κ_{stat}, σ, and the domain size. The numerical result represents a plausible estimate of the critical frequency: $\omega_c/2\pi$ for iron is 2×10^9 or 2×10^7 cps, depending on whether the domain size is taken to be 10^{-4} or 10^{-3} cm.

Equation (19) shows that at low frequencies ($\omega \ll \omega_c$) the walls (and hence the magnetization) follow the field. As ω increases, the magnetization decreases and begins to lag on the field. When $\omega = \omega_c$, the magnetization is reduced by a factor of $(\frac{1}{2})^{\frac{1}{2}}$, and the phase angle is 45°. Thus $\omega_c/2\pi$ is roughly the frequency at which the susceptibility, or permeability, is modified substantially. As the frequency is increased still

further, the susceptibility approaches zero ($\mu \to 1$), and the phase angle approaches 90°.

Becker's theory gives a good semi-quantitative account of the frequency dependence of the initial permeability. Using eqs. (3), (17), and (18), the predictions of equation (19) could be compared with the data of Figs. 1 and 2; but a detailed comparison of this kind would probably not be conclusive. In addition to the many experimental and theoretical difficulties mentioned in the present article, it must be remembered that Becker's theory applies only to wall displacements. It implicitly assumes, furthermore, that the domain size is small compared to the skin depth. The general idea of eddy current damping, however, has not been questioned, and it probably applies to domain rotations also.

Another successful aspect of Becker's theory is its treatment of *irreversible* wall displacements. Here the restoring force on a domain wall is negligible, and a simple calculation indicates that irreversible displacements are damped out at about 10^2 times smaller frequencies than reversible ones. As Becker himself has shown, this prediction is in good agreement with Kreielsheimer's experimental results (cf. Section IV, 2).

The theory of Landau and Lifshitz[7] considers the domain changes induced by a high frequency magnetic field in an anisotropic crystal. In order to simplify the problem, only one preferred direction of magnetization ("easiest direction") is assumed, and hysteresis effects are neglected. The calculation starts with the classical equation of motion

$$(1/\gamma)d\mathbf{M}/dt = \mathbf{M} \times \mathbf{H}_t + (\text{damping torque per unit volume}) \quad (20)$$

describing the precession of a gyroscopic magnetic moment, **M**, in a magnetic field $\mathbf{H}_t$. It should be emphasized that the total field $\mathbf{H}_t$ is the *resultant* of the applied RF field and the anisotropic crystalline field. To describe the approach of **M** toward $\mathbf{H}_t$, a phenomenological damping term (involving an unknown proportionality constant) is included in eq. (20). This damping term leads to a finite μ_2. The gyromagnetic ratio, γ, is usually expressed in units of $e/2mc$, where e and m are the charge and mass of the electron, respectively; thus

$$\gamma = ge/2mc \quad (21)$$

where g is the Landé splitting factor. For electron *spins*, $g = 2$, and $\gamma/2\pi = 2.80$ Mc/oersted.

In solving eq. (20), Landau and Lifshitz make use of the well-known phenomenological representation of crystalline anisotropy energy;[6,25] this representation is based on crystalline symmetry and contains experimentally determined constants. Both domain wall displacements *and*

rotations are considered in the calculation. The former occur when the RF field is parallel to the preferred crystalline axis, and the solution of eq. (20) shows a *damping* of κ (or μ) as a function of ω. Domain rotations, on the other hand, take place when the RF field is normal to the axis; in this case the dependence of κ (or μ) on ω exhibits a typical *resonance* behavior.

The theory of Landau and Lifshitz has not been applied to any of the pre-1945 experiments. This is partly due to the fact that the theory is based on the rather special model of a uniaxial crystal, and has not yet been generalized. As in Becker's theory, moreover, skin effect considerations are neglected.

One of the outstanding contributions of the Landau-Lifshitz theory is its recognition of the effect of internal fields on the rotation of electron spins. The theory indicates, for example, that ferromagnetic resonance may occur in *internal* (crystalline) fields. This important prediction seems to be confirmed by the recent experimental results of Snoek (Section V, 3) and Birks (Section V, 2).

In analyzing experimental data on the dispersion of the *initial* permeability (Section IV, 2), it has been difficult (or impossible) to separate the effects due to domain-wall displacements from those due to domain rotations. Both types of phenomena, however, must be characterized by laws of dispersion that satisfy the so-called Kramers-Kronig equations (cf. Gorter[34]) relating κ_1 and κ_2. *If* one of these susceptibilities is known for *all* frequencies, the other susceptibility may thus be calculated for any desired frequency by means of the Kramers-Kronig relations.

V. Recent Advances

1. Permeability of Metals

a. Skin-Effect Theory of Dispersion. The theories discussed in Section IV, 3 (Landau and Lifshitz, Becker) attribute the dispersion of the microwave permeability to resonance and relaxation phenomena. In Kittel's theory,[9] on the other hand, the experimentally observed effects are ascribed exclusively to the *incomplete* penetration of the domains (near the surface of a metal) by the applied microwave field; but no dispersion or absorption is predicted on the basis of the skin effect alone. It is *assumed*, in effect, that the dimensions of the domains contributing to the microwave magnetization are comparable to or larger than the skin depth. This assumption is indeed justified *if* the usual estimates of domain dimensions[6] (10^{-4} to 10^{-3} cm.) are used. Only the initial (weak field) permeability is treated in this theory, and the magnetization changes are assumed to be solely a result of domain-wall displacements.

In the limiting case considered by Kittel, the surface energy of the domain boundaries exceeds the magnetization energy. Consequently a domain wall moves *as a whole* in the direction of an applied field. This conclusion is valid even if, due to the skin effect, the penetration of the magnetic field is limited to a fraction of the thickness of a surface domain. At microwave frequencies the force acting on a surface-domain wall is obviously smaller than in the corresponding static case, so that the wall displacement, as well as the resultant magnetization and permeability, is decreased as the frequency is increased. It should be emphasized, however, that the theory rests on the crucially important assumption of rigid domain walls, and that the validity of this assumption may be questioned.

Kittel's calculations are based on a simple model. The ferromagnetic metal is replaced by a film which consists of a single layer of domains. This film is parallel to the magnetic field, and each domain is magnetized parallel or antiparallel to the field direction. The dimension of a domain along the surface of the film is assumed to be small compared to the film thickness. A solution of Maxwell's equations appropriate for this model is found, and the surface impedance (or propagation constant) so obtained is put equal to the surface impedance derived from the ordinary eddy current equation. Since the latter contains the effective permeability (cf. Section III), whereas the impedance derived from the film model does not depend on any RF permeabilities, the complex μ or the apparent permeabilities may be evaluated. The final expressions involve, in addition to the frequency, the domain dimension normal to the film surface, and the conductivity and the initial static permeability of the ferromagnetic substance.

Figure 3 shows a comparison of Kittel's theoretical results with the experimental data of Fig. 1 on the frequency dependence of the apparent, initial, permeabilities of iron. The theory is seen to describe adequately the *general trend* of these data. According to Johnson and Rado (cf. Section V, 1, *b*), on the other hand, certain experimental results on the ferromagnetic dispersion and absorption in iron lead to conclusions contrary to the assumptions used in Kittel's theory. The interpretation of their data indicates that the effective domain size in iron is *smaller* than the skin depth at 200 or 975 Mc, and that domain rotations as well as some types of wall displacements contribute to the initial permeability. Although the measurements of Birks[47] on the microwave permeability of mixtures of carbonyl iron *powder* and paraffin wax appear to support the validity of the theory, it should be noted that Birks' final results are based on considerable extrapolations. The permeabilities obey familiar concentration laws for mixtures up to 50% concentration, and data taken in this range are extrapolated to 100% iron. This procedure, as Birks

himself has pointed out, is not in general equivalent to measurements on iron in bulk form.

b. Dependence on Frequency and Polarizing Field. Using a method similar to that of Zottu,[31] Eichholz and Hodsman[48] measured the apparent, initial, permeability of iron, nickel, and steel at microwave frequencies. The ferromagnetic inner conductor of a coaxial line was replaced by a nonmagnetic metal of the same dimensions, and μ_R was

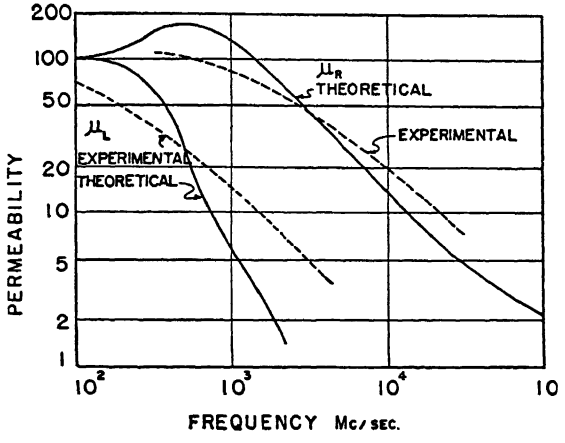


FIG. 3.—Comparison of smoothed experimental values for iron (cf. Fig. 1) with Kittel's theoretical calculations using a film thickness of 2.5×10^{-4} cm. (After Kittel.[9])

deduced from the measured change of attenuation; but μ_L, the other apparent permeability, was not determined. It is evident from the results obtained (Table I) that the value of μ_R (at a given frequency) in

TABLE I. Results of Eichholz and Hodsman[48] on μ_R.

Frequency (Mc) Wavelength (cm.)	0 ∞	2300 13	2600 11.5	3000 10	3400 8.8
Iron, bright	101	50	65	63	61
Iron, white annealed	140	41	60	67	65
Steel, bright	65	38	34	39	25
Steel, white annealed	102	31	31	28	28
Nickel	13	7	8	10	8

various types of iron and steel is quite different. This applies to a comparison of the two types of iron characterized by different static permeabilities, as well as to a comparison of bright iron and annealed white steel whose static permeabilities happen to be in good agreement. Eichholz

and Hodsman's results emphasize the important effect (cf. Sections IV, 1 and IV, 2) of chemical composition, heat treatment, and surface conditions, on the high frequency permeability. The data of Figs. 1 and 2, as mentioned before, should be regarded as representative only.

The method of measurement described by Johnson and Rado[10] differs considerably from previous methods. A determination of *both* components of the complex permeability ($\mu = \mu_1 - i\mu_2$) is accomplished by simultaneously measuring the changes in attenuation and phase velocity introduced into a conducting system by the ferromagnetism of one of its walls. The experimental technique, developed with the assistance of M. Maloof, involves the use of pulsed magnetic fields. Values of μ for samples of magnetic iron at 200 and 975 Mc *as a function of a polarizing magnetic field* (which is parallel to the high frequency field) are compared to the static incremental permeabilities measured on the *same* sample. The interpretation of the experimental results indicates several interesting characteristics of the domain structure and magnetization in iron. Among the other subjects of this investigation are studies of hysteresis phenomena and of the complex permeability of permalloy. Some of the latest experiments of Rado *et al.*[11] include measurements on nickel, and determinations of the effect of surface conditions on the complex permeability.

At the frequencies used so far, the ferromagnetic sample constitutes a part of the inner conductor of a half-wave coaxial resonator. A different resonant cavity is required for each frequency, but the same sample may be used in the various cavities. The measurements are based *essentially* on a substitution method. If the ferromagnetic part of the inner conductor is replaced by a nonmagnetic metal, the measured changes in the "Q"-value and resonant frequency of the cavity suffice to determine the complex permeability of the replaced metal. But the changes in resonant frequency, unlike the changes in Q, can also arise from small changes in the geometry of the resonator and require that the substitution be made with a mechanical exactness that is extremely difficult to achieve. This difficulty has been surmounted by magnetically saturating the sample with a strong static field that is parallel to the RF magnetic field. The incremental permeability (at any frequency) is thus reduced to unity, and the substitution of the magnetic sample by a nonmagnetic metal is effected without any mechanical changes. It is, therefore, possible to determine the complex permeability of the sample at any particular value of the polarizing field, H_0, by measuring the change of Q, and change of resonant frequency, due to an increase of H_0 from the original value to a saturating field. This means, of course, that the results are least accurate at the very large values of H_0 close to saturation.

The polarizing field is generated by a direct current passing through the center conductor. Currents of the order of 1000 amperes are needed to saturate the sample. Heating effects are avoided by employing a pulse technique;[10] this involves the use of a rather complicated "single pulse system" because the transient skin effect prevents the establishment of equilibrium conditions unless each pulse is of relatively long duration (several milliseconds). Since the frequency shifts due to ferromagnetic changes are rather small, the frequency is actually maintained constant, and the cavity is tuned by means of a dielectric bead.

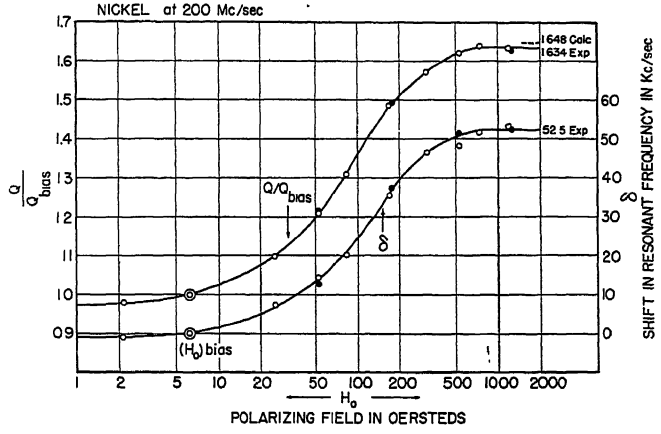


FIG. 4.—Representative data for annealed nickel at 200 Mc. The symbol "Q" denotes the cavity Q at a polarizing field H_0, "Q"$_{bias}$ is the value of "Q" for a biasing field $(H_0)_{bias}$, and δ is the shift of the cavity's resonant frequency due to a change of the polarizing field from $(H_0)_{bias}$ to H_0. Open and full circles refer to two experimental runs. Further explanations are given in the text. (After Rado *et al.*[11])

This procedure amounts to a measurement of resonant wavelength changes and permits (indirectly) a determination of frequency changes as small as 0.001% of the resonant frequency.

Figure 4 shows the type of data obtained for nickel and illustrates the smallness of the frequency shifts. It is interesting to note that the measured quantities, and hence the components of the complex μ, do not change appreciably until H_0 exceeds several oersteds. This behaviour is displayed by all the metals investigated so far (at 200 and 975 Mc); its important implications will be noted in connection with Fig. 5. The data of Fig. 4 indicate another significant fact. Since the value of Q in a saturating field may be calculated[10] in terms of static conductivities, the observed agreement (within the error of measurement) of the calculated and experimental Q-values proves that the RF and static con-

ductivities are equal. This applies to ferromagnetic metals (e.g., nickel) as well as to the nonmagnetic metal (brass) used for part of the cavity. It should be emphasized that this result is *not* affected by unavoidable "Q-deficits" because the calculation makes use of the measured (rather than calculated) Q value of the cavity with brass inner conductor.

The final results for magnetic iron shown in Fig. 5 are derived[10] from data similar to those of Fig. 4. It is seen that μ_1, the real part of the complex permeability, and $\mu_{s'at}$, the static incremental permeability,

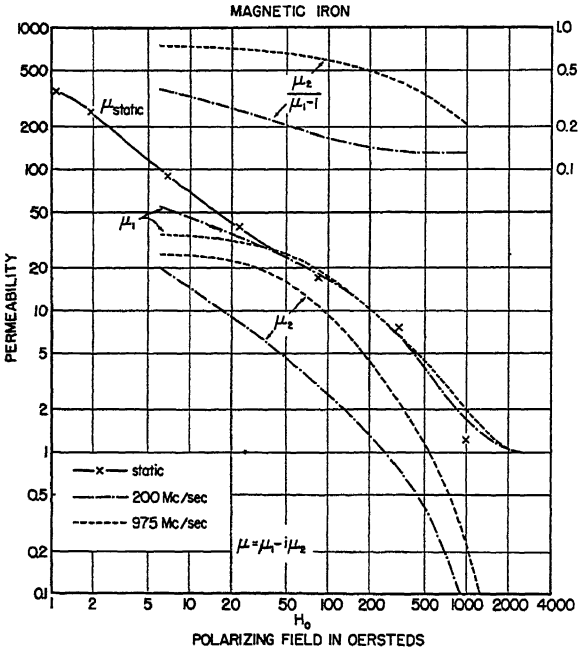


FIG. 5.—Experimental incremental permeabilities as a function of the polarizing field H_0 for annealed magnetic iron at 0, 200, and 975 Mc. (After Johnson and Rado.[10])

agree at large polarizing fields (> 50 oersteds). This equality is not predicted by any theory; for small values of H_0 and for very high frequencies μ_1 is, in fact, smaller than μ_{stat}. But the observed agreement of μ_1 (at two frequencies) with μ_{stat} over a considerable *range* of fields can hardly be coincidental and indicates convincingly that the magnetic properties in the surface layer of the sample are representative of those in the interior.

It should be noted that the domain size must be smaller than the skin depth in the range of fields where $\mu_1 = \mu_{stat}$; if this conclusion were incorrect, the imperfect penetration of the surface domains by the RF

field would result in μ_1 being smaller than μ_{stat}. On the basis of this argument, the data of Fig. 5 lead to an *upper limit of* 10^{-4} *cm.* for the domain size in iron as long as $50 < H_0 < 500$ oersteds. Now the domain dimensions in small fields cannot exceed this limiting value because the domains become larger as H_0 *increases;* neutron experiments indicate that in saturating fields the domain sizes approach those of the

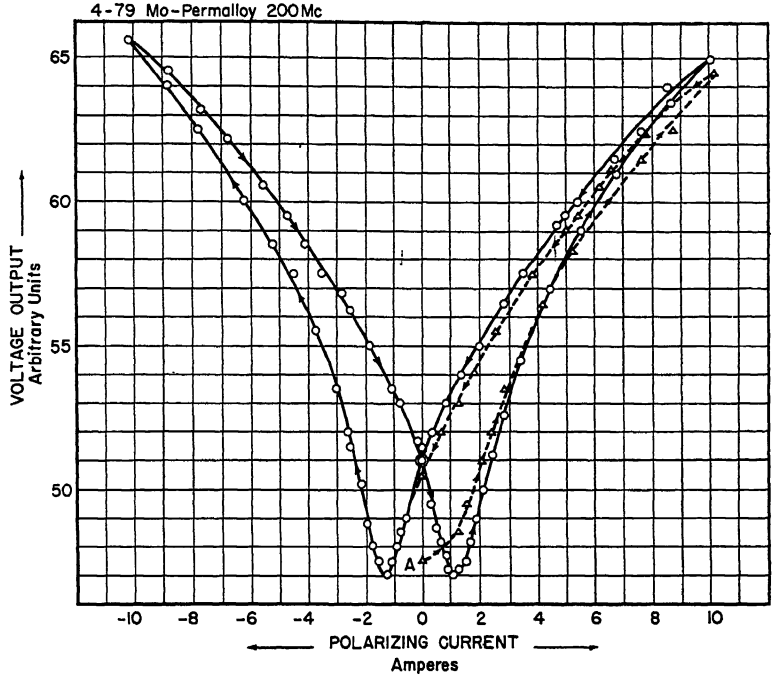


FIG. 6.—Typical butterfly curve (4–79 Mo-Permalloy at 200 Mc) obtained by plotting the rectified voltage output of the cavity as a function of the (unpulsed) polarizing current passing through the center conductor. The starting point (marked "A") represents the demagnetized state. (After Johnson and Rado.[10])

crystallites. Thus the upper limit quoted above applies to all polarizing fields smaller than 500 oersteds provided the domains move "rigidly."

Since the complex μ, unlike μ_{stat}, is practically independent of H_0 for small fields (not shown in Fig. 5), the high-frequency magnetization is a strong-field effect. It is particularly interesting to note that $\mu_1 < \mu_{stat}$ for small H_0, but $\mu_1 = \mu_{stat}$ for large H_0. A fairly detailed discussion of these facts, and of other aspects of Fig. 5, is given in Johnson and Rado's[10] paper. The interpretation adopted therein is based partly on eq. (19) and indicates the following properties of the magnetization in iron:

domains have different degrees of stability in the applied field; "weak-field domain" wall displacements are characterized by a relatively long relaxation time and are practically eliminated at 200 and 975 Mc; spin rotation and "strong-field domain" wall displacements, on the other hand, are only slightly damped at these frequencies, and are *responsible for* the RF magnetization in weak as well as strong fields; both spin rotation and "strong-field domain" wall displacements *contribute* to the *static* magnetization in weak as well as strong fields, but in weak fields the effect of "weak-field domain" wall displacements predominates.

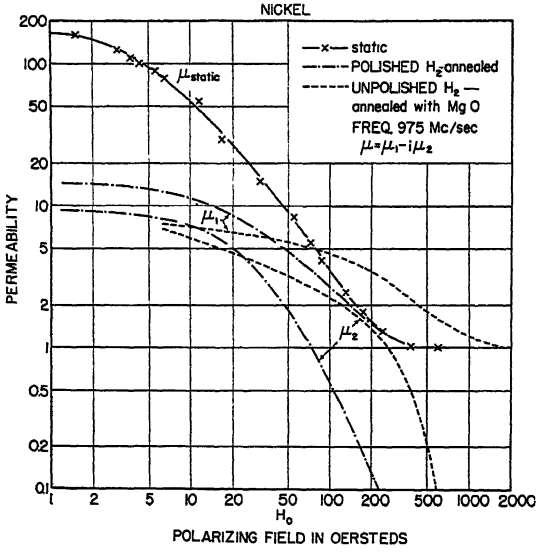


FIG. 7.—Experimental incremental permeabilities as a function of the polarizing field H_0 for nickel at 0 and 975 Mc. The two sets of 975 Mc curves refer to samples prepared under different conditions; the broken lines illustrate the combined effect of the use of MgO powder in the annealing, and of the lack of (mechanical) polishing. (After Rado *et al.*[11])

The RF permeability is found to be independent of the RF field strength for values around 0.01 oersted, so that the results of these investigations refer to *reversible* magnetization processes. Since the static magnetization, however, is subject to hysteresis with respect to the polarizing field H_0, the RF permeability is a double valued function of H_0. This effect does not appear in the data of Fig. 4 and 5 because they refer to polarizing fields applied in pulses starting from a uniquely magnetized state. In taking data of the type shown in Fig. 6, on the other hand, the polarizing field is applied *cyclically* to the ferromagnetic

specimen. (In some experiments a low frequency polarizing field was used to modulate the cavity transmission.) The measured voltage transmission of the cavity is related in a calculable way[10] to the reciprocal of $\mu_R^{\frac{1}{2}}$, and may thus be used to study hysteresis phenomena in terms of their effect on the reversible RF permeability. In the "butterfly curve" of Fig. 6, for example, the minima (which correspond to maximum μ_R and hence approximately to the coercive force due to reversible processes) occur at larger polarizing fields than in the corresponding static butterfly curves obtained by Elmen.[49] Although in the particular case of Mo-Permalloy this discrepancy is undoubtedly due to spurious surface conditions on the samples, it should be noted that a similar situation is found in other metals also. In iron, for example, no disturbing surface effects are observed (cf. the discussion of Fig. 5), and the large coercive force is ascribed to the "strong-field domain" wall displacements operative at 200 and 975 Mc.

Samples of nickel, unlike iron, must be polished[11] in order to obtain a surface representative of the interior of the sample. If the (mechanical or electrolytic) polishing process is omitted, the polarizing field needed to reduce μ_1 to unity is considerably larger than that required for static saturation. This "magnetic hardness" of the surface layer is enhanced if MgO powder (which does not affect μ_{stat}) is used during the anneal. Figure 7 illustrates the importance of proper surface treatment and suggests the possibility of employing microwave magnetic measurements in studies of metallic surfaces.

2. Permeability of Semiconductors

Birks[12] has described a method for the measurement of the complex permeability (and complex dielectric constant) of low-conductivity ferromagnetic materials at centimeter wavelengths. The ferromagnetic compound is mixed with a lossless, nonmagnetic binder (e.g., paraffin wax), and the solid, homogeneous mixture obtained after molding is inserted into a section of the coaxial line, or waveguide, employed for the RF measurements. Ferromagnetic metals may also be studied by Birks' method *if* they are used in the form of finely divided powders dispersed in a suitable insulating material (cf. the reference to Birks' work in Section V, 1, *a*).

Measurements on these types of material are considerably easier than those on metals because the difficulties due to the skin effect are negligible in low conductivity samples. The theoretical interpretation of the results is also simplified when no skin effect is present. It is necessary, on the other hand, to make measurements that permit a distinction between the dielectric constant, ϵ, and the permeability, μ. (In the

previous sections, only metals were considered; there the "apparent" dielectric constant is a large imaginary number because it is almost exclusively due to the conductivity.) Accordingly, the method of Birks involves the determination of *two* complex quantities: the propagation constant and the (relative) characteristic impedance of the transmission line section containing the sample. Both of these parameters depend on the complex ϵ as well as on the complex μ. The quantities actually measured are the (relative) input impedance of the "filled" section when it is terminated in a short circuit *and* when it is terminated in an open circuit. In these two cases the sample is backed by a metal plate and by a closed quarter-wavelength line, respectively. The impedances are measured by determining the standing wave ratio and the (relative) position of the voltage minimum in both cases. Although the techniques used for this purpose are well known,[2] it may be mentioned that Birks took special precautions to establish and maintain the response characteristics of his crystal detectors, and to minimize other sources of experimental error.

One of the problems in this type of work involves the concentration dependence of the electromagnetic properties of the mixtures. On the basis of a comprehensive series of investigations Birks established that in a number of cases the following empirical relations apply

$$\log |\mu| = V \log |\mu_f| \tag{22}$$

$$\log |\epsilon| = V \log |\epsilon_f| + (1 - V) \log |\epsilon_b| \tag{23}$$

$$\tan \theta_\mu = V \tan (\theta_\mu)_f \tag{24}$$

$$\tan \theta_\epsilon = V \tan (\theta_\epsilon)_f \tag{25}$$

The quantities on the left-hand side of these equations refer to properties of the mixture (θ_ϵ is the dielectric loss angle); V is the proportion by volume of the ferromagnetic powder; the subscript "f" refers to the properties of the ferromagnetic powder (extrapolated to 100% concentration), and the subscript "b" to the properties of the binder. As an example for the applicability of eq. (22) to (25), Birks states that they are valid for γFe_2O_3—wax mixtures over the range of wavelengths (60 to 3 cm.) and concentrations ($V < 0.41$) investigated. In preliminary studies the Clausius-Mosotti relation was used for the extrapolation, but it has been found[13] that this formula is of limited validity.

Although the results of Birks[12,13] shown in Figs. 8 and 9 represent extrapolations based on eqs. (22) to (25), the subscript "f" on the components of the complex permeability is omitted for simplicity. It is seen that in γ-ferric oxide as well as in magnetite there is a pronounced ferromagnetic absorption, and a decrease of $|\mu|$ with wavelength; accord-

ing to Birks,[13] the "similarity of behaviour of the two oxides is almost certainly due to similarity of crystal structure." No comprehensive theoretical interpretation of these results has yet been published. Birks[50] has stated, however, that these results, as well as further measurements on similar ferrite-type substances, all appear to be consistent with the

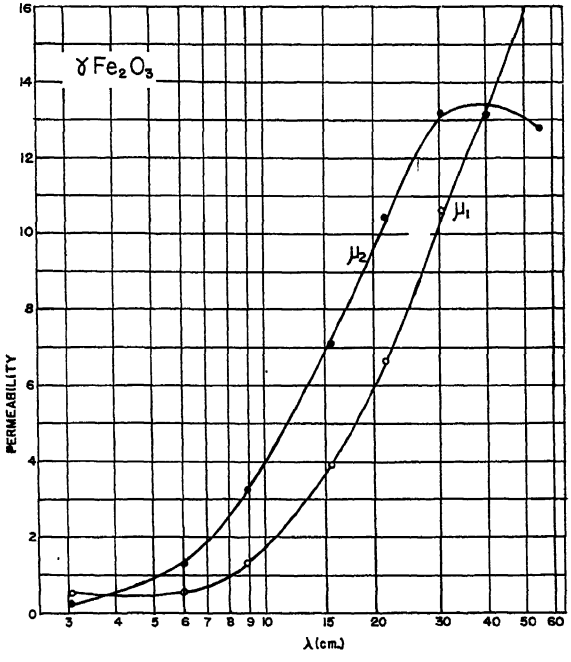


FIG. 8.—Magnetic spectrum of the initial permeability of γFe_2O_3. The results were obtained by extrapolating data taken on γFe_2O_3-wax mixtures to 100% concentration. The components μ_1 and μ_2 of the complex permeability are plotted as a function of wavelength. (After Birks[13]; this figure was prepared on the basis of an original drawing kindly supplied by Dr. Birks.)

general dispersion formula for a damped resonance. This relation may be written in the form

$$\kappa = \kappa_{stat} \frac{\omega_0^2}{\omega_0^2 - \omega^2 + 2i\omega\omega'} \tag{26}$$

where ω_0 is the (circular) resonance frequency, ω' is a damping coefficient, and the other quantities have been defined previously. The resonance is presumably due to internal crystalline fields in the manner predicted by the spin rotation theory of Landau and Lifshitz (cf. Section IV, 3) for a special case.

Figure 10 shows the results obtained by Birks[13] when he applied a static magnetic field perpendicular to a plane through the axis of the coaxial line. A quantitative interpretation of the data is complicated by the fact that this static field has a component both parallel and perpendicular to the RF lines of force. Since the permeabilities are found

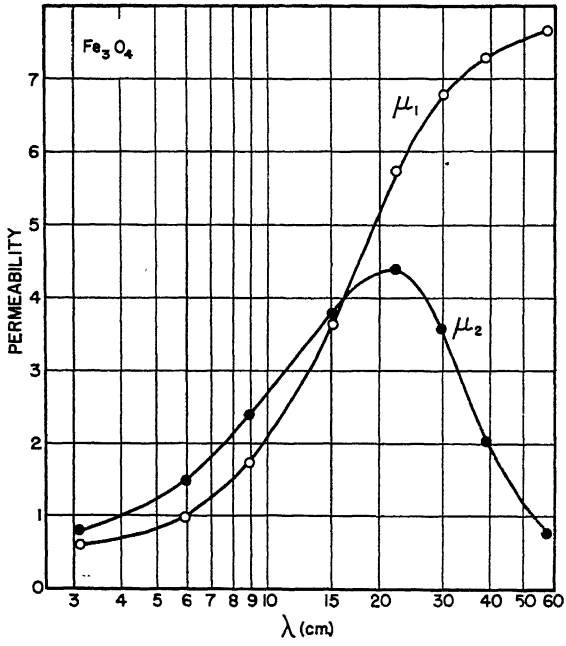


FIG. 9.—Magnetic spectrum of the initial permeability of Fe_3O_4. The results were obtained by extrapolating data taken on Fe_3O_4-wax mixtures to 100% concentration. The components μ_1 and μ_2 of the complex permeability are plotted as a function of wavelength. (After Birks[13]; this figure was prepared on the basis of an original drawing kindly supplied by Dr. Birks and contains a slight correction near the top of the μ_1 curve.)

to be insensitive to applied fields of several hundred oersteds, however, the RF magnetization is revealed to be a strong-field effect.

Although this particular behavior is similar to the situation in metals (cf. Section V, 1, *b*), there are definite differences in the RF characteristics of metals and semiconductors. Phenomena of the type shown in Figs. 8 and 9 can probably be attributed to internal resonance; no such effect has yet been observed in high-conductivity substances. It is interesting to note, on the other hand, that the iron data of Fig. 1 may be recalculated by the use of eqs. (15) and (16). The resulting curves of μ_1 and μ_2 vs.

frequency (or wavelength) are somewhat similar to those obtained in oxides (Figs. 8 and 9), but the maximum of μ_2 in iron is rather broad. However, a discussion of this fact, and of other dissimilarities between metals and oxides, would seem premature as long as the role of the skin effect in metals (cf. Sections V, 1, *a* and V, 1, *b*) is not well understood.

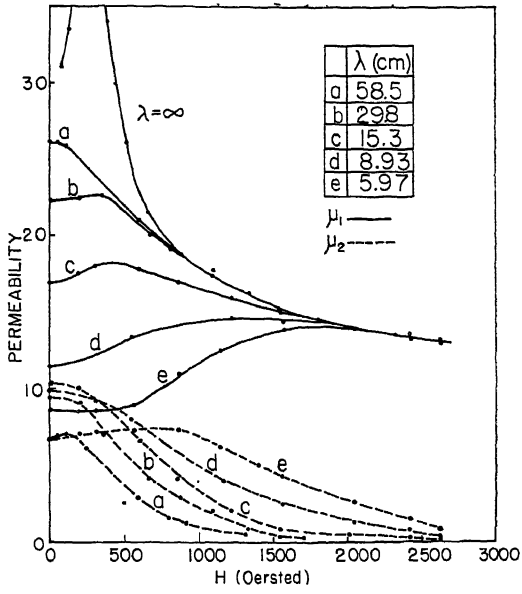


FIG. 10.—Experimental incremental permeabilities of γFe_2O_3 (32.6% by volume in wax) at various wavelengths. The components μ_1 and μ_2 of the complex permeability are plotted as a function of a static magnetic field. This field was applied in such a way that it had, in general, a component both parallel and perpendicular to the RF magnetic field. (After Birks[13]; this figure was prepared on the basis of an original drawing kindly supplied by Dr. Birks.)

3. Ferromagnetic Resonance

The brief description of ferromagnetic resonance given in Section II emphasizes the gyromagnetic character of this phenomenon. In the present discussion the subject is introduced on the basis of the principal facts discovered by Griffiths[14] in his first ferromagnetic resonance experiments at microwave frequencies.

Griffiths determined the apparent permeability μ_R of nickel, iron, and cobalt at wavelengths of about 1 and 3 cm. He deposited, by electroplating, a thin film of the ferromagnetic metal under investigation on a brass disk which formed one end of a cylindrical cavity resonator, and

deduced μ_R (actually μ_R/σ) from Q measurements. The essentially new part of his experiment, however, was the use of a static magnetic field, H_n, *normal* to the RF magnetic lines of force. This additional variable was one of the prerequisites for observing the interesting effect which Griffiths tentatively ascribed to resonance absorption of energy from the RF field.

When μ_R at a given circular frequency ω is plotted as a function of H_n, the resulting curve has the appearance of a broad resonance. Griffiths' results show, in particular, that there is a simple relation between the values of ω and H_n: *at maximum* μ_R (i.e., resonance) the ratio ω/H_n is approximately constant and independent of the ferromagnetic substance under investigation. Now in the case of a strong skin effect a large value of μ_R indicates a large μ_2 (cf. eq. (16)) and thus a considerable ferromagnetic *absorption*. Since the elementary magnetic dipoles are gyroscopic, the approximate constancy of ω/H_n at maximum μ_R suggests that the absorption may be due to a resonance phenomenon similar to that observed in electronic and nuclear paramagnetism. Accordingly, Griffiths attempted to interpret his data by using the relation

$$\omega_0 = \gamma H_n \tag{27}$$

where ω_0 is now written for the value of ω at resonance.

It should be recalled that eq. (27) is the well known expression for the Larmor frequency, ω_0, of a gyroscopic magnetic dipole (which is characterized by the value of γ) in a constant field H_n; ω_0, of course, has the significance of a *natural* precession frequency, and eq. (27) may be derived from the basic eq. (20) by using H_n for H_t, and neglecting the damping. To find that particular *forced* motion of the dipole which corresponds most closely to Griffiths' experimental situation, it is necessary to consider the effect of a small alternating magnetic field, of frequency ω, applied perpendicular to H_n. It may then be shown that the damping term in eq. (20) leads to an absorption (i.e., finite μ_2) which (for small damping) is a maximum when the applied frequency ω equals the natural frequency ω_0.

For an electron the value of γ corresponding to *spin* is given by the use of $g = 2$ in eq. (21). In most ferromagnetic substances, on the other hand, g is slightly smaller than two, and in some cases the deviations are known to be as large as 7 or 8%. These results were obtained largely by Barnett[51] and his collaborators in a series of experiments on the gyromagnetic effects associated with the names of Barnett and of Einstein and de Haas. One of the significant consequences of Barnett's results is the almost exclusive identification of electron spins (rather than orbital motions) as the effective carriers of the elementary magnetic moment in a

ferromagnetic substance. The orbital and spin motions of an electron are associated with the values $g = 1$ and $g = 2$, respectively; the observed g values (slightly smaller than two) are ascribed primarily to spin and to a small extent to orbital motion.

Returning to Griffiths' experiments, his results may be summarized by the statement that the use of his experimentally determined ω_0 and H_n in eq. (27) leads to values of g two to six times larger than the value $g = 2$. Since g should not exceed 2, it is clear that either the observations are not due to resonant absorption, or that the measured field H_n does not correspond to the H_n in eq. (27); the measurement and interpretation of ω_0 cannot reasonably be questioned. One possibility of error is the action of a demagnetizing field of the type mentioned in connection with eq. (4). However, the static field (as well as the RF field) in Griffiths' experiments is parallel to the plane of the sample; since the sample is thin, the demagnetizing effect is negligible. It is apparent, moreover, that a correction of this type decreases the effective value of H_n (and thus increases the apparent g) unless the demagnetizing field is so large as to reverse the field direction. Griffiths attempted to explain his anomalous results by calculating, with the use of Lorentz' formula, the local field acting on an elementary dipole in the metal; but he realized that in this case the Lorentz expression is a crude approximation at best.

It is appropriate to mention one additional fact before considering further interpretations of Griffiths' results. If the ferromagnetic resonance experiment is performed on a polycrystalline specimen, the observed permeabilities should not necessarily be expected to show resonance characteristics. The combined effect of the random orientation of the crystallites, and of crystalline anisotropy, may act to broaden or completely destroy the resonance even if the measured H_n satisfies eq. (27). It seems probable that the success of Griffiths' experiments is due, in part, to the use of frequencies which are sufficiently high to require, by eq. (27), applied fields of several thousand oersteds. (For electron spins $\gamma/2\pi = 2.80$ Mc/oersted.) In such large fields most ferromagnetic materials are practically saturated, so that, as pointed out by Kittel,[15] the whole specimen acts essentially as a single domain magnetized parallel to H_n.

A remarkably successful interpretation of Griffiths' apparently anomalous results was given by Kittel.[15] His theory is based essentially on the following idea. The total field H_t (cf. eq. (20)) acting on the specimen includes not only the static field (H_n) and the small microwave field, which are applied parallel to the plane of the specimen and normal to each other; H_t also includes a *time-dependent* demagnetizing field which is determined largely by the shape of the specimen. This dynamic

demagnetizing effect profoundly influences the motion of the magnetization vector and is not to be confused with the static demagnetizing effect that may modify H_n (cf. eq. (4)); both effects are due to the same basic physical law but only the static effect was understood (and taken into account) before Kittel's work. In the case of a plane specimen (like that in Griffiths' experiment), the component of the instantaneous induction *normal* to the plane must vanish at all times because the divergence of the induction is always zero; but this means that normal to the plane there *must* be a (demagnetizing) field equal to (-4π) times the instantaneous normal component of the magnetization. The use of this additional field in eq. (20) led Kittel to the resonance condition

$$\omega_0 = \gamma(B_n H_n)^{\frac{1}{2}} \tag{28}$$

which replaces eq. (27) in the case of a plane specimen. The quantity B_n is the induction corresponding to H_n, and its value may be obtained from the known static magnetization curve of the ferromagnetic material under investigation. Equation (28) was confirmed, according to Kittel,[16] by a quantum-mechanical calculation due to Van Vleck (to be published).

Kittel has shown that the use of Griffiths' data in eq. (28) leads to g values (≈ 2.2) close to the value $g = 2$ which obtains for an electron spin. This is an impressive result, especially since the theory contains simplifying assumptions that are only approximately fulfilled in the experiments. Kittel assumes, for example, that the sample is magnetically saturated, and that H_n is sufficiently strong to permit neglecting the magnetic anisotropy forces.

To test Kittel's theory, Yager and Bozorth[17] performed an experiment similar to those of Griffiths. However, the experimental conditions were chosen so as to conform closely to Kittel's theoretical assumptions. The material used, Supermalloy, has a very low magnetic anisotropy and it can be saturated in unusually low fields. Since the magnetostriction is also small, all anisotropy effects are presumably negligible. Thus the resonance in this substance is very sharp (Fig. 11), and an accurate determination of the effective g value is possible. Using Kittel's eq. (28), $g = 2.17$ is obtained, whereas Barnett's gyromagnetic measurements on an alloy of similar composition resulted in the value $g = 1.91$.

The shape of the resonance curve depends, of course, on the damping forces acting on the electron spins. According to Kittel,[16] the particular curve of Yager and Bozorth may be accounted for by a certain type of phenomenological damping term in eq. (20); but a detailed physical description of the relaxation process has not been given. It is interesting to note, incidentally, that the results shown in Fig. 11 include μ_R values smaller than unity. This fact may possibly be used in practical applica-

tions because it implies that *under certain conditions* the attenuation of Supermalloy to microwaves is smaller than the attenuation expected on the basis of the known resistivity of this material.

Recent results by Yager[18] show that the width of the resonance curve increases if the material is cold worked; however, the position of the maximum (and hence the effective g value) remains unchanged.

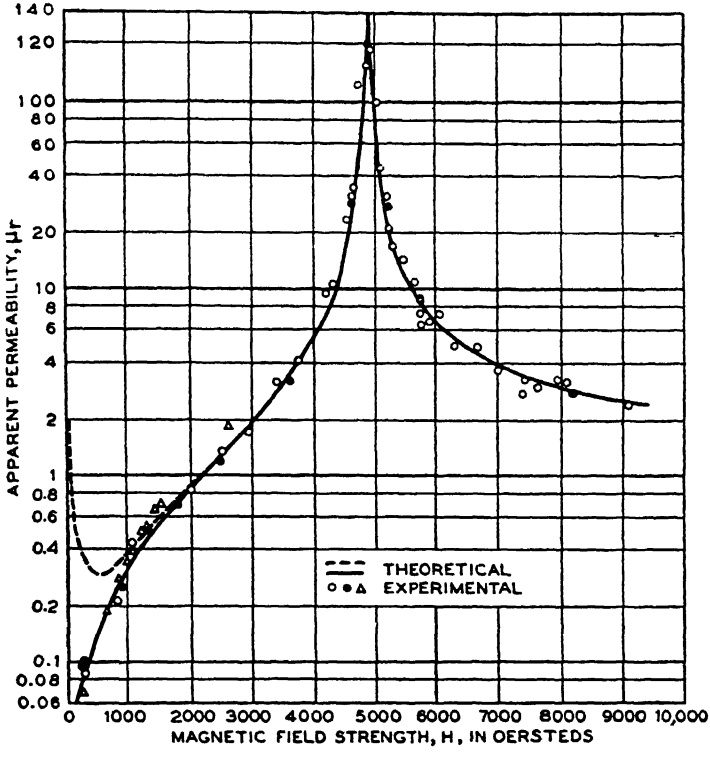


FIG. 11.—Ferromagnetic resonance in Supermalloy at 24,050 Mc. The apparent permeability μ_R is plotted as a function of the normal field H_n; the two theoretical curves refer to two kinds of assumed damping. (After Yager and Bozorth.[17])

Kittel[16] has extended his theory to include the dependence of the resonance condition on the shape of the specimen. In the case of single crystals, moreover, his generalized theory predicts the dependence of the resonance condition on crystal orientation and crystalline anisotropy energy. Some of these effects have been observed experimentally by Hewitt[20] (shape dependence in polycrystals) and by Kip and Arnold[19] (directional dependence in single crystals). Kittel's predictions have been confirmed, at least approximately, in all the cases studied so far.

Figure 12 shows Hewitt's results on ferromagnetic resonance in a ferrite material and includes curves for the components of the complex permeability and of the complex dielectric constant. Since the conductivity of this material is low, μ_1 and μ_2 could be separated easily, so that more information was obtained than in the case of metals where the

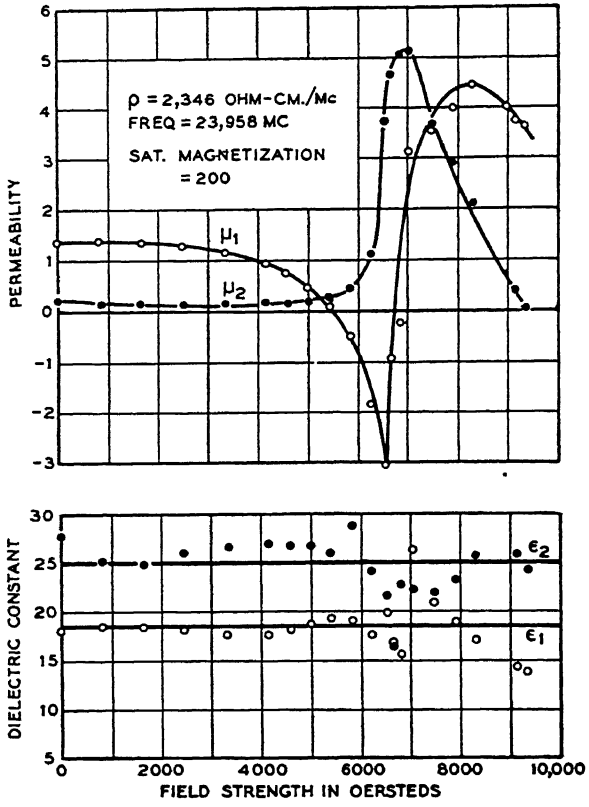


FIG. 12.—Ferromagnetic resonance in $(MnO)(ZnO)\cdot 2Fe_2O_3$ at 23,958 Mc. The real and imaginary parts of the permeability ($\mu = \mu_1 - i\mu_2$) and dielectric constant ($\epsilon = \epsilon_1 - i\epsilon_2$) are plotted as a function of the normal field H_n (without correction for demagnetization). (After Hewitt.[20])

resonance experiments have yielded μ_E only. The curves for μ_1 and μ_2 are seen to be typical of anomalous dispersion and absorption, respectively. At first sight these curves may appear to have an unexpected shape. It should be noted, however, that the abscissae are H_n values; thus they represent variations of ω_0 whereas in the familiar dispersion curves the independent variable is the applied frequency ω. The

specimen used in this experiment was plane, and the resonance data (corrected for demagnetization) yielded an effective g value of 2.12 on the basis of eq. (28).

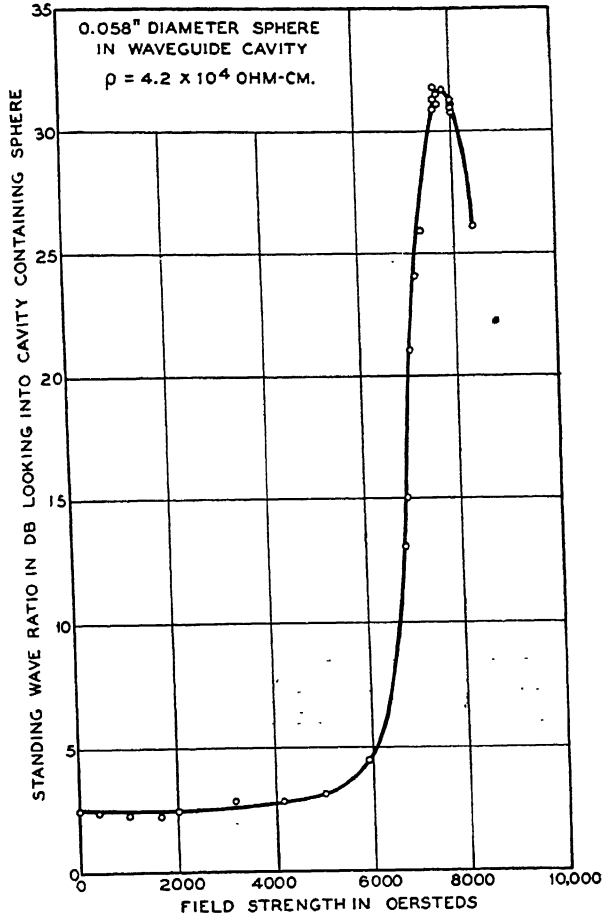


FIG. 13.—Experimental data obtained in one of Hewitt's resonance experiments. The standing wave ratio of a resonant cavity containing a $(MnO)(ZnO)\cdot 2Fe_2O_3$ sphere is plotted as a function of the normal field H_n; the frequency at maximum absorption is 23,730 Mc. (After Hewitt.[20])

In another resonance experiment Hewitt used a small *sphere* of a similar ferrite material. The actual data, reproduced in Fig. 13, illustrate the large value of the standing wave ratio observed at resonance. Kittel's theory predicts that in the special case of a *sphere* consisting of *low*-conductivity material, the correct resonance condition is eq. (27).

Using this equation, Hewitt found $g = 2.16$; since this value is quite close to the $g = 2.12$ obtained with the plane specimen, the measurements may be regarded as a confirmation of Kittel's theory.

No satisfactory explanation of the small, residual deviation of the observed g values from the value $g = 2$ has yet been reported.* It should be mentioned, on the other hand, that Griffiths[21] obtained $g = 2.00 \pm 0.02$ in recent experiments with very thin nickel specimens (< 0.5 micron thick); thicker samples, however, are reported to have yielded larger g values.

Some of Birks' investigations discussed in Section V, 2, as well as an experiment of Snoek,[52] indicate the possibility of ferromagnetic resonance absorption due to *internal* fields in semiconductors. The interpretation of these observations is based on the theory of Landau and Lifshitz (cf. Section IV, 3), but it has not yet progressed as far as that of the resonance experiments with an *external* field considered in the present section.

A preliminary report by Cooke[53] on the Hall and Kerr effects at microwave frequencies may well be mentioned at this point because his interesting observations may partly be due to ferromagnetic resonance phenomena in the metallic specimens used.

* * * * *

Most of the investigations considered in Section V are still in progress, and therefore, it is difficult to evaluate their over-all significance. However, it may be appropriate to make a few remarks concerning probable developments. The present research on *time-dependent* ferromagnetic processes will undoubtedly lead to substantial advances in our *general* knowledge of ferromagnetism, and its results are likely to be particularly important in connection with domain structure, domain behavior, and gyromagnetic effects. A better understanding of domain rotations, domain-wall displacements, and relaxation phenomena may well elucidate some of the problems of technical magnetization curves. It may not be premature, moreover, to expect that the results of current research will soon find applications in microwave apparatus.

* A preliminary communication dealing with this problem has just appeared; see Birks, J. B., *Phys. Rev.*, **74**, 988 (1948). According to Birks' note, the anomalous values of the g factor arise from the fact that effects due to internal anisotropy and/or strain fields have been neglected. Birks gives a formula for combining the internal and external fields; when used in conjunction with his new experimental data on several ferromagnetic compounds of the ferrite type, this relation leads to g values of 2.00 and 2.01. In regard to metallic samples, Birks suggests that the preparation of the thin specimens used by Yager and Bozorth[17] may have caused an artificial strain anisotropy. It should be recalled, however, that Yager's[18] results show no change in the apparent g value after cold working the specimen.

ADDENDUM

Frequency Dependence of the Permeability

Birks' measurements of the dispersion and absorption in Fe_3O_4 (cf. Section V, 2, particularly Fig. 9) were extended by Crouch[54] to a wavelength of 1.2 cm. Crouch suggests that the broadness of the resonance observed in this substance may be due to the effect of dipole interactions on the internal magnetic field which causes the spin precession. New results on the frequency dependence of the apparent permeabilities of metals (cf. Section IV, 2 and the beginning of Section V, 1, *b*) are now available. Smidt[55] found that μ_L in a fine iron wire decreases abruptly to unity at some wavelength between 60 and 70 cm.; the exact value of the critical wavelength depends on the temperature and the heat treatment. His results are discussed by van Leeuwen[56,57] who concludes that the theories of Becker and Kittel (cf. Sections IV, 3 and V, 1, *a*) fail to account for the type of frequency dependence observed. In similar experiments (described in part by van Leeuwen[57]) with a fine wire containing 35% nickel and 65% iron, Wieberdink[58] found a sharp dip at about 87 cm. in the wavelength dependence of a quantity related to μ_L; he states, moreover, that this phenomenon is strongly affected by the application of a static field of about 100 oersteds normal to the RF field. Kronig[59,60] suggests that these results may be due to the resonance absorption accompanying the Larmor precession of electron spins about the internal anisotropy field. It therefore appears that this type of resonance phenomenon may not be restricted to semiconductors (cf. the remarks at the end of Sections V, 2 and V, 3). Hodsman *et al.*[61] have now published a full account of their work (cf. the beginning of Section V, 1, *b*) on the frequency dependence of μ_R in metals. Interpretation of their results by Becker's theory leads to domain sizes of about 10^{-4} cm. In emphasizing the limitations of the Becker and Kittel theories, Hodsman *et al.*,[61] as well as van Leeuwen,[56] advance essentially the same arguments as those given in Sections IV, 3 and V, 1, *a*. Two new theoretical papers should also be mentioned. Döring[62] generalized Becker's theory (cf. Section IV, 3) of domain wall displacements at high frequencies by taking into account the apparent inertia of the moving walls. This effect is basically due to the spin angular momentum of the electrons constituting the walls and leads to a mass term in the equation of motion. Thus the susceptibility due to each type of wall is determined by a damped resonance of the form of eq. (26) rather than by a relaxation mechanism described by eq. (19). According to Döring's calculations the resonance frequency in iron is smaller than the relaxation frequency due to eddy current damping. In principle, therefore, this theory may

account for dispersion and absorption in semiconductors as well as in metals. Akhieser[63] has published a quantum-mechanical theory of the relaxation time in ferro-magnetic substances at low temperatures.

Ferromagnetic Resonance in an External Field

A number of additional ferromagnetic resonance experiments[64–69] (cf. Section V, 3) have been reported. The resulting g-values exceed the value $g = 2$ by 3–12% in the following substances: Supermalloy (static field parallel,[64] and perpendicular,[65] to plane of specimen), nickel-zinc ferrite[66] (sphere), iron[67] (single crystal) and magnetite[68] (single crystal). In the last two cases the anisotropy constant K was also determined, and in magnetite the dependence of g and K on temperature was studied. A particularly interesting result is the value[69] $g = 2.01$ found in a Heusler alloy where the g-value derived from gyromagnetic experiments is 2.00. With regard to the theory, Luttinger and Kittel[70] have given a quantum-mechanical derivation of Kittel's resonance formula (cf. Section V, 3; eq. (28)), and Richardson[71] subsequently derived the same result by a simpler method. Polder[72] has published a generalization of Kittel's classical theory of resonance, and Polder[72] and Kittel[73] have given quantum-mechanical calculations which suggest that the anomalous g-values found in ferromagnetic resonance experiments may be explained by the combined effect of a crystalline field and spin-orbit interaction. According to Polder,[72] furthermore, exchange forces do not usually affect the resonance experiments. Similar considerations (resulting in non-rigid domains) suggest that any supposed connection (cf. refs. 9, 10 and Section V) between skin depth and domain size is highly questionable.

General

Reference should be made to two interesting papers by Williams *et al.*[74] on ferromagnetic domain patterns, and to Morgan's[75] calculations on the effect of surface roughness on the resistance measured at high microwave frequencies.

REFERENCES

1. Cf., for example, Slater, J. C. Microwave Transmission. McGraw-Hill, New York, 1942; also Stratton, J. A. Electromagnetic Theory. McGraw-Hill, New York, 1941.
2. Cf., for example, Slater, J. C. *Rev. Mod. Phys.*, **18**, 441–512 (1946).
3. Hagen, E., and Rubens, H. *Ann. Phys. Lpz.*, [IV], **11**, 873–901 (1903).
4. Arkadiew, W. *J. Phys., USSR*, **9**, 373–378 (1945). This paper contains a summary (with complete references) of Arkadiew's numerous investigations.
5. For a review, see Allanson, J. T. *J. Instn Elect. Engrs.*, **92** [pt. III], 247–255 (1945).

6. Becker, R., and Döring, W. Ferromagnetismus. Springer, Berlin, 1939; Photo-Litho-print Reproduction by Edwards Brothers, Inc., Ann Arbor, 1943. Kreielsheimer's work is discussed in Section 18c.
7. Landau, L., and Lifshitz, E. *Phys. Z.*, **8**, 153–169 (1935).
8. Becker, R., and Döring, W. Ferromagnetismus. Springer, Berlin, 1939, pp. 237–242. (See Ref. 6.)
9. Kittel, C. *Phys. Rev.*, **70**, 281–290 (1946).
10. Johnson, M. H., and Rado, G. T. *Phys. Rev.*, **75**, 841–864 (1949). For preliminary communications (by these authors and M. Maloof) see *Phys. Rev.*, **71**, 472 [Abstracts L4 and L5] (1947); *ibid.*, **71**, 322 (1947); *ibid.*, **72**, 173 (1947).
11. Rado, G. T., Maloof, M., and Rosenblatt, H. *Phys. Rev.*, **74**, 1206 (1948); paper presented at the April, 1948, meeting of the American Physical Society in Washington, D.C.
12. Birks, J. B. *Proc. Phys. Soc., Lond.*, **60**, 282–292 (1948).
13. Birks, J. B. *Nature, Lond.*, **158**, 671–2 (1946); *ibid.*, **159**, 775–6 (1947); *ibid.*, **160**, 535 (1947).
14. Griffiths, J. H. E. *Nature, Lond.*, **158**, 670–1 (1946).
15. Kittel, C. *Phys. Rev.*, **71**, 270–1 (1947).
16. Kittel, C. *Phys. Rev.*, **73**, 155–161 (1948).
17. Yager, W. A., and Bozorth, R. M. *Phys. Rev.*, **72**, 80–1 (1947).
18. Yager, W. A. *Phys. Rev.*, **73**, 1247 (1948).
19. Kip, A. F., and Arnold, R. D. *Phys. Rev.*, **73**, 1247 (1948).
20. Hewitt, W. H., Jr. *Phys. Rev.*, **73**, 1118–9 (1948).
21. Griffiths, J. H. E. Program of the Oxford Conference of the Physical Society, July 23 and 24, 1948.
22. Becker, R., and Döring, W. Ferromagnetismus. Springer, Berlin, 1939, p. 223. (See Ref. 6.)
23. Becker, R., and Döring, W. Ferromagnetismus. Springer, Berlin, 1939, Section 19. (See Ref. 6.)
24. Snoek, J. L. New Developments in Ferromagnetic Materials. Elsevier, New York, 1947. Also: *Philips Tech. Rev.*, **8**, 353–360 (1946).
25. Bitter, F. Introduction to Ferromagnetism. McGraw-Hill, New York, 1937.
26. Stoner, E. C. Magnetism and Matter. Methuen and Company, London, 1934.
27. Van Vleck, J. H. Electric and Magnetic Susceptibilities. Oxford Univ. Press, New York, 1932.
28. Polder, D. *Phys. Rev.*, **73**, 1120–1 (1948).
29. Debye, P. Polar Molecules. Dover Publications, New York, 1945.
30. Epelboim, I. *J. Phys. Radium*, [VIII], **8**, 251–6 (1947).
31. Zottu, P. D. *Union Radio Scientifique Internationale*, **5**, [pt. 1], 190–204 (1938).
32. Glathart, J. L. *Phys. Rev.*, **55**, 833–8 (1939).
33. Maxwell, E. *J. Appl. Phys.*, **18**, 629–638 (1947).
34. Gorter, C. J. Paramagnetic Relaxation. Elsevier, New York, 1947.
35. For a review, see Van Vleck, J. H. *Rev. Mod. Phys.*, **17**, 27–47 (1945).
36. Brown, W. F., Jr. *Rev. Mod. Phys.*, **17**, 15–9 (1945).
37. Hughes, D. J., Wallace, J. R., and Holtzman, R. H. *Phys. Rev.*, **73**, 1277–1290 (1948).
38. Marton, L. *Phys. Rev.*, **73**, 1475 (1948); cf. also *J. Appl. Phys.*, **19**, 687–8 (1948); *ibid.*, **19**, 863–4 (1948).
39. Néel, L. *J. Phys. Radium*, [VIII], **5**, 241–251 (1944); *ibid.*, [VIII], **5**, 265–276 (1944). See also Lifshitz, E. *J. Phys., USSR*, **8**, 337–346 (1944).

40. Williams, H. J. *Phys. Rev.*, **70**, 106 (1946); *ibid.*, **71**, 646–7 (1947); *ibid.*, **72**, 529 (1947). Williams, H. J., and Bozorth, R. M. *ibid.*, **73**, 1246–7 (1948).
41. See, for example, Pauling, L., and Wilson, E. B., Jr. Introduction to Quantum Mechanics. McGraw-Hill, New York, 1935.
42. For references, see Van Vleck, J. H. *Rev. Mod. Phys.*, **17**, 27–47 (1945).
43. Becker, R., and Döring, W. Ferromagnetismus. Springer, Berlin, 1939, particularly Sections 11, 12, 14, 15, and 16. (See Ref. 6.)
44. Kersten, M. Grundlagen einer Theorie der ferromagnetischen Hysterese und der Koerzitivkraft, S. Hirzel, Leipzig, 1943. Lithoprinted by Edwards Brothers, Inc., Ann Arbor, 1946.
45. Gans, R. *Ann. Phys., Lpz.*, [V], **15**, 28–44 (1932).
46. Lawton, H., and Stewart, K. H. *Proc. Roy. Soc.*, **A193**, 72–88 (1948).
47. Birks, J. B. *Phys. Rev.*, **74**, 843–4 (1948).
48. Eichholz, G., and Hodsman, G. F. *Nature, Lond.*, **160**, 302–3 (1947).
49. Elmen, G. W. *Bell Syst. Tech. J.*, **15**, 113–135 (1936).
50. Birks, J. B. Private communication, July, 1948; the author wishes to thank Dr. Birks for communicating this result prior to publication, and for sending the original drawings used for Figs. 8, 9, and 10 of the present article.
51. Barnett, S. J. *Amer. J. Phys.*, **16**, 140–7 (1948); this review of gyromagnetic investigations contains a summary of recent results as well as many references to the literature.
52. Snoek, J. L. *Nature, Lond.*, **160**, 90 (1947); *Physica*, **14**, 207–217 (1948).
53. Cooke, S. P. *Phys. Rev.*, **74**, 701–2 (1948).
54. Crouch, G. E., Jr. *Phys. Rev.*, **75**, 525–6 (1949).
55. Smidt, J. *Appl. Sci. Res.*, **B1**, 127–134 (1948).
56. van Leeuwen, H. J. *Appl. Sci. Res.*, **B1**, 135–138 (1948).
57. van Leeuwen, H. J. *Physica*, **15**, 258–263 (1949).
58. Wieberdink, A. *Nature, Lond.*, **162**, 527–528 (1948).
59. Kronig, R. *Nature, Lond.*, **162**, 528 (1948).
60. Kronig, R. *Physica*, **15**, 264–5 (1949).
61. Hodsman, G. F., Eichholz, G., and Millership, R. *Proc. Phys. Soc. Lond.*, **B62**, 377–390 (1949).
62. Döring, W. *Z. Naturforsch.*, **3a**, 374–9 (1948).
63. Akhieser, A. *J. Phys., USSR*, **10**, 217–230 (1946).
64. Yager, W. A. *Phys. Rev.*, **75**, 316–7 (1949).
65. Kittel, C., Yager, W. A., and Merritt, F. R. *Physica*, **15**, 256–7 (1949).
66. Beljers, H. G. *Physica*, **14**, 629–641 (1949).
67. Kip, A. F., and Arnold, R. D. *Phys. Rev.*, **75**, 1556–1560 (1949). For a preliminary communication, cf. Ref. 19.
68. Bickford, L. R., Jr. *Phys. Rev.*, **76**, 137–8 (1949).
69. Yager, W. A., and Merritt, F. R. *Phys. Rev.*, **75**, 318 (1949).
70. Luttinger, J. M., and Kittel, C. *Helv. Phys. Acta*, **21**, 480–2 (1948).
71. Richardson, M. *Phys. Rev.*, **75**, 1630–1 (1949).
72. Polder, D. *Phil. Mag.*, **40**, 99–115 (1949). For preliminary communications, cf. Ref. 28 and *Phys. Rev.*, **73**, 1116 (1948). A brief summary of Polder's work has appeared in *Physica*, **15**, 253–5 (1949).
73. Kittel, C. *Phys. Rev.*, **76**, 176 (1949).
74. Williams, H. J., Bozorth, R. M., and Shockley, W. *Phys. Rev.*, **75**, 155–178 (1949). Williams, H. J., and Shockley, W. *ibid.*, **75**, 178–183 (1949). For preliminary communications, cf. Ref. 40.
75. Morgan, S. P., Jr. *J. Appl. Phys.*, **20**, 352–362 (1949).

Microwave Spectroscopy

DONALD K. COLES

Westinghouse Research Laboratories, East Pittsburgh, Pennsylvania

CONTENTS

I. Introduction

The war time development of microwave techniques has opened the way to an entirely new branch of spectroscopy. In microwave absorption spectroscopy, radiation of known frequency is passed through the material to be studied, and absorption of the radiation is detected by a reduction in the intensity of the transmitted radiation.* As the frequency of the radiation is slowly varied, the attenuation at each frequency is noted. In this article we shall be concerned primarily with the absorption of microwaves by gases.

In the case of a gas at low pressure, it is found that absorption takes place only at very sharply defined frequencies. Most commonly these are rotational frequencies of the gas molecules. The distinctness of these rotational frequencies is intimately associated with the quantum theory of matter and energy.

1. Quantized Rotation of Molecules

According to classical statistics, gas molecules at room temperature should possess all gradations of translational and rotational energy. The distribution of rotational frequencies depends on the molecular moments of inertia and the temperature of the gas. For a diatomic molecule the most probable frequency of rotation is given by

$$\nu_{\text{prob}} = \left[\frac{kT}{4\pi^2 I}\right]^{\frac{1}{2}}$$

where k is the Boltzmann constant,

T is the absolute temperature,

I is the moment of inertia of the molecule.

For a molecule of bromine chloride at room temperature, the value of ν_{prob} is 3×10^{11} revolutions/second. Electric dipoles rotating at these speeds should emit and absorb electromagnetic waves of all wavelengths in the neighborhood of 1 mm. Moreover, as Lord Rayleigh pointed out in 1892, the relatively low-frequency rotation of molecules should "modulate" their high-frequency absorption or emission, and the resulting side bands should have the effect of broadening a high frequency spectral line, such as is emitted or absorbed in the ultraviolet or near infrared regions.

* We shall not discuss in this article the powerful methods of molecular beam spectroscopy in which the absorption or emission of radiation is detected by a change in the molecular beam intensity. These methods are surveyed by Estermann and by Kellogg and Millman in the 1946 *Reviews of Modern Physics*.

Quantum statistics, first used by Planck in 1900 to explain the frequency distribution of black body radiation, led to suggestions that the rotational frequency of a molecule should assume only discrete values. It was then pointed out by Bjerrum, in 1912, that a compound in the gaseous state should possess not a continuous rotation spectrum, but one consisting of discrete lines. Furthermore, Bjerrum theorized that a molecular absorption band in the near infrared or ultraviolet regions should appear not as a single strong line, but as a whole series of fine lines. The frequency shifts of these fine lines from the band center should be equal to the rotational frequencies of the molecule.

In 1913 this quantum theory of molecules was confirmed by the infrared studies of Eva v. Bahr. By improving the resolution of her infrared spectrograph, she found that the near-infrared absorption band of HCl gas could be resolved into a series of fine lines. The absorption band is associated with the vibration of the hydrogen atom against the chlorine atom in the HCl molecule, so that it is known as a rotation-vibration band. The *pure* rotation spectrum of HCl, in the far infrared, has also been observed to consist of a series of discrete lines, although this spectrum was not resolved until somewhat later. It has been found experimentally that for any diatomic molecule, the frequencies of rotational absorption lines may be represented approximately by the equation:

$$\nu = \frac{nh}{4\pi^2 I} \tag{1}$$

where $n = 1, 2, 3, \cdots$,

h = Planck's quantum of action,

I is the moment of inertia of the molecule.

This relationship may be interpreted in the light of Bohr's quantum theory, which was announced the same year (1913). According to Bohr's theory a spectral line emitted or absorbed is associated with a transition between two stationary states, the frequency of the line being related to the energy difference of the stationary states by the equation

$$h\nu = W_2 - W_1 \tag{2}$$

where h is Planck's constant $= 6.624 \times 10^{-27}$ erg-sec.

The rotational energy of a rigid molecule in a stationary state depends on its moments of inertia I_A, I_B, I_C, and on certain rotational quantum numbers (integers). For example, the laws of quantum mechanics give for the rotational energy of a rigid symmetric top[1]

$$W_{J,K} = BJ(J+1) + (C - B)K^2 \tag{3}$$

where

$$C = \frac{h^2}{8\pi^2 I_C},$$

I_C = moment of intertia of symmetric top about its axis of symmetry,

$$B = \frac{h^2}{8\pi^2 I_B},$$

I_B = moment of inertia of symmetric top about an axis through its center of gravity, perpendicular to its axis of symmetry.

$J = 0, 1, 2, 3, \cdots$

$K = -J, -(J-1) \cdots 1, 0, 1, \cdots + (J-1), J$

$J\frac{h}{2\pi}$ is a measure of the total angular momentum of the top, and $\frac{Kh}{2\pi}$ is the component of total angular momentum parallel to the axis of symmetry of the top.

The selection rules for transitions between energy levels, as derived by quantum mechanics, are

$$J' = J, J \pm 1,$$
$$K' = K, K \pm 1,$$

where J, K are quantum numbers associated with the initial rotational level; J', K' are quantum numbers associated with the final rotational level.

In the case of a *rotational absorption* line we observe the transition $J, K \rightarrow J', K'$, where

$$\begin{aligned} J' &= J + 1, \\ K' &= K. \end{aligned} \tag{4}$$

By substituting (3) and (4) in (2) we obtain

$$\nu_J^{J+1} = (J+1) \cdot 2B/h \tag{5}$$

This result agrees with eq. (1).

It is found experimentally that molecules which possess an axis of symmetry, such as PH_3 or $HCCl_3$, can be treated as almost rigid symmetric tops. The effect of centrifugal distortion of symmetric-top molecules is usually small, at least for J values corresponding to microwave absorption lines. However, at room temperature, a fraction of the molecules may be in excited vibrational states, and the moment of inertia of a molecule in an excited vibrational state is not quite the same as that of a molecule in the ground state. The effect of centrifugal distortion in molecules and the calculation of vibration frequencies is discussed in Section III, 1.

Linear molecules, in which all the atoms lie on a single straight line, may be classified as a special class of symmetric-top molecules. Because of the low moment of inertia about the axis of symmetry for these molecules, the rotational angular momentum about the symmetry axis is always equal to zero. The quantum number K is therefore equal to zero for linear molecules. However, if the molecule is in an excited vibrational state it may possess a vibrational angular momentum about the axis of symmetry.

Most molecules do not possess an axis of symmetry, so that their three moments of inertia are, in general, all different. For these molecules the rotational frequencies depend on all three moments of inertia, and the rotational spectrum generally has a rather irregular appearance. Examples of the two different types of absorption spectra are shown in

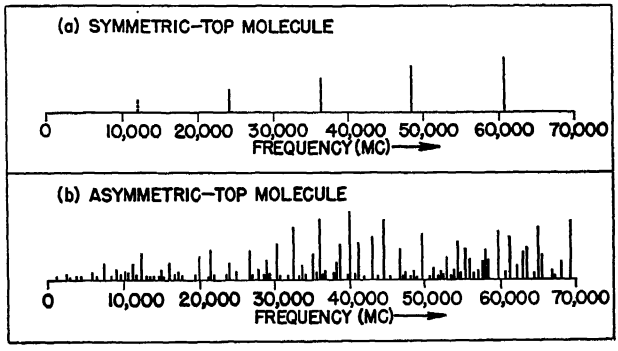


FIG. 1.—Two types of molecular rotation spectra.

Fig. 1. The upper spectrum is that of carbon-oxy-sulfide. (O=C=S.) The three atoms of this molecule are arranged in a straight line, with the carbon atom in the middle. The four absorption lines shown solid have already been observed.

In the case of asymmetric-top molecules, the angular momentum associated with vibrations is usually not large, but the effect of centrifugal distortion of the molecule may be rather serious. Semi-rigid molecules are treated rather extensively in the excellent book on infrared and Raman spectroscopy by Herzberg.[1]

2. *Relationship of Microwave Spectroscopy to Other Branches of Spectroscopy*

Because of the very weak intensities of infrared sources in the far infrared region, where the *pure* rotation spectra of molecules are found, it has so far been possible to study the pure rotation spectra of only about

half a dozen molecules. These are molecules consisting of a single heavy nucleus bonded to one or more hydrogen atoms, which therefore have small moments of inertia.

It has also been possible to measure the moments of inertia for about one hundred of the lighter molecules by resolving the fine structure of their spectra in the near infrared, visible, and ultraviolet regions. This fine structure, as mentioned above, is associated with molecular rotation. In infrared and ultraviolet spectroscopy the experimental inaccuracy of a careful rotational frequency determination is approximately 5 Mc. In microwave spectroscopy the experimental inaccuracy of a careful rotational frequency measurement is less than .01 Mc.

We stress the importance of frequency measurements rather than wavelength measurements because of the relation between frequency and energy, expressed in eq. (2).

In order to avoid the necessity of constantly multiplying cycles per second by Planck's constant to obtain energies, it is often convenient to omit the constants of proportionality and to think of the megacycle per second (Mc) as a unit of energy.

$$1 \text{ Mc} = 6.6242 \times 10^{-21} \text{ ergs} = 4.135 \times 10^{-9} \text{ electron volts.}$$

Since the microwave frequencies of greatest interest are in the range of tens of thousands of megacycles, it is sometimes convenient to use as a unit of frequency the kilomegacycle (1 kmc = 10^9 cycles/sec.). Another unit, used occasionally in spectroscopy of the visible and ultraviolet, is the fresnel (1F = 10^{12} cycles/sec.). A "frequency" unit which is more commonly used in other branches of spectroscopy is the number of waves per centimeter in a vacuum. This unit, called the *reciprocal centimeter* is also used as a convenient unit of energy (1 cm.$^{-1}$ = hc ergs). For the relationship between the three convenient frequency units, we have

$$1 \text{ kmc} = .033358 \text{ cm.}^{-1} = .001 \text{ fresnel}$$
$$1 \text{ cm.}^{-1} = 29.978 \text{ kmc} = .029978 \text{ fresnel}$$

The relationship of the rotational region of molecular spectra to the infrared and ultraviolet regions is shown in Fig. 2. On this frequency scale the visible region falls in the range 375 — 750 fresnel. Molecular absorption frequencies associated with the electronic transitions appear predominantly in the ultraviolet region. The interatomic vibrational frequencies usually lie in the near infrared.

The rotational frequencies of molecules fall in a rather inaccessible part of the electromagnetic spectrum. A few of the lightest molecules

have rotational absorption lines at the extreme low frequency end of the infrared region, which are strong enough to be detected. Rotational frequencies of the heavier molecules are too low to be detected by conventional infrared methods, but some of these frequencies are low enough to fall in the accessible microwave region.

The rotational spectrum of methyl chloride is shown in Figure 2 as an example of a spectrum that contains a few faint lines at the low frequency end of the infrared region, in addition to observable lines in the microwave region below 100 kmc. The near infrared (vibrational) spectrum is also shown, as is the ultraviolet (electronic) spectrum so far as it is known. Because of the logarithmic frequency scale used, it is, of course, not possible in Fig. 2 to see the rotational fine structure in the near infrared and ultraviolet regions, although it has been resolved experimentally.

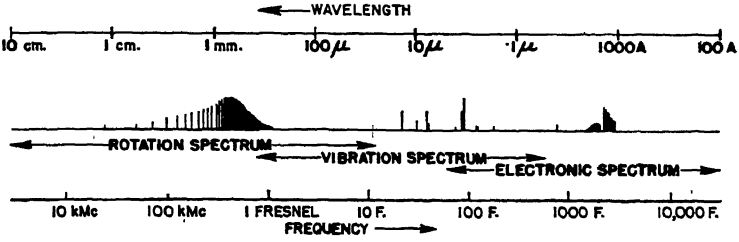


FIG. 2.—Absorption spectrum of CH_3Cl. The electronic absorption occurs in the ultraviolet region, the interatomic vibration occurs in the near infrared, and the pure rotational absorption overlaps the far infra-red and the microwave regions. The rotational structure superimposed on the vibrational and electronic lines cannot be shown because of the logarithmic frequency scale used above.

It should be emphasized that molecules will not show a pure rotation spectrum unless they have a permanent electric or magnetic moment. Molecules with magnetic dipole moments (such as the O_2 molecule) are rare, and these absorb very weakly. The great majority of molecules have electric dipole moments, but in the case of hydrocarbons, for example, the dipole moments are very small. A symmetrical molecule such as CO_2 will have no permanent dipole moment at all. However, if the molecule is vibrating in an unsymmetrical manner, it will have a vibrating dipole, which allows it to absorb in the near infrared region.

In the case of a symmetrical diatomic molecule such as H_2, even an interatomic vibration cannot produce an electric dipole. Such a molecule will, therefore, not absorb appreciably in the infrared region, although it does exhibit absorption in the ultraviolet region. Ultraviolet absorption is to be associated with an electronic transition to an upper electronic state, which is unsymmetrical in the electronic distribution.

II. The Beginning of Microwave Spectroscopy

1. Investigations at the University of Michigan

The first attempt to use electronic oscillators as sources in molecular spectroscopy was made in 1933 at the University of Michigan. The group of infrared spectroscopists at the University of Michigan had discovered that ammonia possessed a very unusual absorption spectrum. Professor Dennison and associates at the University of Michigan interpreted the data to mean that the ammonia molecule was able to turn itself inside out. They predicted that this remarkable property of ammonia would give rise to a strong absorption line at a frequency of about 20 kmc, considerably stronger than any rotational absorption line at this low frequency. It would be impossible to detect absorption at this frequency by using conventional infrared sources and receivers.

Accordingly, in 1933, Cleeton and Williams[2] designed and built a set of magnetron oscillator tubes covering the region 7 kmc to 28 kmc. Radiation from one of these sources was passed through 40 cm. of ammonia gas at atmospheric pressure, and was then detected by an iron pyrites crystal detector. The frequency at which the magnetron was oscillating was determined by means of a large diffraction grating.

Cleeton and Williams succeeded in proving that ammonia did indeed possess a broad absorption band at about 1.25 cm. wavelength. In 1939 this work at the University of Michigan was taken up again by H. S. Howe,[3] who studied the ammonia absorption band at reduced pressures. At the lowest pressure, $\frac{1}{4}$ atmosphere, the line had become much narrower.

One of the great advantages that microwave spectroscopy enjoys over other types of absorption spectroscopy is that it is not necessary to use a monochromator. This is because the source emits only one frequency at a time. Instead of using a diffraction grating, Howe used a cavity wavemeter to determine the wavelength. He also used large cylindrical waveguide.

So far as the writer is aware, these were the only attempts at microwave spectroscopy prior to the second world war. During the war, the efforts of hundreds of physicists and radio engineers were concentrated on the production and propagation of very high frequency radiation.

2. Wartime Investigations of Attenuation by H_2O and O_2

Theoretical calculations on the attenuation of radar beams by the oxygen and water vapor of the earth's atmosphere were carried out by Van Vleck in 1942.[4,5] He pointed out that, although nitrogen and oxygen molecules do not possess permanent electric dipole moments, the oxygen molecule possesses a *magnetic* moment, which makes possible

a weak absorption in the microwave region. The water vapor molecule has a permanent electric dipole moment, and one of its rotational lines falls in the neighborhood of the radar K-band. At atmospheric pressure the water vapor line and the oxygen absorption band are so broad that they overlap each other somewhat.

Measurements on the absorption of oxygen were made in 1944 by Beringer,[6] using the microwave frequency bridge shown in Fig. 3. In order to study the O_2 absorption band at 60 kmc he used a 30 kmc oscillator and a crystal frequency-doubler. The microwaves traversed 6 meters of waveguide filled with oxygen or a mixture of oxygen and nitrogen at pressures down to a few centimeters of mercury.

The absorption due to water vapor was estimated by measuring the intensity of radar signals transmitted through the earth's atmosphere on

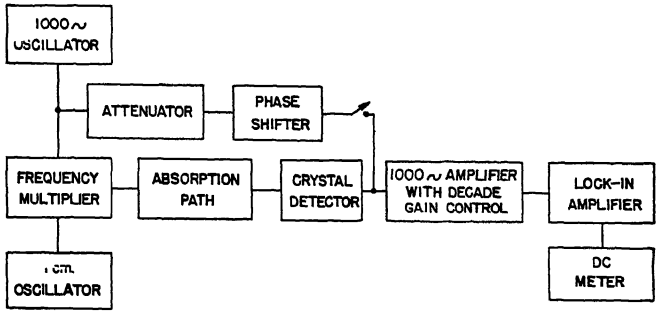


FIG. 3.—Microwave frequency bridge. Apparatus used by Beringer[8] to study the absorption band of O_2 in the vicinity of $\frac{1}{2}$ cm. A 1-cm. oscillator is used, and the 5 mm. harmonic is generated in a crystal frequency multiplier. The crystal also serves to modulate the intensity of the 5 mm. waves at a frequency of 1000 cps.

days of high and low humidity. In order to check these results, Becker and Autler[7] measured the attenuation of water vapor in a large untuned cavity that was a cube 8 feet on an edge. Many different modes of the cavity were excited simultaneously, and the average energy density was measured by means of a large number of thermocouples scattered throughout the cavity. Since the large cavity used in this investigation could not be conveniently evacuated, attenuation measurements were made at atmospheric pressure.

3. The First High-Resolution Microwave Spectroscopy

The end of the war marked the beginning of high resolution microwave spectroscopy. The first really sharp-line absorption spectrum was obtained by Bleaney and Penrose[8] at Oxford University, England. These investigators introduced ammonia into a resonant cavity, which

effectively is a short length of waveguide in which the radiation is reflected back and forth thousands of times. The effect of the ammonia gas was to damp out the waves and thus to reduce the energy density in the cavity. The change in energy density could be measured by means of a crystal detector.

Bleaney and Penrose found that at low pressures (below 1 mm. of mercury) the broad absorption band of ammonia consisted of several series of very sharp lines. This was in general agreement with the prediction of Dennison and others at the University of Michigan. When these results were announced, early in 1946, similar results were already being obtained by W. E. Good[9] at the Westinghouse Research Laboratories and C. H. Townes[10] at the Bell Telephone Laboratories.

Other molecules were investigated immediately, so that by the end of 1946 sharp absorption lines had been found for some six different compounds. Absorption data on some of these compounds had been obtained earlier by W. D. Hershberger at the Radio Corporation of America Laboratories, who made attenuation measurements at fixed frequencies, but at many different pressures.

TABLE I. Some rotational parameters determined by means of microwave spectroscopy

Linear molecules	B_0(Kmc)	Symmetric top molecules	B_0(Kmc)
$Br^{79}CN$	4.120 190	AsF_3	5.833 0
$Cl^{35}CN$	5.970 821	$B^{11}H_3CO$	8.657 21
HCN	44.336	CF_3CH_3	5.185
ICl^{35}	3.422 300	CH_3Br^{79}	9.568 100
ICN	3.225 527	CH_3Cl^{35}	13.292 89
N_2O	12.561 64	CH_3CN	9.198 45
OCS^{32}	6.081 453	CH_3F	25.530
$OCSe^{80}$	4.017 677	CH_3I	7.501 250

Asymmetric-top molecules	A_0(Kmc)	B_0(Kmc)	C_0(Kmc)
C_2H_4O	25.484	22.121	14.098
SO_2	61.10	10.40	8.870
CH_2CF_2	11.0018	10.4285	5.3456

By the end of 1947 more than twenty different compounds had been investigated rather intensively, by workers at the above-mentioned laboratories, and at the Massachusetts Institute of Technology, Harvard, and Duke University, and at an increasing number of other laboratories.

An excellent review of the accomplishments of microwave spectroscopy up to the middle of 1948 has been given by Gordy.[11]

A short list of rotational constants determined by means of microwave spectroscopy is given in Table I, by way of illustration. A much more complete list with references, including many different isotopic species of molecule, and molecules in excited vibrational states, has been compiled by Kisliuk and Townes.[12] This listing includes an interpretation of every observed line, if known, and an estimate of its intensity.

4. Microwave Spectroscopy Applied to Atoms

The methods of microwave spectroscopy allow determinations not only of low-lying molecular energy levels, but also of low-lying atomic energy levels, usually due to noncoulombic interaction between the atomic nucleus and the surrounding electrons.

Transitions between these low-lying atomic energy levels are associated with magnetic dipole moments, rather than electric dipole moments, hence the interaction with radiation of microwave frequency is weak. Such energy levels are affected by relatively weak magnetic fields. Roberts *et al.* took advantage of this fact by using a variable magnetic field to frequency modulate the absorption lines.[13] In this manner they were able to study the hyperfine structure of the cesium atom. The cesium vapor was contained in a resonant cavity, which was excited by 3 cm. microwaves. By imposing a steady magnetic field on the cavity it was possible to make an independent determination of the nuclear spin of cesium and of the interaction between the magnetic moment of the cesium nucleus and that of the valence electron in the neutral cesium atom.

5. Stark Effect for Molecules

When a molecule is placed in an electric field, the component of its total angular momentum in the field direction is quantized, being equal to $M \frac{h}{2\pi}$ where $-J \leq M \leq J$. Due to the interaction between the electric field and the dipole moment of the molecule, the original energy level is split up into $2J + 1$ sublevels.

The effect of an electric field in breaking up an energy level into several components is called the Stark effect. Generally, the frequency shift of a rotational level component from the original level is small compared with the frequency separation between the original level and its nearest neighbors. In this case, the Stark shift is approximately proportional to $E^2(A + BM^2)$, so that the level breaks up into only $J + 1$ separate components.

The magnitude of this quadratic Stark effect depends on the electric dipole moments of the molecule parallel to its three inertial axes, and on the frequency separation between the rotational level under consideration and its nearest neighboring levels. The Stark effect for an asymmetric top was first treated by W. G. Penny. His treatment has been extended by S. Golden and E. B. Wilson, Jr.,[14] who have also provided tables for approximating rotational levels up to and including $J = 2$. They give general rules facilitating the identification of spectral lines.

For symmetric top molecules the first and second order Stark effect has been calculated by R. de L. Kronig and others.[15] For a molecule with an electric dipole moment μ in an electric field E, the change of energy is

$$W_{JKM} = -E\mu \frac{KM}{J(J+1)} + \frac{E^2\mu^2}{2B}\left[\frac{\left\{\frac{3K^2}{J(J+1)} - 1\right\}\left\{\frac{3M^2}{J(J+1)} - 1\right\}}{(2J-1)(2J+3)} - \frac{M^2K^2}{J^3(J+1)^3}\right] \quad (6)$$

with $W_{000} = -\frac{1}{6}\frac{E^2\mu^2}{B}$.

If E is expressed in volts per centimeter and μ is expressed in Debye units, then the product should be multiplied by .5032 to obtain $(E\mu)$ in Mc.

At low electric fields, the first order term on the right hand side of eq. (6) corresponds to the classical expression $W = -E(\mu_z)$. The second order term is due to distortion of the orbital motion of the molecule by the electric field.

In the waveguide method of microwave spectroscopy, it is most convenient to study the Stark effect when the electric vector of the microwave is parallel to the constant electric field on the gas. Under these conditions the selection rule for M during a transition is

$$M' = M \quad (7)$$

If the constant electric field has a component perpendicular to the electric vector of the microwave, Stark components are observed corresponding to

$$M' = M \pm 1. \quad (8)$$

The Stark effect of OCS and of ammonia were first studied by Dakin *et al.*[16] and by Coles and Good.[17] The effect of an electric field on the lowest rotational levels of a linear molecule is indicated in Fig. 4. The

case of a linear molecule in an excited vibrational state has been discussed by Strandberg *et al.*[15]

6. *Zeeman Effect for Molecules*

The large Zeeman effect, as usually observed for atoms, or molecules in excited electronic states, is a change in energy level due to the effect of a magnetic field on the electronic magnetic moments of the atom or molecule. This *large* Zeeman effect is usually not present in molecules in the lowest electronic state, since most molecules possess an even number of electrons, and the magnetic moments due to these electrons tend to cancel each other. However, a rotating molecule may possess a small

$$W_{JM} = BJ(J+1) + \frac{\mu^2 E^2 \left[1 - \frac{3M^2}{J(J+1)}\right]}{2B(2J-1)(2J+3)}$$

$$W_{00} = BJ(J+1) - \frac{1}{6}\frac{\mu^2 E^2}{B}$$

$$(-J \leq M \leq J)$$

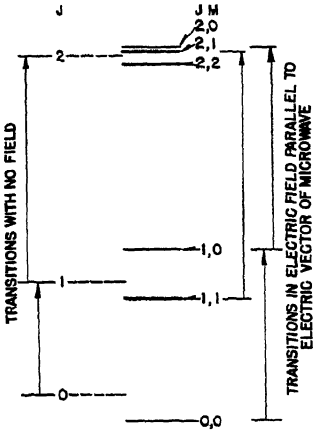


FIG. 4.—Effect of an electric field on the energy levels of a linear molecule. Transitions giving rise to absorption lines are indicated by vertical arrows.

unbalanced magnetic dipole moment, due to its rotation. The interaction of this dipole with a magnetic field was first observed by Coles and Good[17] for ammonia, and also for OCS and H_2O. More accurate measurements on NH_3 and on CH_3Cl have been made by Jen,[18] who also detected effects due to the interaction of the nitrogen and chlorine nuclei with the magnetic field. The magnitude of the magnetic dipole moment which is produced by molecular rotation has not been calculated theoretically, but it appears certain that its average component in any given direction is proportional to the component of the molecular angular momentum in that direction,

$$(\mu_z)_{av} = g_J M. \tag{9}$$

μ_z is the component of the magnetic dipole moment parallel to the field,

g_{J_i} is the gyromagnetic ratio of the molecule in a particular rotational level (i.e., the ratio of magnetic moment to angular momentum),
i is an inner quantum number,
M is the outer quantum number.

As in the Stark effect, when the electric vector of the microwave is parallel to the magnetic field, the outer quantum number does not change. When the electric vector of the microwave is perpendicular to the magnetic field we have $M' = M \pm 1$.

The frequency shift $\Delta\nu$ of a Zeeman component of a rotational level is obtained from the relation

$$h(\Delta\nu) = -H(\mu_z)_{av} \tag{10}$$

The Zeeman frequency shift of a rotational absorption line is thus

$$\begin{aligned}
\nu_{J,i,M}^{J',i',M+1} &= -H\frac{e}{4\pi mc}\left[M(g_{J',i'} - g_{J,i}) + g_{J',i'}\right] \text{ (perp.)} \\
\nu_{J,i,M}^{J',i',M} &= -H\frac{e}{4\pi mc}M(g_{J',i'} - g_{J,i}) \cdots \text{ (parallel)} \\
\nu_{J,i,M}^{J',i',M-1} &= -H\frac{e}{4\pi mc}\left[M(g_{J',i'} - g_{J,i}) - g_{J',i'}\right] \text{ (perp.)}
\end{aligned} \tag{11}$$

Where g_{J_i} is defined as the ratio of magnetic moment in nuclear magnetons $\frac{eh}{4\pi mc}$ to the angular momentum in units of $\frac{h}{2\pi}$,

m is the mass of the proton,
c is the velocity of light,

$$\frac{e}{4\pi mc} = 762.2 \text{ emu/g.}$$

This value is obtained from the Faraday in emu and the atomic weight of hydrogen on the same mass scale, but corrected for the mass of the electron.

The gyromagnetic ratio of a rigid symmetric top with a fixed charge distribution has been calculated by Condon[20]

$$g_{JK} = a\frac{K^2}{J(J+1)} + b \tag{12}$$

where the constants a and b depend on the distribution of charge within the top. The form of this equation appears to be appropriate to a symmetric top molecule, although it is doubtful whether the constants a and b are determined by the molecular charge distribution alone.

For a nonvibrating linear molecule, $K = 0$, so that from eq. (12), $g_{JK} = b$, a constant. For such a molecule, no Zeeman effect should be observed when the electric vector of the microwave is *parallel* to the magnetic field. No detectable shift was observed for OCS under these conditions. When the electric vector of the microwave was *perpendicular* to the magnetic field, the $J = 1$ to $J = 2$ rotational line broke up into a doublet. The frequency shift of either component of the doublet was

$$\frac{\nu \text{ (cycles/sec.)}}{H \text{ (oersteds)}} = 27$$

This corresponds to a gyromagnetic ratio $g = \pm .035$. The sign of the gyromagnetic ratio may be determined by observations with circularly polarized microwaves. The Zeeman effect observed for this linear molecule is the "normal" Zeeman effect with intensities and polarizations just as predicted by Lorentz in 1898. Photographs of the oscilloscope screen before and after applying the magnetic field are shown in Fig. 5.

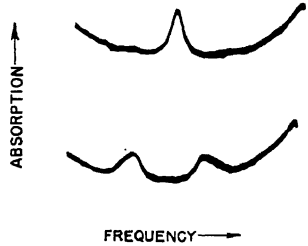


FIG. 5.—Zeeman effect for O-C-S. (*a*) With no magnetic field, or with magnetic field parallel to the electric vector of the microwave (upper curve). (*b*) With magnetic field applied perpendicular to the electric vector of the microwave (lower curve).

The Zeeman effect of N_2O has recently been studied by Jen,[19] but this case is complicated at low magnetic fields by a weak interaction between the molecule and the N^{14} nuclei.

The 22.235 kmc line of water vapor and several lines of SO_2 have been observed under both conditions of polarization. However, in no case was it possible to completely resolve the splitting.

For satisfactory resolution of microwave lines in general, it would be desirable to use a magnetic field of 25,000 oersteds, with orientations both parallel to the electric vector of the microwave, and parallel to the direction of propagation of the microwave.

7. *Hyperfine Structure of Molecular Energy Levels*

The rotational absorption lines of some molecules exhibit a hyperfine structure. According to our present interpretation, this is due to a coupling between the angular momentum of the molecule as a whole and the nuclear spin associated with one or more of the atoms making up the molecule. If such a coupling exists, the angular momentum of the molecule as a whole will no longer be a constant vector, but will slowly

precess about a fixed axis of total (nuclear plus rotational) angular momentum.[21]

The magnitude of the nuclear spin is given in units of $\frac{h}{2\pi}$ by the spin quantum number "I", which can assume only positive integral and half-integral values. If only one of the nuclei is coupled to the orbital angular momentum of the molecule then the total angular momentum (spin plus orbital) is given in the same units by the quantum number "F", where F assumes positive integral or half integral values, depending on whether I is integral or half integral. The upper and lower bounds of F are given by

$$(J - I) \lesseqgtr F \lesseqgtr (J + I) \tag{13}$$

Different values of F are associated with different orientations of the nuclear spin with respect to the orbital angular momentum of the molecule, and, therefore, to different average values of the interaction energy. A rotational level is thus broken up into a close multiplet. The main cause for hyperfine ground structure of microwave absorption lines is an interaction between the electric field gradient of the molecule and the electric quadrupole moment of the nucleus.

The magnitude of this interaction was first worked out by H. B. G. Casimir[22] for the case of a single atom. The change of energy due to a nuclear quadrupole on the axis of a symmetric top molecule has been given by Van Vleck[23,24] and Coles and Good.[17]

$$W = eQ \frac{\partial^2 V}{\partial z^2} \left[\frac{3K^2}{J(J+1)} - 1 \right] \cdot \frac{\frac{3}{4}C(C+1) - I(I+1)\,J(J+1)}{2I(2I-1)\,(2J-1)\,(2J+3)} \tag{14}$$

where $C = F(F+1) - I(I+1) - J(J+1)$,

e = electronic charge,

Q = quadrupole moment of nucleus as defined by Casimir,

V = electrostatic potential produced by all charges in the molecule except those inside a small sphere surrounding the nucleus,

z = coordinate along the molecular axis of symmetry.

A small deviation from this formula has been observed by Simmons and Gordy[25] in the hyperfine structure of NH^3. The deviation has been attributed to an effect of interaction of the molecular magnetic field arising from molecular rotation upon the N^{14} nuclear magnetic moment. The interaction has been calculated by Jauch[26] and Henderson[27]. It is

$$W = \left[a \frac{K^2}{J(J+1)} + b \right] [F(F+1) - J(J+1) - I(I+1)] \tag{15}$$

For cases where the quadrupole interaction is not sufficiently small compared with the distance between adjacent rotational energy levels,

second order corrections to this formula have been calculated by Bardeen and Townes.[28] These authors have also given methods for calculating the hyperfine structure due to two quadrupoles on the molecular axis.[29] The product $eQ\frac{d^2V}{dz^2}$ is called the quadrupole coupling. The interaction of nuclear electric quadrupole moments with asymmetric top molecules has been studied by Bragg, by Knight and Feld, and by Bragg and Golden.[30]

The Stark effect of symmetric top molecules with one nuclear quadrupole has been studied by Fano, Karplus and Sharbaugh[31] and Low and Townes.[32] The Stark effect on the hyperfine structure of $N^{14}H_3$ has been calculated by Jauch.[33]

The Zeeman effect on the hyperfine structure of $N^{14}H_3$ and of CH_3Cl has been observed by Jen,[18] who calculated the g factor of the molecules and confirmed the magnetic moment of the N^{14} and Cl nuclei. The nuclear magnetic moments had been determined previously from molecular beam experiments.

III. Applications of Microwave Spectroscopy

In addition to the immediate application of microwave absorption data to the determination of molecular structure, there are a number of other important applications in the fields of chemistry and of nuclear physics. We are able to give here only a short survey of some of these applications.

1. Determination of Interatomic Bond Distances and Molecular Configuration

In the case of a diatomic molecule, such as C-O, the moment of inertia is related to the interatomic distance r_{CO} by the relation

$$I_B = \frac{m_C \cdot m_O}{m_C + m_O} \cdot r^2_{CO} \tag{16}$$

where m_C and m_O are the masses of the carbon and oxygen atoms respectively. If we assume the molecule to be perfectly rigid, then we may calculate its moment of inertia from measurements of its absorption line frequencies.

$$I_B(10^{-40}\text{ g. cm.}^2) = 838.964/B\text{ (kmc)} = 1677.929\,\frac{J+1}{\nu_J(\text{kmc})} \tag{17a}$$

Or, if the masses of the atoms are given in physical atomic weight units (AU) and the distance in Angstrom units (A.), then the moment of inertia is

$$I_B \text{ (AU-A.}^2\text{)} = \frac{505.425}{\text{B(kmc)}} = 1010.85 \frac{(J+1)}{\nu_J\text{(kMc)}} \tag{17b}$$

Thus in the case of a diatomic molecule, a measurement of 1 rotational frequency and a knowledge of the J value provide an accurate value for the distance between the two atoms. In the case of a linear triatomic molecule such as O-C-S, it is possible to deduce the two bond distances by measuring the rotational frequency for two different isotopes of sulfur, and assuming that the interatomic distances for O-C-S^{32} and O-C-S^{34} are the same.[34,53]

A small error is incurred here, primarily due to the fact that, even in the lowest vibrational state, there is a certain amount of "zero-point" vibration. The amplitude of the zero-point vibration depends on the atomic masses involved. The effect of transverse vibration on the effective moment of inertia of the O-C-S molecule has been discussed by Townes *et al.*[35] They estimate that if the two bond distances are deduced from a single isotopic substitution (on the sulfur atom) that they may be in error by as much as $.01 \times 10^{-8}$ cm. In general, more accurate results may be derived by making as many isotopic substitutions as possible.

In any rigid linear molecule, if an atom of mass m is replaced by an isotope of mass m', then the distance of these atoms from the center of mass of the *original* molecule is given by

$$z^2 = \frac{I'_B - I_B}{m' - m} \frac{(m' - m) + M}{M} \tag{18}$$

where M is the total mass of the original molecule.

The positions of the different atoms in an asymmetric top molecule may also be found in this way, referred to inertial axes of the molecule which pass through its center of mass. When an isotopic substitution is made in an asymmetric top molecule, it is necessary, of course, to take into account both the translation and *rotation* of the inertial axes from their original position. Rather accurate values for the bond distances may thus be calculated, provided that the molecule is sufficiently rigid.

The inertial properties of the asymmetric-top molecule are conveniently described in terms of the fundamental dyadic

$$F = \begin{pmatrix} \Sigma mx^2 & \Sigma mxy & \Sigma mxz \\ \Sigma myx & \Sigma my^2 & \Sigma myz \\ \Sigma mzx & \Sigma mzy & \Sigma mz^2 \end{pmatrix}$$

whose principal values are F_A, F_B, and F_C. In terms of the above dyads the inertia dyadic is

$$I = \begin{pmatrix} F_{yy} + F_{zz} & -F_{xy} & -F_{xz} \\ -F_{yx} & F_{zz} + F_{xx} & -F_{yz} \\ -F_{zx} & -F_{zy} & F_{xx} + F_{yy} \end{pmatrix}$$

and its principal values are $I_A = F_B + F_C$, $I_B = F_C + F_A$, and $I_C = F_A + F_B$.

Interpretation of the rotational spectrum of the molecule results in experimental values for the principal moments of inertia I_A, I_B, I_C. From these, the principal moments of the fundamental dyadic are

$$\begin{aligned} F_A &= \tfrac{1}{2}(I_B + I_C - I_A) \\ F_B &= \tfrac{1}{2}(I_C + I_A - I_B) \\ F_C &= \tfrac{1}{2}(I_A + I_B - I_C) \end{aligned}$$

Now, if the original molecule has the principal moments F_A, F_B, F_C, and substitution of a different isotope results in new principal moments F'_A, F'_B, F'_C, then the position of that atom, referred to the inertial axes of the *original* molecule, may be easily calculated on the assumption of perfect rigidity. If no two of the principal moments of inertia are approximately equal, we may use the following approximation for calculating the distance of the isotope from the yz-plane.

$$\begin{aligned} x^2 = \frac{F'_A - F_A}{m' - m} \cdot \left[1 + \frac{m' - m}{M}\right] \left[1 + \frac{F'_B - F_B}{F_B - F_A} + \frac{F'_C - F_C}{F_C - F_A}\right. \\ \left. + (F'_B - F_B)(F'_C - F_C) \left\{\frac{1}{(F_B - F_A)(F_C - F_A)} + \frac{1}{(F_B - F_C)(F_A - F_B)}\right.\right. \\ \left.\left. + \frac{1}{(F_C - F_A)(F_B - F_C)}\right\} + 0\left(\frac{\Delta m}{M}\right)^3\right] \end{aligned} \tag{19}$$

where m is the mass of the original isotope,

m' is the mass of the new isotope,

M is the total mass of the original molecule.

Similar equations hold for y^2 and z^2, so that the position of the interchanged atom with respect to the inertial axes of the original molecule is completely determined. In fact, by using the auxiliary equations $\Sigma mx = 0$, $\Sigma mx^2 = F_A$, etc., it is possible to completely fix the interatomic distances by making isotopic substitutions on $n - 2$ of the atoms in an n-atomic molecule. If the molecule possess elements of symmetry, a smaller number of isotopic substitutions is necessary for a complete determination.

It should be emphasized that these simple equations are subject to small errors, due to the slight change of interatomic distance when an isotopic substitution is made. The magnitude of the vibrational effect

on the average internuclear distance may be determined by measurements of the rotational spectra of vibrating molecules, provided that these are sufficiently intense to be observable. The effective moments of inertia of the vibrating molecules are slightly modified by the vibration so that rotational spectra due to the vibrating molecules appear slightly shifted from that associated with the ground state. The different spectra may be distinguished by changing the temperature of the absorption cell and noting the change in intensity of the absorption lines.

The deduction of molecular moments of inertia from the microwave spectrum of an asymmetric top molecule presents some difficulty. The energy levels up to $J = 12$ for a rigid asymmetric top have been given by King *et al.* for eleven different values of the asymmetry parameter.[36] In the interpretation of the SO_2 microwave spectrum, it was found that the Stark effect was a great help in fixing the values of J associated with a particular absorption line.[37] Lines associated with high rotational levels could be segregated by their behavior at low temperatures. If the molecule possesses a hyperfine structure, this may assist in assigning rotational quantum numbers to an absorption line. The hyperfine structure for an asymmetric top molecule was first calculated by Bragg.[30] The calculation of higher rotational energy levels, including the effect of centrifugal distortion, has been considered recently by Golden.[38]

The effect of centrifugal distortion is indicated by the second term in the expression for the rotational energy of a diatomic molecule

$$W(J) = BJ(J + 1) - DJ^2(J + 1)^2 \tag{20}$$

$$D = \frac{4B^3}{\omega^2} \tag{21}$$

where ω is the vibrational frequency of the diatomic molecule. Since ω is usually much greater than B, D is very much smaller than B. The notation is that of Herzberg.[39] For further details see references 1 and 11.

In the absence of zero-point vibration, and at low rotational speeds, we might imagine that the internuclear distance would assume the equilibrium value r_e, and the moment of inertia of the molecule would be I_e. The value of B_e in kmc would then be

$$B_e(\text{kmc}) = \frac{h}{8\pi^2 I_e} 10^{-9} = \frac{8.3896 \times 10^{-38}}{I_e(\text{g. cm.}^2)} \tag{22}$$

In a vibrational state v the rotational constant B_v is to a first approximation

$$B_v = B_e - \alpha_e(v + \tfrac{1}{2}) + \cdots v = 0, 1, 2 \cdots \tag{23}$$

α_e is a small constant representing the influence of vibration of the effective moment of inertia. The subscript e refers to *equilibrium* (not to *effective*). Thus we may calculate the *ideal* quantity B_e by means of the relation

$$B_e = B_0 - \frac{B_1 - B_0}{2} \tag{24}$$

provided that the effective rotational constant can be measured in both the ground state and the first excited vibrational state.

Since vibration may affect also the small rotational constant D, which represents the influence of centrifugal force, we should use

$$D_v = D_e + \beta_e(v + \tfrac{1}{2}) \tag{25}$$

where β_e is small compared to D_e.

By measuring rotational frequencies corresponding to two or more different values of J, it is possible to determine the quantity D_v experimentally. In the case of a diatomic molecule, this provides an independent determination (by using eq. 21) of the molecular vibrational frequency.

Obviously, vibrational frequencies of polyatomic molecules cannot all be obtained by the above scheme. However, in the case of the COSe molecule, Strandberg and Wentink[40] successfully calculated the frequency ω_2 by making intensity-ratio measurements on rotational lines associated with both the excited vibrational state and with the ground state. Since the intensity is proportional to the square of the electric dipole moment, they made allowance for the fact that the dipole moment, as measured by the Stark effect, was slightly different in the two different vibrational states.

Isotopic substitutions within a molecule may also be used to determine the relative positions of different atoms in the molecule. For example, in the linear molecules C-N-Cl and C-N-Br, it was possible to decide that the carbon atom (and not the nitrogen) occupies the central position.[41]

2. Determination of Electronic Structure of Molecules

The electric dipole moment of a molecule can theoretically be determined by making careful measurements of the intensities of its microwave absorption lines, since these are proportional to the square of the electric dipole moments. A more exact determination of the dipole moment can be made by measuring the shift in energy levels of the molecule produced by a constant or low frequency electric field imposed on the absorbing gas. This "Stark" effect is discussed in Section II, 6.

The measurement of electric dipole moments by the methods of microwave spectroscopy appears to be more direct than by previous methods and the data obtained refer to a particular rotational state, rather than being statistical results inferred from the total behavior of all states.

Since the microwave methods are free from large errors due to traces of impurities in the gas, or to small temperature variations, the methods of microwave spectroscopy promise to provide many values for molecular dipole moments which are more reliable than those obtained by other methods. According to Wentink *et al.*[42] relative dipole moments can be measured to better than one part in a thousand. The dipole strength appears to depend on the vibrational state of the molecule.

The gradient of the electric field at the position of the different atomic nuclei may be determined by studing the hyperfine structure of absorption lines, provided that the atomic nuclei involved possess quadrupole moments, and that these are known. Studies of different molecules by these means have made important contributions to our knowledge of the electronic structure of molecules.[43]

The weak magnetic fields in a molecule may also interact with the atomic nuclei, making possible a determination of the magnitude of the field of the molecule at the nuclear positions.

The magnetic moment of the molecule as a whole, due to its rotation, can be determined experimentally by studying the Zeeman effect of microwave absorption lines, but this quantity has not been correctly calculated theoretically so far for any molecule.

3. Study of Internal Rotation in Molecules

Although the simpler molecules rotate essentially as rigid bodies, in many molecules it is possible for one part to rotate with respect to another part. A molecule for which the internal rotation should be almost unhindered is $CF_3—C{\equiv}C—CH_3$. On the other hand, the internal rotation of a molecule such as $CCl_3—CH_3$ is, apparently, strongly hindered by potential barriers, which tend to keep the hydrogen group from rotating with respect to the chlorine group. The effect of free and hindered rotation on the spectra of molecules is discussed by Herzberg.[1] In the case of symmetric top molecules, the microwave absorption spectrum should not be affected by internal rotation about the symmetry axis, except insofar as it slightly affects the moment of inertia of the molecule about an axis perpendicular to the symmetry axis. However, for asymmetric top molecules, microwave spectroscopy provides an excellent means of studying internal rotation. Up to the present time, our knowledge of the hindrance to rotation comes mainly from measure-

ments of the entropy and low-temperature heat capacity of gases. The important case of CH_3OH has been studied by Dennison, who finds that the internal rotation of the OH group is hindered by a potential barrier of the order of 10F.[44]

4. *Calculation of Thermodynamic Quantities*

The values of thermodynamic quantities such as the specific heats of gas may be calculated with great precision from spectroscopic data, since these depend on the molecular energy levels.

The fraction of molecules that occupy a particular quantum state W_j is

$$\frac{N_j}{N} = \frac{g_j e^{-\frac{W_j}{kT}}}{\sum_i g_i e^{-\frac{W_i}{kT}}} \tag{26}$$

where k is Boltzmann's constant

g_j is a statistical weight.

The quantity $Z = \sum_i g_i e^{-\frac{W_i}{kT}}$ is called the sum-over-states or *partition function* of the gas.

The internal energy per mole of a perfect gas is related to this partition function by the equation

$$E = E_0 + RT^2 \frac{d(\ln Z)}{dT} \tag{27}$$

where E_0 is the zero-point energy at the absolute zero of temperature,

R is the gas constant per mole.

All other thermodynamic quantities can also be expressed in terms of this quantity.

For a perfect gas the partition function can be separated into a product of translational and internal partition functions

$$\begin{gathered} Z = Z_{\text{tr}} \cdot Z_{\text{int}} \\ Z_{\text{tr}} = 1.8794 \times 10^{20} M^{\frac{3}{2}} T^{\frac{3}{2}} V \\ Z_{\text{int}} = \Sigma g_{\text{int}} e^{-\frac{W_{\text{int}}}{kT}} \end{gathered} \tag{28}$$

M = molecular weight of the gas,

V = the volume of the gas per mole

The internal energy of a molecule is the sum of the electronic, vibrational, and rotational energies. The energies of excited electronic states are almost always so high that the corresponding Boltzmann

factors $e^{-\frac{W}{kT}}$ are entirely negligible at room temperature. If the interaction of rotation and vibration are neglected, the internal partition function may then be written

$$Z_{\text{int}} = Z_{\text{rot}} \cdot Z_{\text{vib}} \tag{29}$$

The Boltzmann factor corresponding to excited vibrational states must be considered for the very low vibrational frequencies even at room temperature. If we make the assumption that the vibrations are harmonic (usually valid at room temperatures) then the vibrational partition function is

$$Z_{\text{vib}} = \left[1 - e^{-\frac{\omega_1 h}{kT}}\right]^{-1} \cdot \left[1 - e^{-\frac{\omega_2 h}{kT}}\right]^{-1} \cdot \left[1 - e^{-\frac{\omega_3 h}{kT}}\right]^{-1} \tag{30}$$

where ω_1, ω_2 . . . are vibrational frequencies of the molecule. If two or more of the ω_i coincide, the appropriate factor must be repeated the same number of times in the above product.

The rotational partition functions, which are of greatest interest at low temperatures, are given below for all classes of molecules *which have a permanent electric dipole moment.*

For a symmetric top molecule (that is, a molecule with a p-fold symmetry axis, where $p \geq 3$), the rotational partition function is

$$Z_{\text{rot}} = \sum_{J=0}^{\infty} \sum_{K=-J}^{J} \sum_{M=-J}^{J} \frac{g_I e^{-[BJ(J+1)+(C-B)K^2]\cdot\frac{1}{kT}}}{p} \tag{31}$$

Here the weight factor g_I is due to identical nuclei symmetrically disposed about the axis of symmetry. This weight factor actually depends on the spins of the identical nuclei, and on the rotational states. Its averaged value over several consecutive rotational states is unity, so the results of practical statistical calculations are independent of nuclear spin (see Section IV, 2). For molecules such as NH_3, in which each rotational level is doubled by the inversion phenomenon, we count twice as many rotational levels. However, the averaged statistical weight per level is then halved, so that the partition function for a molecule with inversion doubling is the same as that for ordinary molecules.

The factor kT and the rotational constants are conveniently expressed in ergs, kmc, or cm.$^{-1}$

$$\begin{aligned} k &= 1.3805 \times 10^{-16} \text{ ergs/deg.} \\ &= 20.84 \text{ kmc deg.} \\ &= .6951 \text{ cm.}^{-1}\text{/deg.} \end{aligned} \tag{32}$$

at $T = 300°$,

$$\begin{aligned} kT &= 6252 \text{ kmc} \\ &= 208.53 \text{ cm.}^{-1} \\ &= 4.14 \times 10^{-21} \text{ joules.} \end{aligned} \tag{33}$$

Since the value of B for a molecule such as CH_3Cl is only 13 kmc or .43 cm.$^{-1}$, it is evident that the Boltzmann factor in eq. 31 is close to unity for low values of J. We may then use the following expansion for a symmetric top molecule with a p-fold axis of symmetry and a *permanent electric dipole moment.*

$$Z_{\text{rot}} = \frac{e^{\frac{Bh}{4kT}}}{p}\sqrt{\pi}\left[\frac{(kT)^3}{h^3CB^2}\right]^{\frac{1}{2}} \left[1 + \frac{1}{12}\left(1 - \frac{B}{C}\right)\frac{hB}{kT} + \frac{7}{480}\left(1 - \frac{B}{C}\right)^2\left(\frac{hB^2}{kT}\right) + \cdots\right] \tag{34}$$

which applies down to extremely low temperatures. At higher temperatures this expression approaches the value

$$Z_{\text{rot}} = \frac{168.7}{p}\left[\frac{T^3}{B^2C}\right]^{\frac{1}{2}} \tag{35}$$

where B, C are expressed in kmc.

When the value for Z_{rot} from eq. (35) and the value for Z_{tr} are substituted into eq. (27) we have the classical result

$$E = E_0 + 3RT \tag{36}$$

Thus it is only the low temperature *deviation* from the classical value for the internal energy that depends on the moments of inertia of the molecule.

According to eq. (36) a zero moment of inertia I_C will give a zero value for the rotational partition function. Therefore, in this special case (of a linear molecule with a permanent electric dipole moment) we should use the expansion

$$Z_{\text{rot}} = \frac{kT}{hB} + \frac{1}{3} + \frac{1}{15}\cdot\frac{hB}{kT} + \frac{4}{315}\left(\frac{hB}{kT}\right)^2 + \cdots \tag{37}$$

At room temperatures this approaches the classical value

$$Z_{\text{rot}} = 20.84 \cdot \frac{T}{B} \tag{38}$$

where B is in kmc.

The rotational partition function for arbitrary asymmetric top molecules cannot be calculated explicitly, but may be approximated by the expression

$$Z_{\text{rot}} = e^{\frac{h\sqrt{BA}}{4kt}} \sqrt{\pi} \left(\frac{kT}{h}\right)^{\frac{3}{2}} \left[\frac{1}{ABC}\right]^{\frac{1}{2}} \left[1 + \frac{1}{12}\left(1 - \frac{\sqrt{BA}}{A}\right)\frac{h\sqrt{BA}}{kT} + \cdots\right] \quad (39)$$

If the molecule possesses a twofold symmetry axis, this partition function should be divided by 2. When $B = A$ this equation is equivalent to eq. (34).

The important contributions to the specific heat which are associated with internal rotation of the molecules are discussed in the more complete treatments of this subject by Wilson[45] and Herzberg.[1]

5. *Determination of Nuclear Moments, Spin, and Statistics*

If one or more of the atomic nuclei in a molecule possesses an electric quadrupole moment, each energy level exhibits a hyperfine structure. The interpretation of this structure provides definite values for the nuclear spins involved, and for the coupling constants associated with each nucleus. Unfortunately, the electric quadrupole moment of the nuclei cannot be inferred from these coupling constants without a knowledge of the gradient of the electric field in the neighborhood of the particular nuclei. An attempt to estimate the electrostatic potential in certain molecules from atomic spectra of the atoms involved has been made by Townes.[46] The gradient of the electric field at the atomic nucleus is usually reduced when the atom is bonded into a molecule. In case the bond is strongly ionic in nature, the gradient of the electric field at the nucleus may be greatly reduced.[43]

If the quadrupole coupling can be measured for several different isotopes in the same position of the molecule, then the ratio of the electric quadrupole moments of the several nuclei is immediately known. This information is of interest for the theory of nuclear structure. An electric quadrupole moment equal to zero indicates that the nucleus in question probably has a spin of 0 or $\frac{1}{2}$. (These two spin values automatically give an effectively spherical charge distribution over the nucleus.)

The *magnetic dipole* moment of an atomic nucleus may be determined by placing the molecule containing it in a magnetic field. If the nucleus in question has an *electrical quadrupole* moment, giving rise to hyperfine structure in the microwave absorption lines of the molecule, then the

behavior of this structure under the influence of a magnetic field makes possible a determination of the nuclear magnetic moment. The hyperfine structure of $N^{14}H_3$ and of CH_3Cl has been studied in this manner by Jen.[18]

If hyperfine structure is present, it is usually resolved more easily in the microwave region than is possible in the ultraviolet region of atomic spectroscopy, and the resulting determination of nuclear spin is more conclusive. Thus the observation of hyperfine structure of Cl^{35} and Cl^{37} in several different molecules,[47,11] which are in accordance with a spin of $\frac{3}{2}$ are much more to be trusted than the previous determinations of $\frac{5}{2}$ from intensity measurements on Cl_2 in the visible and ultraviolet regions. The lack of hyperfine structure observed for C^{13}, S^{34}, Se^{78}, Se^{80}, Se^{82} suggests that these nuclei possess spins of $\frac{1}{2}$, 0, 0, 0, 0 respectively. Assignments of these spins remain somewhat doubtful at present, however, because the lack of observable coupling may possibly be due to a very small molecular anisotropy at the positions of these nuclei.

If a molecule contains two or more identical nuclei, then different rotational states have different statistical weights. These statistical weights may be determined by comparing the intensities of absorption lines or of Stark components[48] associated with these states, and the nuclear *spins* thereby deduced.

The *statistics* obeyed by nuclei may be inferred from observations on asymmetric top molecules which possess twofold axes of symmetry, or from the study of certain symmetric top molecules which possess twofold axes of symmetry. For example, the NH_3 molecule possesses an *oscillating* electric dipole moment which gives rise to a microwave absorption spectrum in the region of 20 kmc. From the long term point of view this molecule has three twofold axes of symmetry and a dipole moment equal to zero. A study of the absorption lines of NH_3 revealed that one component of the "K-type doublets" was missing.[49,50] Calculations concerning this doublet separation by H. H. Nielsen and D. M. Dennison[51] provided confirmation of the fact that the hydrogen nuclei have spins of $\frac{1}{2}$ and obey Fermi statistics. In the case of ND^3 the statistical weights of the rotational energy levels would be quite different, because of the different spin and the Bose statistics obeyed by the deuterons.

6. Determination of Mass Differences of Isotopes

In a diatomic molecule, substitution of a different isotope provides information about the ratio of the two isotopic masses. The bond distance depends slightly on the isotopic masses, but this vibrational effect can be determined if the effective moments of inertia in the ground vibrational state and in the first vibrational state are both determined.

Thus Townes *et al.*[52] by studying the $J = 3$, to $J = 4$ transition of ICl^{35} and $ICl,^{37}$ found that the ratio of the two chlorine nuclear masses was .945,9811 ± .000,00050. In the study of OCS^{32}, OCS^{33} and OCS^{34}, Townes *et al.*[53] assumed that the vibrational correction was proportional to the masses, and found the ratio of mass differences $S^{33} - S^{32}/S^{34} - S^{32}$ to be 0.50066 ± .00015.

In the case of heavier polyatomic molecules, it may be possible to find rotational lines for molecules in all the first excited vibrational states, so that the mass ratios of the isotopes in these molecules may be determined. Thus Strandberg and Wentink[40] have found rotational lines for two of the vibrational modes of COSe. If a rotational line for molecules in the third mode of vibration is found, it will be possible to calculate mass differences of the different isotopes within 10% of the packing fraction.

7. Chemical and Isotopic Analysis

The absorption lines of a gas at low pressure are so sharp that there appears to be room for a million separate lines in the presently accessible microwave region. Since not more than 100 lines have been reported for any one chemical compound, it might appear to be possible to make chemical analyses on mixtures containing twenty or thirty components without mutual interference of the different components. However, this is certainly not practical at the present time. The reason that a larger number of lines has not been reported for any single compound is mainly due to the insensitivity of detection of weak absorption lines. In particular, the detection of heavy molecules with small dipole moments, such as the heavier hydrocarbons, is extremely difficult, even in the pure state. However, as the sensitivity of detecting apparatus and the workable frequency range is increased, it should be possible to detect all polar molecules (except very large molecules).

The use of microwave spectroscopy for analysis has been discussed by Dailey.[53a] He estimates that gases which absorb strongly will be detectable in dilutions greater than 10^7. We discuss sensitivity of detection further in Section V, 9.

Molecules containing isotopes of different masses are widely separated in the absorption spectrum, so that a microwave spectroscope appears rather better suited to isotope analysis than an infrared or ultraviolet spectrograph.

Although the sensitivity of detection of minor impurities may never be as high as that achievable in infrared and ultraviolet spectroscopy, or in mass spectroscopy, there will almost certainly be many analyses that can be carried out most profitably by means of a microwave spectroscope.

8. *Study of Liquids and Solids*

In solids and liquids, molecular rotation is strongly hindered by potential barriers, so that microwave rotation lines have not been observed. It seems likely that the increased damping forces in liquids will result in a rotational line breadth of 100 kmc or more, so that in the presently accessible microwave region it would be difficult to observe individual rotation lines.

Regions of strong absorption and anomalous dispersion have been observed, but these are generally associated with relaxation times of the molecular dipole moments. For dilute solutions of polar molecules in simple nonpolar solvents, it is possible to apply the Debye theory, in which a single relaxation time τ is assumed. The dielectric loss in such a solution is given by:

$$\tan \delta = \frac{(\epsilon + 2)^2}{\epsilon} \frac{4\pi}{27} \frac{\mu^2}{kT} \frac{Nc 2\pi\nu\tau}{[1 + (2\pi\nu\tau)^2]} \tag{40}$$

where δ = loss angle (complement of the phase angle),
ϵ = dielectric constant of the solvent,
N, c = Avogadro's number, concentration of the solute,
k, T = Boltzmann's factor, absolute temperature,
ν, μ = frequency of the radiation, electric dipole moment.

In many cases the above equation provides values for the dipole moments which agree with determinations in the gaseous phase. For concentrated solutions, however, and for most pure polar liquids, the interaction between neighboring dipoles has an important effect, which is difficult to treat theoretically.

In many solids and liquids, molecules do not exist as such; only the atoms or ions retain any semblance of individuality. Since atoms do not have permanent electric dipole moments, they can experience a torque in a microwave field only by virtue of their magnetic dipole moments. Usually the atoms or ions do not possess a permanent magnetic dipole moment either, since their electrons tend to be paired off in such a way with those of neighboring atoms that the magnetic moments cancel. A relatively small number of ions possess a net magnetic moment. Since this gives them the property of paramagnetism, they are generally called paramagnetic ions.

A particular atom or ion may (in an approximate way) be regarded as in a strong electric field representing the effect of the neighboring ions. In some crystals, the combination of the strong electric field and the spin-orbit coupling may produce (in addition to large Stark splittings)

small splittings of the order of 10 to 100 kmc. Such ions exhibit resonance absorption in the microwave region.[54]

A magnetic field imposed on the crystal splits the energy levels into separate components. In the absence of electron spin-orbit coupling, the resonance frequency for transitions between different magnetic components satisfies the Larmor equation for electron spins.

$$\frac{\nu}{H} = \frac{e}{2\pi mc} = 2.80 \text{ Mc per oersted} \tag{41}$$

Thus a magnetic field of 10,000 oersteds will produce resonance absorption in the region of 28 kmc.

The combination of an internal electric field and spin-orbit coupling may result in an effective gyromagnetic ratio even greater than that indicated by eq. 41. In the case of single crystals, the measured gyromagnetic ratio depends on the orientation of the crystal axis with respect to the magnetic field.[55,56] Measurements of this type provide information about the strength and symmetry of the intercrystalline electric fields, the magnetic dipolar interactions, spin-lattice relaxation times, and of the magnitude of exchange coupling between the electrons in neighboring ions.

Ferromagnetic materials exhibit a resonance absorption in accordance with eq. 41, but where the effective value of H differs from the external magnetic field in a way which depends on the shape of the specimen, and its electrical conductivity.[57]

9. Other Applications

The above list of applications of microwave spectroscopy is by no means complete. The pressure broadening of absorption lines has provided values for the collision cross-section of molecules[58] and will doubtless lead to a better understanding of collision processes in gases. Attempts will be made to study, and possibly to influence, the mechanism of gas reactions.

The microwave absorption lines of gases have been used to stabilize the frequencies of microwave transmitters and receivers to one part in 10^7. There is some hope that an oscillator may be stabilized by a microwave absorption line to a long time accuracy of one part in 10^{10} or even 10^{12} and that an oscillator so stabilized may provide the primary standard of time.[53,59]

The study at low temperatures of the properties of solids (particularly the properties of solids containing trapped electrons) offers interesting possibilities. Microwave spectroscopic studies outdoors will have applications to radar, meteorology, and possibly even to astrophysics.

We cannot discuss here all the possible side branches of microwave spectroscopy.

IV. The Width and Intensity of Absorption Lines

1. The Causes of Line Width

The most important contributions to the width of microwave absorption lines are usually:

(a) Doppler effect.

(b) High microwave power level in the waveguide.

(c) Collisions of the molecules with each other or with the waveguide.

If the microwave power level is kept low, so that it does not affect thermal equilibrium, then Bouguer's law of absorption is valid. That is, for a plane wave travelling a distance x through an absorbing medium,

$$\frac{I}{I_0} = e^{-\gamma x} \qquad \text{or} \qquad \log_e\left(\frac{I_0}{I}\right) = \gamma x \tag{42}$$

where I_0 is the incident microwave power at a particular frequency,

I is the transmitted microwave power,

γ is a constant called the absorption coefficient, which is a function of the microwave frequency.

This law is sometimes called Lambert's law. The absorption coefficient in cm.$^{-1}$ is equal to .2303 times the attenuation constant of the medium in decibels per centimeter or twice the (amplitude) attenuation constant of the medium in nepers per centimeter.

The most serious limitation on line sharpness is due to the Doppler effect. A molecule which is rotating at a frequency ν_0 and moving with a velocity v toward an observer is apparently (to that observer) rotating with a frequency ν given by $\dfrac{\nu - \nu_0}{\nu_0} = \dfrac{v}{c}$ where c is the velocity of light. This Doppler shift, for molecules moving randomly in all directions, gives rise to an absorption line of finite width. The absorption coefficient at a frequency ν, which is close to the resonant frequency ν_0 is

$$\gamma_\nu = a(2\pi b^2)^{-\frac{1}{2}} e^{-\frac{(\nu-\nu_0)^2}{2b^2}} \tag{43}$$

where $a = \int \gamma_\nu d\nu$ whose value is given in eq. (57),

$b^2 = \dfrac{kT\nu_0^2}{mc^2}$,

T is the absolute temperature,

m is the mass of the molecule,

k is Boltzmann's constant.

The half-maximum line width for such a line is given by

$$\frac{w}{\nu} = 7.15 \times 10^{-7} \left(\frac{T}{M}\right)^{\frac{1}{2}} \tag{44}$$

where w is the total width at half maximum,

M is the molecular weight in atomic weight units.

At 300° absolute (room temperature) this is

$$\frac{w}{\nu} = 1.23 \times 10^{-5} \left(\frac{1}{M}\right)^{\frac{1}{2}} \tag{45}$$

As an example, the room-temperature width of an absorption line of SO_2 in the neighborhood of 30,000 Mc should be .046 Mc, provided that the gas pressure and the microwave power level in the waveguide are kept sufficiently low.

If, at sufficiently low gas pressure and microwave power level, the observed width of an absorption line is greater than that calculated from the Doppler effect, then the single broad line may actually be composed of several sharp lines lying very close together. This internal structure could be due to imperfect degeneracy of the energy levels involved, or to weak coupling between the angular momentum of the molecule as a whole and the spins of its atomic nuclei.

An apparent widening of absorption lines is caused by high microwave energy densities in the absorption cell (see Section IV, 4). This is explained as a redistribution of the energy level population, which is particularly effective at frequencies near the center of the line, but less effective at frequencies in the wings of the line.[60,61,62,63] Using the 10.7×4.3 mm. waveguide which is standard at 24,000 Mc, and a gas pressure of 10^{-2} mm. Hg, the broadening of the stronger ammonia lines is noticeable at powers of 10^{-4} watts. At a pressure of 5×10^{-2} mm. Hg, the effect is noticeable at 25×10^{-4} watts.

When both the gas pressure and the microwave power level in the waveguide are too low to influence the line width, then Beer's law of absorption is valid. The absorption coefficient γ is proportional to the gas pressure. At pressures of 10^{-2} mm. and higher the collisions of the gas molecules with each other begin to contribute to the line width. At somewhat higher pressures, the width of an absorption line becomes accurately proportional to the gas pressure.

When the pressure of an absorbing gas is sufficiently high that the absorption line width is due entirely to collisions, but not so high that the width of an absorption line is comparable to the resonant frequency of

the line, then the total width of the line at half intensity is equal to $\frac{1}{\pi\tau}$, where τ is an average of the time between collisions.

From the kinetic theory of gases, the relation between the mean collision time and the effective molecular collision cross section s is

$$\tau_{\kappa}(\text{microsecs.}) = \frac{.04944}{s} \frac{(MT)^{\frac{1}{2}}}{P\ (\text{mm.})} \tag{46}$$

where s is the collision cross section in square Angstrom units (1A = 10^{-8} cm.),

M is the molecular weight in atomic weight units,

P is the gas pressure in mm. Hg,

T is the absolute temperature.

By measurements on the line widths in the inversion spectrum of ammonia, Bleaney and Penrose[8] and Townes[60] found that the effective collision diameter for ammonia molecules was about four times that found by ordinary kinetic measurements. They attribute this to the large dipole-dipole forces between ammonia molecules. This theory received some confirmation by Townes' measurements on ammonia diluted by argon, which gave a mutual collision diameter for ammonia and argon molecules in agreement with that obtained from kinetic measurements. The pressure broadening of ammonia lines as a function of rotational quantum number has been discussed by Bleaney and Penrose[58,64] and by Masatake Mizushima.[65] The broadening at high pressures has been studied by Smith[66] and Weingarten[78] Bleaney and Loubser, Anderson, and Margenau.[67]

Even at extremely low gas pressures, there remains a small line breadth due to collisions of the molecules with the walls of the waveguide. According to Bleaney and Penrose,[64] the half-intensity width due to this cause is about .040 Mc for NH_3 at room temperature, in a waveguide with a narrow dimension of 4.3 mm. Thus if the waveguide dimensions are not smaller than those customary in each frequency range, the line broadening due to collisions with the walls will not be more than half that due to the Doppler effect.

2. *The Absorption Coefficient for Lines Broadened by Collisions Alone*

The considerations of statistical weights in this section exclude molecules of such high symmetry that they could not possibly absorb in the microwave region. Moreover, the $(2I + 1)$-fold degeneracy due to each nucleus of spin $I > 0$ is omitted from the statistical weight factors. The absorption coefficient for an absorption line is given as if the hyper-

fine structure associated with nuclear spins $I > \frac{1}{2}$ were never resolved. In case hyperfine structure *is* resolved, the sum of the hyperfine structure component intensities (in the absence of an electric field) is equal to the intensity of the whole line as calculated here (for example, from eq. 54). The *relative* strengths of the hyperfine structure components may be found in tables given by Condon and Shortley.[21] The method of doing this has been described by Bardeen and Townes.[21]

Van Vleck and Weisskopf[63] have calculated the absorption coefficient for a microwave absorption line on the assumption that the main cause of line broadening is "sharp" collisions. That is, that the duration of a collision is short compared to one cycle of the microwave field. This assumption appears to be justified, at least for the lower microwave frequencies. Their results apply to gases not only at low pressure, but to gases at pressures up to atmospheric pressure. (See, however, ref. 67.)

Because of the usefulness of the Stark effect in detecting and interpreting molecular microwave spectra, it is often necessary to calculate intensities of the individual Stark components of a line. The initial and final rotational states associated with a particular Stark component may be designated by J, i, M and J', i', M', where i is an inner quantum number and M is the outer quantum number. The approximate expression given by Van Vleck and Weisskopf for the absorption coefficient of a single Stark component is

$$\gamma_\nu = C \cdot \mu^2 \cdot \nu^2 \cdot \frac{N}{T} \cdot F(\nu_{JiM}^{J'i'M'}, \nu) \frac{g_I e^{\frac{-W_{\nu JiM}}{kT}}}{pZ} \cdot |D_{JiM}^{J'i'M'}|^2 \qquad (47)$$

where α_ν = absorption coefficient in cm.$^{-1}$,

$C = \frac{8\pi^3}{ck} = 5.994 \times 10^7$ deg. sec.^{3} cm.$^{-3}$ g.$^{-1}$,

W = average energy of the two energy levels involved in transition,

c, k = velocity of light, Boltzmann constant,

μ = electric dipole moment in esu or magnetic dipole moment in emu,

ν, T = microwave frequency in cycles per second, absolute temperature,

τ = mean interval between collisions,

N = number of molecules per cm^3 in the absorption cell.

$$F(\nu_{JiM}^{J'i'M'}, \nu) = \frac{1}{\pi} \left\{ \frac{\frac{1}{2\pi\tau}}{(\nu_{JiM}^{J'i'M'} - \nu)^2 + \left(\frac{1}{2\pi\tau}\right)^2} + \frac{\frac{1}{2\pi\tau}}{(\nu_{JiM}^{J'i'M'} + \nu)^2 + \left(\frac{1}{2\pi\tau}\right)^2} \right\}$$

$\nu_{JiM}^{J'i'M'} = \frac{1}{h}\{W_{J'i'M'} - W_{JiM}\}$ = the resonant frequency in cycles per second

$\mu^2 |D_{JiM}^{J'i'M'}|^2$ = squared matrix element of the dipole moment with reference to the two states involved

p, Z = symmetry number, partition function. The value of pZ for different types of molecules is given in Section III, 4.

g_I = weight factor is equal to unity except when the molecule contains identical nuclei occupying symmetrical positions in the molecule. It depends on the over-all species (apart from nuclear spin) of the two levels involved in the transition. The two levels involved always have the same over-all species (apart from nuclear spin). This law is elucidated in reference 1 (page 62).

The quantum number v, which does not change during a microwave transition, is used to designate the vibrational state of the molecule. Generally, most of the molecules are in the ground vibrational state at room temperature.

As mentioned in Section II, 5, it is often convenient, in studying the Stark effect, to keep the electric vector of the microwave parallel to the constant electric field. In this case we need consider only the matrix elements $D_{JiM}^{J'i'M'}$, where $M' = M$. For a rotational transition of a symmetric top molecule, observed under these conditions, we have the $J' = J + 1$, $K' = K$, so that the appropriate squared element of the direction-cosine matrix[69] is

$$|D_{J,K,M}^{J+1,K,M}|^2 = \frac{[(J+1)^2 - M^2][(J+1)^2 - K^2]}{(J+1)^2(2J+1)(2J+3)}. \tag{48}$$

When an electric field is not applied, the Stark components corresponding to different values of the outer quantum number M coincide. The sum of the squared matrix elements is then (for a symmetric top)

$$\sum_{M=-J}^{M=J} |D_{J,K,M}^{J+1,K,M}|^2 = \frac{1}{3}\frac{(J+1)^2 - K^2}{(J+1)} \tag{49}$$

For some purposes, it is convenient to consider the K values to be greater than (or equal to) zero, and to multiply the statistical weight by 2 for all values of K (except $K = 0$). This is done, for example, in Table II.

The most important class of symmetric top molecules consists of those which possess a threefold symmetry axis. n-atomic molecules of

TABLE II. Rotation-vibration species for molecules with a threefold symmetry axis and a permanent electric dipole moment

Vibration state	Rotation state (Here K is considered to be a positive number or zero)	Vibr-rot species for each theoretically distinct level or microwave absorption line
Ground state or any combination of excited symmetrical vibrations	$K = 0$	A
	K divisible by 3	A A
	K indivisible by 3	E
One degenerate vibration singly excited plus any combination of excited symmetrical vibrations	$K = 0$	E
	K divisible by 3	E E
	K indivisible by 3	A A E
One degenerate vibration doubly excited plus any combination of excited symmetrical vibrations	$K = 0$	A E
	K divisible by 3	A A E E
	K indivisible by 3	A A E E
One degenerate vibration triply excited or two degenerate vibrations singly excited	$K = 0$	A A E
	K divisible by 3	A A A A E E
	K indivisible by 3	A A E E E

this class can execute $n-1$ vibrations which maintain the symmetry axis, and the same number of doubly degenerate vibrations, which destroy it. The relationship between the type of vibration, the rotational quantum number K, and the rotation-vibration species is given in Table II for molecules which have a permanent electric dipole moment. The statistical weight g_I of each theoretically distinct level is given in Table III. Certain unusual molecules such as NH_3 possess a "semi-permanent" electric dipole moment, due to an "inversion" phenomenon. In such molecules the type A levels indicated in Table II are split into two components, whose statistical weight depends on the type of statistics obeyed by the identical nuclei. The sum of the statistical weights for the two component levels is equal to the weight for a type A level given in Table III. The type E levels designated in Table II are also split by

TABLE III. Statistical weight factor g_I for molecules with a threefold axis of symmetry and a permanent electric dipole moment.

Rot-vib. species	One set of three identical nuclei with spin I	Two sets of three identical nuclei with spins I_1 and I_2
A	$1+\frac{2}{(2I+1)^2}$	$1+\frac{2}{(2I_1+1)^2(2I_2+1)^2}$
$\frac{1}{2}E$	$1-\frac{1}{(2I+1)^2}$	$1-\frac{1}{(2I_1+1)^2(2I_2+1)^2}$

inversion into two components. Although these components are each actually doubly degenerate, we assign each component the weight corresponding to only $(\frac{1}{2})E$ in Table III (instead of twice that value), so that an unresolved inversion doubling may be ignored.

For asymmetric top molecules, Cross *et al.*[70] have published a table of "line strengths" up to $J<13$. The squared elements of the direction-cosine matrix may be obtained from these line strengths ($\Sigma\phi$) by means of multiplying factors given in their paper. The multiplying factors for $M'=M$ are given below.

$$
\begin{aligned}
|D^{J+1,s',M}_{J,s,M}|^2 &= \left(\Sigma\phi\right)\frac{(J+1)^2-M^2}{(J+1)(2J+1)(2J+3)}\\
|D^{J,s',M}_{J,s,M}|^2 &= \left(\Sigma\phi\right)\frac{M^2}{J(2J+1)(J+1)}\\
|D^{J-1,s',M}_{J,s,M}|^2 &= \left(\Sigma\phi\right)\frac{J^2-M^2}{J(2J-1)(2J+1)}
\end{aligned}
\qquad (50)
$$

When no electric field is applied to the gas, the Stark components are superimposed. We then find, for example,

$$\sum_{M=-J}^{J} \frac{(J+1)^2 - M^2}{(J+1)(2J+1)(2J+3)} = \frac{1}{3}.$$

The other multiplying factors in eq. 50 satisfy the same sum rule. Thus, if the line strengths themselves are to be used in the above formula for calculating the total absorption coefficient due to all Stark components, an extra factor of $\frac{1}{3}$ should be included in the constant C.

3. Discussion of Absorption Coefficients and Line Intensities

At atmospheric pressure, the mean collision frequency may be of the same order of magnitude as the molecular resonance frequency. However, for pressures below a few millimeters of Hg, the mean collision interval τ is so large that the second term in F may be neglected. By substituting for N its perfect gas value $N = .9658 \times 10^{19} P(\text{mm.Hg})/T$, we have for each Stark component:

$$\gamma = 3.685 \times 10^{22} \mu^2 \nu^2 \frac{P}{w} \frac{S}{T^2} \frac{g_I e^{\frac{-W_{J'i'M'}}{kT}}}{pZ} |D_{JiM}^{J'i'M'}|^2 \tag{51}$$

where S = shape factor $= \dfrac{1}{1 + \left[\dfrac{\nu_{JiM}^{J'i'M'} - \nu}{w/2}\right]^2}$

$w = \dfrac{1}{\pi\tau}$ = width of absorption line at half maximum (in Mc),

ν = microwave frequency in Mc,

$\nu_{JiM}^{J'i'M'}$ = resonant frequency in Mc,

If we assume, tentatively, that the mean collision interval τ of eq. 47 is the same as the mean collision interval of eq. 46, the constant ratio $\dfrac{P}{w}$ may be related to a molecular collision cross section s.

$$\frac{P\ (\text{mm.})}{w\ (\text{Mc})} = .1553 \frac{\sqrt{MT}}{s(A^2)} \tag{52}$$

Thus the absorption coefficient at the resonant frequency for each Stark component is

$$\gamma_0 = 5.723 \times 10^{37} \frac{\mu^2}{s} \nu^2 \frac{\sqrt{M}}{\sqrt{T^3}} \frac{g_I e^{\frac{-W_{vJiM}}{kT}}}{pZ} |D^{J'i'M'}_{JiM}|^2 \tag{53}$$

where μ is the electric dipole moment in esu or the magnetic dipole moment in emu,

s is the effective collision cross section in square Angstrom units,

ν is the absorption line frequency in kmc,

M is the molecular weight of the gas,

$e^{\frac{-W_{vJiM}}{kT}}$ is the average Boltzmann factor for the two states.

For symmetric-top molecules we have, by substituting for $\sum_{M=-J}^{J} |D^{J'KM}_{JKM}|^2$ from eq. 49, and for the rotational part of the partition function from Section III, 4,

$$\gamma_0 = .4577 \mu^2 \nu^3 \frac{\sqrt{M}}{sT^{\frac{3}{2}}} \frac{g_I e^{\frac{-W_{vJ}}{kT}}}{Z_{\text{vib}}} \quad \text{(linear molecule)} \tag{54}$$

$$\gamma_0 = 1.271 \frac{\mu^2 \nu^3}{sT^3} \sqrt{\frac{M}{I_C}} \frac{g_I e^{-W_{vJK}}}{Z_{\text{vib}}} \left[1 - \frac{K^2}{(J+1)^2}\right] \quad \text{(symmetric top molecule)} \tag{55}$$

where μ is the dipole moment in Debye units (1 Debye unit $= 10^{-18}$ esu),

ν is the absorption line frequency in kmc,

I_C is the moment of inertia of the molecule about its axis of symmetry in mass units $\times$ Angstrom units^2 (1MUA^2 $= 1.660 \times 10^{-40}$ g.cm^2).

Z_{vib} is the vibrational partition function (see Section III, 4).

The absorption coefficient for asymmetric top molecules cannot be expressed so simply. However, the partition function for asymmetric top molecules has the same temperature dependence as that for symmetric top molecules. Therefore, for the lower rotational levels (for which the Boltzmann factor is essentially unity) the absorption coefficient varies inversely as T^3, provided that s is independent of temperature. For molecular collisions involving the long range dipole-dipole forces predominately, s is proportional to $T^{-\frac{1}{2}}$, and γ to $T^{-\frac{3}{2}}$.

The same dependence on temperature applies at low gas densities and power levels, where the line width is determined by the Doppler effect alone.

It may be seen from eq. (53) that the absorption coefficient is very low for molecules with small dipole moments. The accepted value for the electric dipole moment of N_2O is only $.16 \times 10^{-18}$ esu, compared with 1.86×10^{-18} esu for CH_3Cl.[70a] In the case of atoms, which have only *magnetic* dipole moments, the magnetic dipole strength is very low,

because the velocities of the valence electrons are usually much lower than the velocity of light. Thus the magnetic moment of the Cs atom in the lowest electronic state is 1 Bohr magneton, or $.9174 \times 10^{-20}$ emu.

On the other hand, the partition function for Cs at ordinary temperatures is only 2. This is an electronic partition function. As stated in the previous section, we are not including here the contribution of nuclear spin to the partition function. The reason that molecules with their (higher) electric dipole moments do not have much stronger absorption lines is that molecules at room temperature are divided among so many different rotational states. For a molecule such as SO_2 the rotational partition function is approximately 6,000 at room temperature. Inspection of equations 35 and 39 show that for larger molecules the partition functions will be very much larger. Thus absorption lines of the larger hydrocarbon molecules, which have large collision cross sections and moments of inertia, but small dipole moments, will be very difficult to detect.

Moreover, molecules which possess hyperfine structure may have their total absorption divided among a large number of lines. In general, a compound will be more difficult to detect when it possesses several nuclei with quadrupole moments, and when it can occur in several different isotopic modifications of comparable abundance.

For example, we may calculate the maximum absorption coefficient of the $J, K = 1, 0$ to $J, K = 2, 0$ transition of CH_3Cl (methyl chloride) from eq. 55 from the known values of I_B, I_C, and μ and by assuming an effective collision cross section of 1000 A^2. The pressure need not be known exactly, and we assume an absolute temperature of 300°. The frequency region in which this absorption line falls may be calculated from eqs. 5 or 17.

Since the three hydrogen nuclei have spins of $\frac{1}{2}$, the weight factor g_I is, from Tables II and III, for $K = 0$, $g_{(\frac{1}{2})} = \frac{1}{2}$.

The maximum absorption coefficient calculated from eq. 55 is then equal to 4.78×10^{-6} cm.$^{-1}$, for the sum of all isotopes and hyperfine structure components.

We now recall that there are two isotopes of chlorine, with approximate relative abundances of 3:1. This means that, for the most abundant isotope, the absorption coefficient must be multiplied by $\frac{3}{4}$.

Finally, the line possesses hyperfine structure, due to the quadrupole moment of the chlorine nucleus which has a spin of $\frac{3}{2}$. The rotational level $J, K = 1, 0$ is a triplet, while the level $J, K = 2, 0$ is a quartet. According to the tables given by Condon and Shortley (reference 21, p. 241), the relative strengths for the allowed transitions ($F = 0, \pm 1$) are:

	$F' = \frac{1}{2}$	$F' = \frac{3}{2}$	$F' = \frac{5}{2}$	$F' = \frac{7}{2}$
$F = \frac{1}{2}$	20.8	20.8		
$F = \frac{3}{2}$	4.2	26.7	52.5	
$F = \frac{5}{2}$	....	2.5	22.5	100.0

The strongest hyperfine structure component has .40 of the total CH_3Cl^{35} intensity. Thus the maximum absorption coefficient of the strongest hyperfine structure component of the transition $J, K = 1, 0$ to $2, 0$ in CH_3Cl should be approximately 1.44×10^{-6} cm.$^{-1}$

In some cases, the absorption coefficients may be made high enough to detect by raising the pressure until several lines overlap. For example, in a symmetric top molecule, the different frequencies $\nu_{J,K}^{J+1,K}$ (for fixed J) are approximately equal. When these lines overlap, it may sometimes be convenient to neglect the dependence of g_I on I, and to use an expression for the sum of the squared matrix elements.

$$\sum_{K=0}^{J} \sum_{M=-J}^{J} |D_{J,K,M}^{J+1,K,M}|^2 = \frac{(2J+1)(2J+3)}{9} \tag{56}$$

Equation 53 shows that the maximum absorption coefficient is independent of pressure over a large range of pressures. A quantity which is proportional to the gas density in this range is $\int \gamma_\nu d\nu$, which we may give the symbol "a" and call the "line intensity." Since $\int S d\nu = \frac{\pi w}{2}$ in this pressure range, the absorption line intensity is

$$a = \int \gamma_\nu d\nu = 5.788 \times 10^{-4} \mu^2 (\nu_{Ji}^{J'i'})^2 \frac{P}{T^2} \frac{g_I e^{-\frac{W_{vJiM}}{kT}}}{pZ} \sum_{M=-J}^{J} |D_{JiM}^{J'i'M'}|^2 \tag{57}$$

where a is in cm.$^{-1}$ Mc,

μ is the dipole moment in Debye units,

$\nu_{Ji}^{J'i'}$ is the resonant frequency in Mc,

P is the pressure in millimeters of mercury.

At very low pressures the *absorption coefficient* also decreases, until, at sufficiently low pressures and power levels, the line width is determined by Doppler effect alone, and the absorption coefficient is proportional to the gas density. This coefficient may be evaluated by substituting the value of "a" from eq. 57 into eq. 43. It will be found that in the Doppler effect limit, the temperature dependence of the maximum absorption coefficient is the same as for collision broadened lines at temperature equilibrium.

4. The Effect of High Microwave Energy Density in the Absorption Cell

Microwaves passing through a gas induce both absorption and emission. Since the induced emission is coherent with the original wave train, it is effectively substracted from the induced absorption. This effect has already been taken into account approximately in eq. 47. The net absorption observed at temperature equilibrium is due to the slightly different Boltzmann factors for the upper and lower levels. At 300°K. the ratio between the Boltzmann factors for the upper and lower states involved in a transition is $e^{-\frac{\nu(\text{kmc})}{6252}}$, where ν is the absorption line frequency.

Thus even at a frequency of 100 kmc, the ratio between the populations of the upper and lower levels is more than .98, and the net absorption at thermal equilibrium is less than 2% of the total stimulated absorption.

The effect of high microwave energy densities is to reduce even this difference in population, particularly at low gas pressures where intermolecular collisions are insufficient to maintain thermal equilibrium. An additional effect of the induced emission and absorption is to broaden the absorption line (see Section IV, 1) and possibly to further reduce the average absorption coefficient. These power distortion or "saturation" effects have been discussed by Townes,[60] Pond and Cannon,[61] Swinger and Karplus,[62] Carter and Smith,[63] Karplus,[71] and others.

According to Karplus and Schwinger, the shape factor F in eq. 47 should be changed at low gas pressures to

$$F_M = \frac{1}{\pi}\left\{\frac{\frac{1}{2\pi\tau}}{(\nu - \nu_{J'i'M}^{J'i'M})^2 + \frac{1}{(2\pi\tau)^2} + \frac{E^2\mu^2}{h^2}|D_{JiM}^{J'i'M}|^2}\right\} \tag{47a}$$

where E is the peak value of the electric vector of the microwave, considered to be traversing free space and τ is the mean interval between collisions. $\mu|D_{JiM}^{J'i'M'}|$ is the matrix element of the electric dipole moment.

Accordingly, when $\frac{E^2\mu^2}{h^2}|D_{JiM}^{J'i'M}|^2$ becomes comparable with $\frac{1}{(2\pi\tau)^2}$, the absorption coefficient at the center of an absorption line will begin to decrease.

The derivation of eq. (47a) rests on Van Vleck and Weisskopf's assumption of sharp collisions, and makes no distinction between collisions that broaden the line and collisions that produce thermal relaxation. Other investigators have suggested that the thermal relaxation time may be somewhat longer than τ, but Karplus[71] has pointed out that eqs.

(47) and (47a) together are in agreement with the careful measurements of Bleaney and Penrose on the inversion spectrum of ammonia.

Spontaneous transitions, which are important in spectroscopy of the visible and ultraviolet, are unimportant in microwave spectroscopy unless, possibly, when the molecules are in very highly excited vibrational or electronic states.

V. Apparatus

1. Simple Microwave Spectroscope

The essential components of a microwave spectrograph are: a microwave generator, with means for determining its frequency of oscillation,

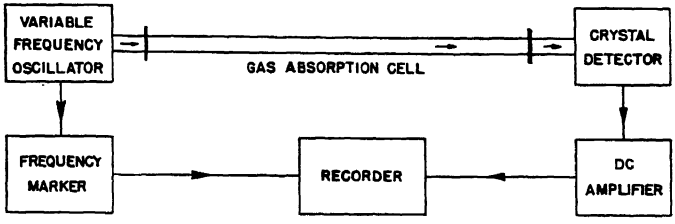


Fig. 6.—Block diagram of simple microwave spectrograph.

an absorption cell consisting of a section of waveguide which can be evacuated and filled with the gas to be studied, and a microwave detector. A block diagram showing the arrangement of these components is given in Fig. 6, while a photograph of the microwave components is given in Fig. 7.

The reflex Klystron oscillator shown in Fig. 7, which was manufactured by Western Electric, can be adjusted to oscillate at any frequency between 22 kmc and 25 kmc. A series of reflex Klystrons is manufactured by Raytheon to cover the complete range between 6 and 18 mm.

The variable attenuator adjacent to the oscillator serves the double purpose of controlling the power output to the absorption cell and of absorbing any energy reflected from the other end of the absorption cell. A crystal detector is shown. The vertical screw adjustments adjacent to the crystal mount are used to reduce reflections from the crystal end of the waveguide to a minimum.

When the cavity wavemeter is tuned to the oscillator frequency, it absorbs a few per cent of the energy from the waveguide. After tuning to resonance, the frequency of the oscillator is known from the reading of the micrometer screw. Various microwave components, including frequency measuring apparatus, have been described in Volume 11 of

the M.I.T. Radiation Laboratory Series.[72] Apparatus and techniques useful in the region of 3 to 5 mm. have been discussed by Smith *et al.*[73]

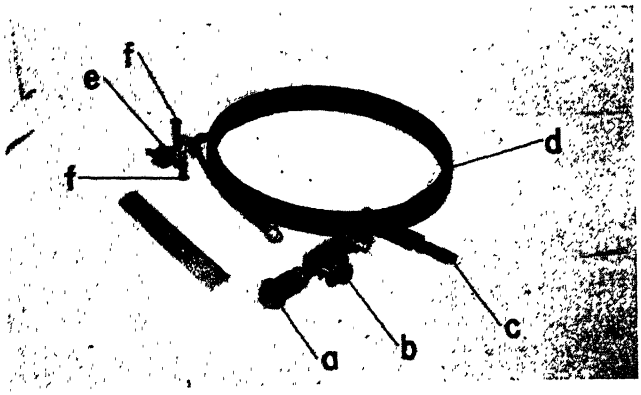


FIG. 7.—Photograph of microwave components.
a—Microwave generator
b—Variable attenuator
c—Cavity wavemeter
d—Waveguide absorption cell
e—Impedance matcher
f—Crystal detector

Instead of recording the absorption spectrum, one may display a small part of the spectrum at a time on the screen of a cathode-ray oscilloscope. A block diagram for this type of spectroscope is shown in Fig. 8.

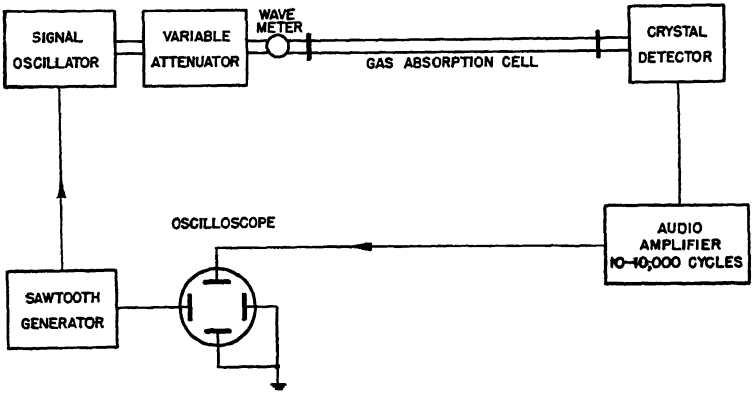


FIG. 8.—Microwave spectroscope using frequency sweep.

The frequency of the oscillator is swept in step with the horizontal spot on the oscilloscope screen, at any convenient frequency from, say, 120 cps to a fraction of a cycle per second. At the lower sweep frequencies

it is convenient to use an oscilloscope with a long persistence screen. Figure 9 shows a photograph of such an oscilloscope pattern, taken by this method.

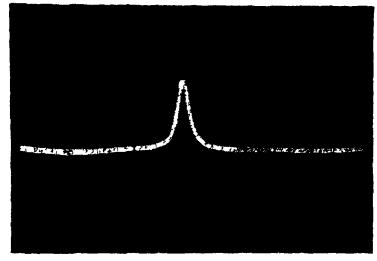


FIG. 9.—Microwave absorption line. A photograph of the oscilloscope screen indicated in Fig. 8. Intensity of transmitted power is indicated vertically downward, the frequency of the microwave oscillator is indicated horizontally, and increasing to the right.

2. The Signal Due to an Absorption Line

When the amount of microwave power traversing the absorption cell is extremely small, and no local oscillator is used at the detector, then the signal output from the detector (due to an absorption line) is proportional to the amount of microwave *power* absorbed by the gas. However, if a local oscillator is used, or if the amount of microwave power traversing the cell is sufficient to ensure a reasonable power conversion gain in the microwave detector, then it is convenient to consider separately the conversion gain at the detector and to think of the *signal due to an absorption line* as the change in microwave amplitude due to absorption by the gas.

The power absorption coefficient is defined in eq. 42. The coefficient for *amplitude* attenuation (in nepers/cm.) is equal to one half the coefficient (in cm.$^{-1}$) for power absorption.

$$A = A_0 e^{-\frac{1}{2}(\gamma_s+\gamma_w)l} \quad (58)$$

where A_0 = amplitude of microwave (in watts$^{\frac{1}{2}}$) entering the absorption cell

A = amplitude of microwave (in watts$^{\frac{1}{2}}$) leaving the absorption cell

γ_s = power absorption coefficient for a plane wave (in cm.$^{-1}$) due to the sample gas

γ_w = effective power absorption coefficient for a plane wave due to losses in the waveguide walls etc.

l = effective length of the absorption cell (in centimeters) for a plane wave.

The effective length of a waveguide is greater than the shortest distance between its two ends, because of the zig-zag motion of the plane wave as it is propagated down the waveguide. For the dominant mode of propagation in rectangular waveguide, the effective length l is related to the physical length L by the relation

$$l = \frac{L}{\left[1 - \frac{\nu_c^2}{\nu^2}\right]^{\frac{1}{2}}} \tag{59}$$

where ν_c = cut-off frequency for the waveguide, below which the microwaves will not be propagated.

In a superheterodyne type of receiver, the signal amplitude (or power) output from the detector is directly proportional to the signal amplitude (or power) input, even though it may be at a lower frequency. If one then speaks of the amplitude gain $\sqrt{G}$ of the detector as the ratio of the square root of the output power to the square root of the available microwave input power, one has for the signal amplitude output

$$A_{out} = \sqrt{G}\, A_0 e^{-\frac{1}{2}\gamma_w l}(1 - e^{-\frac{1}{2}\gamma_c l}) \tag{60}$$

If a superheterodyne receiver contains a crystal detector and a (microwave frequency) local oscillator, the amount of local oscillator power reaching the detector can usually be adjusted to give an amplitude conversion gain $\sqrt{G}$ as high as .5, the optimum amount of microwave power depending on the particular crystal contact.

Instead of using a local oscillator in the receiver, it is possible to generate sidebands in the primary oscillator (by frequency or amplitude modulation) so that one sideband frequency can be slowly swept over an absorption line, while the carrier frequency traverses the same (or a different) path and beats against the sideband in the detector. The absorption of frequency modulated microwave signals has been discussed theoretically by Townes and Merritt,[74] Karplus,[75] and others. A frequency modulation system with a modulation frequency several times as large as an absorption line width has been used by Strandberg and collaborators.[76]

When only small amounts of power are available, it may be impossible to find a detector sufficiently sensitive to operate at an amplitude conversion gain as high as .5. In the limit of low microwave powers, the conversion gain of a square law detector is proportional to both the signal frequency amplitude, and to the carrier amplitude, so that if both originate in the same (signal) oscillator, the output signal from the detector is proportional to the power generated by this oscillator.

At the lower microwave frequencies, the absorption coefficients tend to be rather low. However, at these frequencies, the amount of microwave power available is comparatively high, so that a high conversion gain can be assured. For example, in the region of 24 kmc, the signal amplitude due to an absorption line may be only 10^{-5} watts$^{\frac{1}{2}}$, but the

carrier amplitude may be over .02 watt$^{\frac{1}{2}}$ which is sufficient to obtain an amplitude conversion gain of .5 in a good 1N26 crystal.

In order to raise the signal level due to a weak absorption line, it may be desirable to send a large microwave amplitude through the absorption cell. In this case, the unabsorbed amplitude may be sufficient to overload a sensitive crystal detector. The remedy for this is to use some method of amplitude concelling, such as a microwave-frequency bridge, or to use a detector which is not so easily overloaded, such as a less sensitive crystal or a bolometer.

Bolometers give a high conversion gain only when sufficient microwave signal power is available to raise the detecting element considerably above the surrounding temperature.

Thermocouples are simple heat engines, so that they cannot possibly operate with a conversion gain higher than that of a perfect heat engine operating between the temperatures of the active and passive thermojunctions.

Infrared detectors of the Golay pneumatic type are of course subject to this limitation on conversion gain, but the pneumatic action is followed by a highly efficient photoelectric amplifier incorporated into the instrument. Such detectors have been used by Loubser and Townes for frequencies up to 200 kmc.[77]

In order to minimize the effect of thermal drifts with thermal detection, it is desirable to employ some method of low frequency modulation, with a tuned AC amplifier. Such a detection system has been used effectively by Weingarten.[78]

3. Sources of Noise

The small fractional change of microwave power due to an absorption line may be difficult to detect because of erratic variations of the generator frequency and amplitude with time. This is called generator noise. The generator noise-power at the detector is proportional to the amount of power allowed to reach the detector, and may amount to several times 10^{-5} of this microwave power for frequencies in the audio range. Klystron noise is discussed in Volume 7 of the M.I.T. Radiation Laboratory Series.[79]

If the sensitivity of detection of an absorption line is limited by generator amplitude fluctuation, then eq. 56 shows that the length of the absorption cell should be made as large as possible. Another method of reducing the effects of generator amplitude fluctuations is to use a microwave frequency bridge such as that shown in Fig. 3. However, both of these methods tend to increase the frequency dependence of the signal

reaching the detector, so that erratic frequency modulation at the generator has a greater chance of being converted into amplitude modulation.

Other possible sources of noise are the crystal rectifier and the amplifier. Crystal and receiver noise measurements have been discussed in Volumes 15 and 23 of the M.I.T. Radiation Laboratory Series.[80,81] The noisepower available from an ordinary resistor at a temperature T in a frequency band df is

$$dP_N = kTdf \tag{61}$$

At 300°K. (room temperature) this is equal to 4.14×10^{-21} watts per cycle of bandwidth.

The noise-power output dP_N from the crystal may be defined with reference to the Johnson noise-power available from an equivalent resistance at room temperature by means of a factor t called the noise-temperature ratio.

$$dP_N = kTt(V, p, Z, f)df = 4.14 \times 10^{-21}t(V, p, Z, f)df \tag{62}$$

The noise-temperature ratio depends on (among other things) any biasing voltage V applied to the crystal, on the microwave power p available at the crystal, on the load impedance Z which couples the crystal to the amplifier, and on the frequency f to which the amplifier is tuned. The crystal is assumed to be matched as nearly as possible to the waveguide.

The noise-temperature ratio may approach unity at a frequency of 30 Mc, but below 300 kc it increases approximately as the inverse of the frequency to which the amplifier is tuned.[80]

If we assume that, for detection of an absorption line, the converted signal amplitude must be at least as large as the noise amplitude $dA_N = (tkTdf)^{\frac{1}{2}}$, then the minimum detectable absorption coefficient is given by

$$A_{out} = \sqrt{G}\, A_{in} = \sqrt{G}\,[A_0 e^{-\frac{1}{2}\gamma_w l}(1 - e^{-\frac{1}{2}\gamma_s l})] = (tkTdf)^{\frac{1}{2}} \tag{63}$$

Usually, the absorbed fraction $1 - e^{-\frac{1}{2}\gamma_s l}$ of the microwave amplitude is very small, and practically equal to $\frac{1}{2}\gamma_s l$. Thus, at room temperature, the minimum detectable absorption coefficient is

$$\gamma_s\ (\text{cm.}^{-1}) = \frac{2\sqrt{ktTdf}}{l\sqrt{G}\, A_0 e^{-\frac{1}{2}\gamma_w l}} = \frac{1.248 \times 10^{-10}\ \text{joules}^{\frac{1}{2}}}{A_0 e^{-\frac{1}{2}\gamma_w l}\text{watts}^{\frac{1}{2}}} \times \frac{(df)^{\frac{1}{2}}\ \text{sec.}^{-\frac{1}{2}}}{l\ \text{cm.}} \times \left(\frac{t}{G}\right)^{\frac{1}{2}} \tag{64}$$

Essentially this equation has been given by Townes and Geschwind[82] and by Gordy.[11] Values of T as low as unity may be attained by using crystals with IF frequencies of 10 Mc or higher, or by using bolometer elements with a thickness of 1 micron, or perhaps much less. Values of G as high as $\frac{1}{4}$ may be obtained with either crystals or bolometers, but in the case of bolometers the power necessary for this conversion efficiency (of the order of a milliwatt) must be obtained from the signal oscillator, since the slow response of the bolometer (usually less than 1 kc) makes a local oscillator impractical. To avoid thermal drifts, a low frequency modulation method should be employed with bolometer detectors.

Equation 64 indicates that it is desirable to make A, the microwave amplitude leaving the absorption cell, as large as possible. The upper limit on A may be set by:

(1) The microwave power available at the generator.
(2) Power distortion or "saturation" effects in the gaseous absorption cell.

At wavelengths near 1 cm., the power output of commonly used reflex Klystrons may be in the range 1–50 milliwatts. However, the conveniently available power may be less than this, because it is desirable to insert several decibels of attenuation or "padding" between the generator and the absorption cell, in order to absorb microwave energy reflected from the far end of the absorption cell or the detector, and in order to keep the generator from being "pulled" into resonance with the absorption cell. The power limitation is likely to be rather serious when higher harmonics of an oscillator are used to obtain frequencies higher than those obtainable from microwave oscillator directly.

The second microwave power limitation (power distortion) can often be avoided by maintaining a rather high gas pressure in the absorption cell. This ensures thermal equilibrium and minimizes the effect of microwave-induced transitions on the absorption line width. However, when the compound being studied does not have a sufficient vapor pressure at the desired absorption cell temperature, or when sharp absorption lines are desired, then power distortion or saturation may be a serious limitation.

In the two cases above of microwave power limitation, the optimum effective length l of the absorption cell is approximately

$$l = \frac{2}{\gamma_W} \tag{65}$$

where γ_w is the effective absorption coefficient of the waveguide used, due to conductor losses in the sides of the waveguide. If a separate local oscillator is not used, the optimum effective length of the absorption cell may be as low as $l = \frac{1}{\gamma_w}$.

In order to reduce the conductor losses, the waveguide should be fabricated of high conductivity material—preferably copper or silver. The attenuation may be further reduced by using waveguide of larger lateral dimensions. The use of waveguide of large cross section has the additional advantage of lowering the microwave energy density (for a given microwave power) and of thus reducing power distortion in the absorption cell.

4. Difficulties Due to Standing Waves

When either of the internal lateral dimensions of a rectangular waveguide is greater than λ (where λ is the free-space wavelength of a microwave) then the microwave may be propagated along the waveguide in more than one "mode." Microwave energy which is converted into these other modes of propagation may, because of multiple reflections, be strongly reflected or absorbed in the absorption cell at certain frequencies, thus producing a spurious "absorption spectrum."

Even at frequencies such that only one mode of propagation in the waveguide is possible, there is usually present in the waveguide a small standing wave pattern which is very frequency-sensitive. The resulting undulating baseline of the absorption spectrum may be noted in Fig. 9. This phenomenon is due to reflections from absorption cell windows and other irregularities in the waveguide.

The absorption cell windows are usually made of .001 or .002-inch mica sheet, cemented to the ends of the waveguide section. Waveguide windows which are non-reflecting over a broad frequency range have been described.[83] However, in spite of all attempts to keep the inside of the waveguide smooth, and to use nonreflecting attenuators, it appears to be impossible to completely eliminate the standing wave pattern over a reasonable range of frequencies.

When high amplifier gain is being used to detect weak absorption lines, the absorption lines appear as small "pips" on the sides of very steep waves. The tendency to use larger and longer absorption cells (in order to increase the sensitivity of detection) has the adverse effect of increasing the frequency sensitivity of the standing wave pattern. Under some conditions the "virtual absorption spectrum" due to the apparatus almost completely hides absorption lines due to the gas.

5. *Stark Effect Modulation*

The best solution to the above difficulty which has been found to date is a method described by Hughes and Wilson,[84] in which the gaseous absorption line is periodically shifted in frequency by means of an electric field applied to the gas. The electric field does not appreciably influence the virtual absorption pattern due to the apparatus, so that this is no longer observed. This "Stark" effect has been discussed in Sections II, 5 and IV, 2.

In practice, a sinusoidal voltage may be applied to a central electrode in the waveguide at a frequency from several kilocycles to several hun-

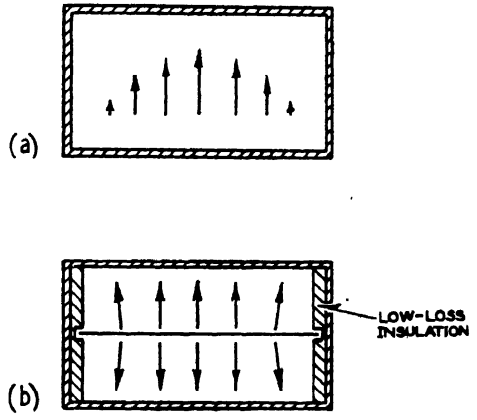


FIG. 10.—Cross sectional view of waveguide.

dreds of kilocycles. The resulting frequency modulation of the absorption line produces an amplitude modulation of the microwaves, which is detected by means of a tuned amplifier in the crystal circuit.

A commonly used arrangement of the internal electrode in the waveguide is illustrated in Fig. 10b. With this arrangement the low frequency electric field is approximately parallel to the smaller lateral dimension of the waveguide. The electric field distribution of the microwave is shown in Fig. 10a; the DC field in Fig. 10b. It may be seen from these sketches that:

(1) The electric field distribution of the microwave in the gas will not be seriously affected by the presence of the central electrode.
(2) The inhomogeneity of the DC or low frequency field will be serious only at the sides of the waveguide, where the microwave field is rather weak anyway.

The amplifier should be tuned to twice the frequency of the oscillator. An alternative method is to impose a constant electric field on the gas, in addition to the sinusoidally varying field, thereby producing a component of crystal current at the fundamental frequency of the oscillator. The Stark effect of a molecule may be conveniently studied by using a constant electric field considerably larger than the sinusoidal field. Under these conditions a much smaller sinusoidal field may suffice to shift the absorption line the same fraction of its width. A block diagram of a system employing Stark effect modulation is given in Fig. 11.

W. D. Hershberger[85] has examined the conditions under which maximum sensitivity can be obtained, assuming a line width limited by

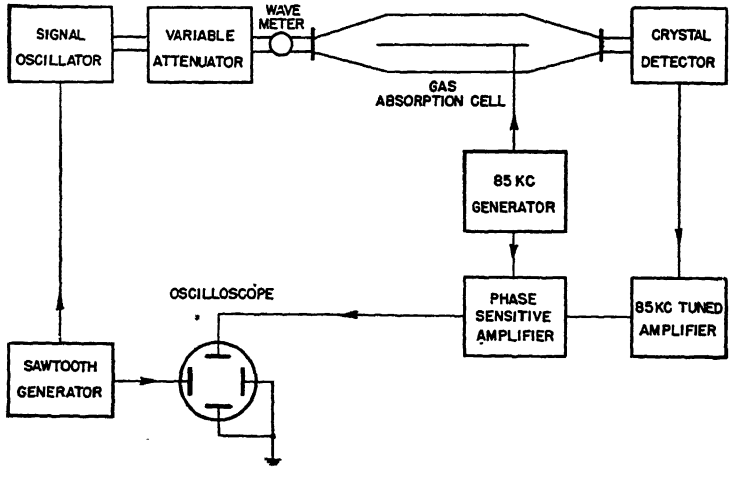


FIG. 11.—Absorption line modulation by means of the Stark effect.

pressure broadening, a single Stark component, and a square-law detector. He finds that the detector output is a maximum when the maximum Stark shift in frequency is twice the half-maximum line width. Under these conditions he finds that the absorption line component appears to be shifted 1.36 line widths, and that the total change in detector output during a cycle is equal to that which would be observed by a DC method. However, most absorption lines have several Stark components, and the amplitude modulation due to one component may be cancelled by that due to other components. This is particularly true for a linear Stark effect, so that no signal is observed at the fundamental frequency unless a DC bias is superimposed on the sinusoidal field.

Instead of using an electric field varying sinusoidally with the time, one may use a square wave based on zero voltage, together with an

amplifier tuned to the repetition rate of the square wave. If a good square can be generated, it is possible to obtain essentially 100% modulation without appreciably widening the component lines. The theory of frequency modulation given in references 73 and 75 applies also to modulation by means of the Stark effect.

Schemes in which the Klystron frequency is modulated the order of a line width or less (instead of the absorption line being modulated) have been described by Hersberger,[85] Gordy and Kessler,[86] and Watts and Williams.[87] Klystron frequency modulation is rather easily accomplished, but it does not eliminate effects due to standing waves. The two systems have been combined very effectively by Strandberg *et al.*[76]

6. Resonant Cavity Absorption Cell

Instead of a long section of waveguide designed to produce very small reflections at both ends, an absorption cell may consist of a relatively short section of waveguide terminated in such a way as to reflect almost completely at its two ends, so that the microwave are reflected back and forth hundreds of times, and the total distance travelled by the microwave train before reaching the detector is very high. The microwave frequency is slowly swept over a frequency region that includes an absorption cell resonance. This resonant cavity type of absorption cell is very well suited for studying gases in a magnetic field, very rare gases, or highly radioactive gases.

Hershberger[85] has shown that the ultimate sensitivity of detection of weak absorption lines is the same for a resonant cavity absorption cell of this type as for a long waveguide absorption cell of the same lateral dimensions. He finds that the minimum detectable gaseous absorption coefficient is proportional to the effective absorption coefficient of the waveguide, due to conductor losses and solid insulation losses.

The deduction of eq. 64 assumed that the power in the absorption cell is not high enough to produce power distortion effects. In the case of a resonant cavity, the microwave energy reflected back and forth across the absorption cell may build up to a rather high energy density, so that power distortion effects are much more likely to occur in a resonant cavity of small volume than in a large waveguide absorption cell. These distortion effects are proportional to the energy density, to μ^2, and to the squared elements of the direction cosine matrix.[62] Thus a fairly large cell should be used for studying a molecule such as methyl chloride, while a small resonant cavity may be quite satisfactory for studying molecules with small electric dipole moments, or magnetic dipole moments. When an absorption "line" possesses a narrow hyperfine structure, power distortion tends to decrease the intensity of

the strong components of the line, while leaving the weak components unaffected. This sometimes provides a convenient test for the presence of power distortion.

7. Attenuation in Waveguide Due to Conductor Losses

The attenuation of guided waves, due to conductor losses, is treated in textbooks on microwave engineering; for example that by Ramo and Whinnery.[88] The (amplitude) attenuation constant γ_w for microwave propagated in the TE_{10} mode of empty rectangular waveguide is in *MKS* units

$$\gamma_w \text{ (nepers/meter)} = \frac{R_s}{120\pi} \frac{\left[\frac{1}{b} + \frac{2}{a}\left(\frac{\nu_c}{\nu}\right)^2\right]}{\left[1 - \left(\frac{\nu_c}{\nu}\right)^2\right]^{\frac{1}{2}}} \tag{66}$$

where R_s = skin effect surface resistivity in ohms of the waveguide metal
$= 2.61 \times 10^{-7} \sqrt{\nu}$ for copper,
a = lateral internal dimension of the waveguide *perpendicular* to the electric vector of the microwaves (in meters),
b = lateral internal dimension of the waveguide *parallel* to the electric vector of the microwaves (in meters),
ν = microwave frequency in cycles/sec.,

$$\nu_c = \frac{c}{2a} = \text{cut off frequency in cycles/sec.}$$

In the waveguide commonly used $\left(b = \frac{a}{2}\right)$ the TE_{10} mode is called the dominant mode, since microwave energy can be propagated in this mode at lower frequencies than is possible in any other mode.

The *effective power absorption coefficient* for the TE_{10} mode of empty rectangular waveguide, corresponding to the power absorption coefficient for gases defined in equations 42, 51, 53 is $2\left[1 - \left(\frac{\nu_c}{\nu}\right)^2\right]^{\frac{1}{2}}$ times the amplitude attenuation constant γ_w.
That is:

$$\gamma_w \text{ (meter}^{-1}) = \frac{R_s}{60\pi}\left[\frac{1}{b} + \frac{2}{a}\left(\frac{\nu_c}{\nu}\right)^2\right] \tag{67}$$

or, for copper waveguide,

$$\gamma_w \text{ (cm.}^{-1}) = 1.39 \times 10^{-5}\nu^{\frac{1}{2}}\text{(Mc)}\left[\frac{1}{b \text{ (mm.)}} + \frac{4.5 \times 10^{10}}{a^3 \text{ (mm.)}\nu^2\text{(Mc)}}\right] \tag{68}$$

It is apparent from this equation that, for minimum conductor losses,

both lateral dimensions should be as large as possible. The effective absorption coefficient per cm is plotted in Fig. 12 for several sizes of copper waveguide. For other metals the plotted quantities should be multiplied by $\left(\frac{\mu\sigma_0}{\mu_0\sigma}\right)^{\frac{1}{2}}$, where μ, σ are the permeability and conductivity of the other metal, and μ_0, σ_0 are the permeability and conductivity of copper.

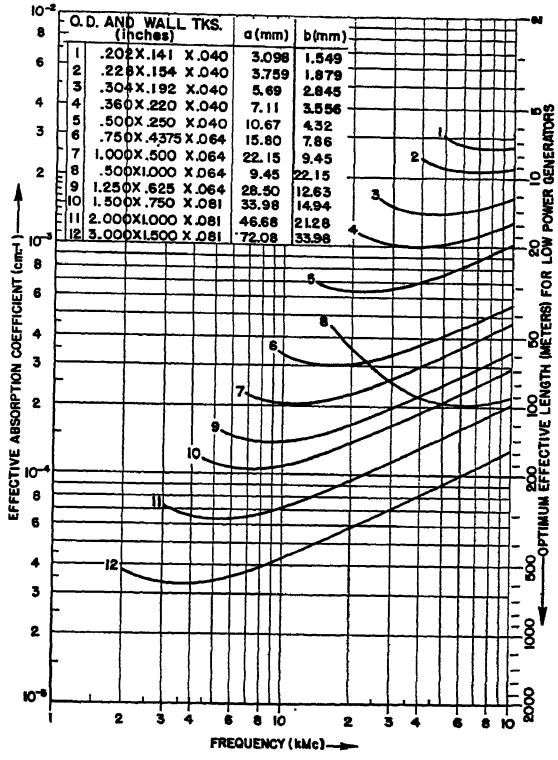


FIG. 12.—Effective power absorption coefficient of rectangular copper waveguide. The effective absorption coefficient is shown for the T_{10} mode of propagation, in which the electric vector of the microwave is parallel to the lateral dimension b. The low frequency termination of each curve is the cut-off frequency for the T_{10} mode for each size of waveguide. The internal dimensions a and b are given in millimeters while the outer dimensions and wall thickness are given in inches.

At high frequencies and a fixed waveguide size, α_w approaches the asymtotic value

$$\gamma_w\,(\text{cm.}^{-1}) = \frac{7.60 \times 10^{-3}}{b\,(\text{mm.})\,\lambda^{\frac{3}{2}}\,(\text{mm.})} \quad \begin{matrix}(\text{copper waveguide,}\\ \text{high frequencies})\end{matrix} \qquad (69)$$

where γ is the free space wavelength.

It may be seen from eq. 68 that the attenuation due to losses in the walls of the waveguide is higher at the shorter wavelengths, especially if the waveguide dimension is decreased along with the wavelength. At the shorter wavelengths, however, the absorption coefficient of the gaseous absorption lines increases even more than that of the waveguide, so that shorter absorption cells can be used satisfactorily.

The effective absorption coefficient for waveguide containing a Stark electrode will be somewhat greater than that given above. The Stark electrode should be made of copper, preferably. The insulation may consist of Teflon tape.

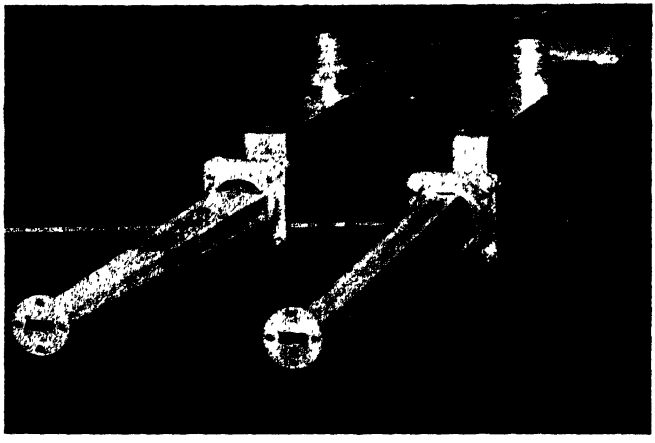


FIG. 13.—Transitions to waveguide absorption cells of large cross section. The small end of the taper fits a size of waveguide in which only the dominant mode can be propagated. The electric vector of the microwave is here in a vertical plane. Pumping and introduction of gas is through a long narrow slot in the center of the uppermost side.

In Fig. 13 are shown the ends of two waveguide absorption cells which have been used for the wavelength range 9–17 mm. The absorption cell on the right side of Fig. 13, which has a dimension parallel to the electric vector of 22 mm., gives the lower attenuation for frequencies above 28 kmc. This may be seen in Fig. 12 by a comparison of curves 7 and 8. It has the further advantage that higher modes of propagation may be attenuated by means of slots cut in the center of the narrow side of the waveguide. Absorption line broadening, due to collisions of the molecules with the waveguide, is slightly smaller for this absorption cell on the right; but in either case the effect would be observable only at low pressures and in the wings of the absorption line. The absorption cell on the left, with the 9.45 mm. dimension parallel to the electric vector of the microwave, has the lower attenuation below 28 kmc and is better suited

for Stark effect modulation. An even larger size of waveguide would be desirable for use with high microwave power, in order to avoid power saturation.

8. *Untuned-Cavity Absorption Cell*

A different line of attack was successfully taken by Becker and Autler,[7] who built an absorption cell in the shape of an 8-foot cube. Their cavity was not tuned to any particular resonant mode, and in a frequency interval corresponding to a (rather broad) absorption line, many modes of resonance could exist simultaneously. The problem in this case is to be sure that there are a large enough number of resonant modes and places for introducing and sampling the power, so that a statistical interpretation of the results is possible. It appears that, by using an absorption cell with a capacity of several tens of cubic meters, and high-power microwave generators, that high resolution and very high sensitivity may be simultaneously achieved.

This type of apparatus has been used by Weingarten[78] to study the absorption spectrum of ammonia at high pressures. In order to distribute the modes uniformly in space and frequency, he used mechanical "mode-mixers." The microwave energy density was sampled by a well distributed array of bolometer wires. Weingarten found bolometers to give a very satisfactory conversion gain at the power levels he used.

9. *Minimum Detectable Absorption*

The noise-temperature ratios for crystals vary markedly from one crystal to another, and they vary with the frequency to which the amplifier is tuned.[80] By using a heterodyne system of detection, with a microwave frequency local oscillator, and an intermediate frequency of 30 Mc, it is possible to achieve crystal noise-temperature ratios of approximately unity. This is commonly done in radar receivers. The power conversion gain of the crystal converter is often in the range $\frac{1}{2}$ to $\frac{1}{4}$. Superheterodyne receivers of this type have been used in several laboratories for studying microwave absorption lines.[61] Low frequency Stark effect modulation may be used to eliminate standing waves and the low frequency noise from the primary microwave generator.[76]

Although bandwidths of several kc are often used when viewing absorption lines on an oscilloscope screen, it is possible to reduce the bandwidth to 1 cps or less, while rather slowly recording the absorption spectrum. Phase sensitive amplifiers have been effectively used for feeding the narrow band recorder.

In order to obtain a figure for a minimum detectable absorption coefficient, let us assume a power conversion gain of $\frac{1}{4}$, a noise tempera-

ture ratio of unity, a bandwidth of 1 cps and an effective absorption path of 20 meters. Then if the gas pressure and the waveguide cross section are sufficient to prevent power distortion and to transmit microwave power of .4 milliwatt at the crystal, the minimum detectable absorption coefficient calculated from equation 64 is 6.4×10^{-12} cm.$^{-1}$ However, the writer does not at the present time know of any *detected* absorption lines with coefficients as low as 10^{-10} cm.$^{-1}$

Equation 64 shows that in order to push the minimum detectable absorption coefficient to its limit, the microwave power should be increased and the absorption cell lengthened. One soon reaches a condition, then, where the cross sectional area of the absorption cell must be increased in order to avoid excessive attenuation and power saturation. Thus it appears that the ultimate sensitivity for weak absorption lines may be attained in a large untuned-cavity absorption cell, such as that described in the previous section. Such a cell would require a rather large gaseous sample.

When trying to detect absorption lines in minute samples of gas, it at first appears desirable to decrease both the size of the absorption cell and the gas pressure, and, of course, to use a considerable amount of microwave power. In this case, power saturation will tend to greatly reduce the absorption coefficient of the line.

Equation 64 suggests that a limit of detectability is set on the product γA. If we therefore minimize the product γA, as given by eqs. 47 and 47a (Section IV, 4), with $\nu = \nu_{JiM}^{J'i'M'}$, and τ considered to be a constant, we find that in rectangular waveguide with negligible conductor losses, the optimum microwave power is given by

$$A^2 = 0.654 \times 10^{-3} S \frac{\lambda}{\lambda_g} \frac{w^2}{\mu^2 |D_{J\gamma M}^{J'\gamma' M}|^2} \tag{70}$$

where A^2 is the microwave power in watts,
S is the cross sectional area of the cell in cm.^{2},
λ is the free space wavelength,
λ_g is the wavelength in the waveguide,
w is the line width in Mc, and
$\mu|D_{J\gamma M}^{J'\gamma' M}|$ is the dipole moment in esu.

With this value of the microwave power, the peak absorption coefficient is reduced to 0.585 of its value at low power levels. By combining this relation with eqs. 47, 64, and 70, one finds that the minimum detectable number of molecules is given approximately by the relation

$$(NV) \frac{g_I e^{-W_{JiM}}}{pZ} = \frac{5.18 \times 10^7}{\nu^2} \sqrt{df} \sqrt{S} \sqrt{\frac{t}{G}} \frac{T}{\mu |D_{JiM}^{J'i'M}|} \tag{71}$$

where V is the volume of the sample in cm.^{3},
(NV) is the total number of molecules in the sample,
df is the band pass in sec.$^{-1}$,
ν is the microwave frequency in kmc,
S is the cross-sectional area of the waveguide in cm.^{2},
t is the noise-temperature ratio of the detector, and
G is the power conversion gain of the detector.

This equation says nothing about the gas pressure, but indicates that the cross-sectional area of the absorption cell should be small. If we then introduce the limitation on cross section imposed by conductor losses in the walls, we find the results

$$S^{\frac{1}{2}} = 0.06 V^{\frac{1}{3}} \nu^{\frac{1}{3}} (\text{kmc}) \tag{72}$$

From this result and assuming a single Stark component, we find that the minimum detectable number of molecules is

$$(NV)_{\min} = 4.5 \left[\frac{s}{w}\right]^{\frac{1}{2}} \left[\frac{pZ}{g_I e^{-W_{vJiM}} \mu |D_{JiM}^{Ji'M'}|}\right]^{\frac{1}{2}} \left[\frac{T^7 (df)^3 t^3}{\nu^{11} M G^3}\right]^{\frac{1}{4}} \tag{73}$$

where w is the line width in Mc,
s is the collision cross section in A^2,
M is the molecular weight,
ν is the absorption frequency in kmc,
T is the absolute temperature,
df is the badwidth in sec.$^{-1}$,
p, Z, $g_I e^{-W_{vJiM}}$ are explained in Section III, 4,
t, G are explained in Section V, 2 and V, 3,
$\mu |D_{JiM}^{Ji'M'}|$ is the matrix element of the dipole moment in Debye units.

In the case of pure OCS gas at 200°K., with a line frequency of 24 kmc, a line width of 1 Mc, a bandwidth of 1 sec.$^{-1}$, a noise figure of 4, and a collision cross section of 250 A.^{2}, it should be possible to detect 10^7 molecules. However, if it were necessary to detect $OC^{13}S$ molecules constituting only 1% of the sample, then the total sample would have to be 1000 times larger.

10. Precision Frequency Measurements

The absorption wavemeter shown in Figs. 6 and 7 is a very simple and convenient method of measuring frequencies to about one part in 5000. Wavemeters are described in reference 72. A general discussion of microwave frequency measurements has been given by Gordy.[11]

The precision frequency measuring equipment used at Westinghouse is shown in Figs. 14 and 15. The secondary frequency standard used is a crystal controlled oscillator. This is adjustable to a frequency of 5,000,000.0 ± .1 cps by zero-beating it against a standard frequency signal broadcast from the National Bureau of Standards Station WWV in Washington, D.C.

This 5 Mc frequency is multiplied by a factor of one or two hundred by a process of frequency quintupling and doubling, using high-frequency

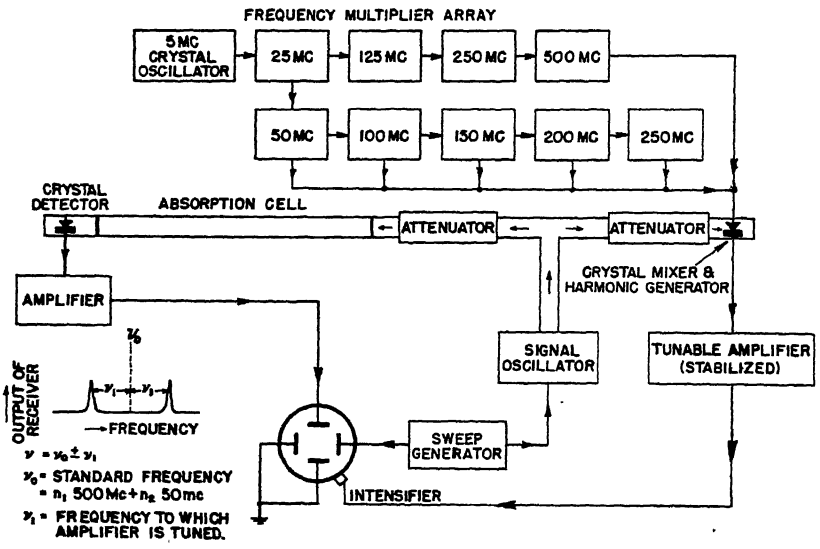


FIG. 14.—Microwave frequency-measuring equipment. Crystal-controlled frequency markers are positioned 50 Mc apart throughout the microwave frequency range 18,000–36,000 Mc. Interpolation between the fixed markers is by means of a tunable receiver, which measures difference frequencies between the microwave generator and a fixed frequency marker.

triodes. The output from the last doubling stage is then taken to a crystal harmonic generator, which generates standard frequency signals throughout the microwave region.

Part of the output signal from the microwave oscillator is imposed on this same crystal. The difference frequencies generated, due to the standard frequency signals beating against the signal oscillator, are then amplified and measured by means of a communications receiver.

As shown in the lower left corner of Fig. 14, a signal is transmitted by the communications receiver whenever the signal is tuned to a frequency $\nu = \nu_0 \pm \nu_1$ where ν_0 is one of the standard frequencies, and ν_1 is the frequency to which the communications receiver is tuned.

By feeding a 50 Mc signal from the frequency multiplier array into the crystal mixer, it is possible to generate standard frequency markers throughout the region 18,000–36,000 Mc, at intervals of only 50 Mc.

If the absorption line is viewed on the screen of an oscilloscope, then the signal from the communications receiver may be applied to the

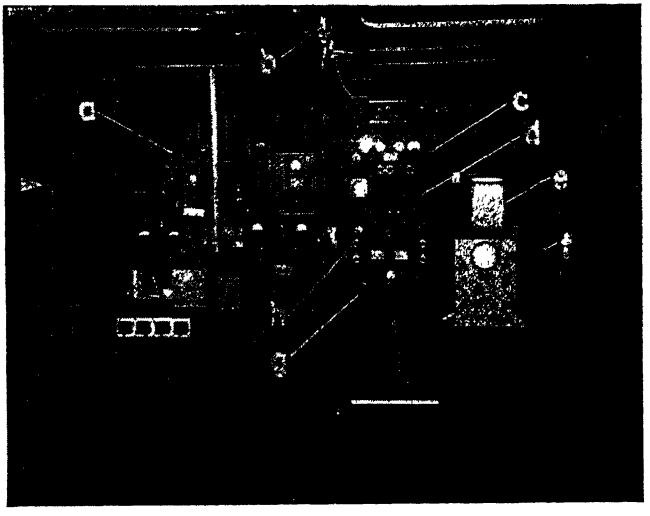


FIG. 15.—Microwave spectrograph employing Stark effect modulation
a—Gas-handling apparatus
b—Waveguide absorption cell
c—Power supply for microwave generator
d—Oscilloscope for viewing absorption lines
e—Spectrum recorder
f—Monitoring oscilloscope
g—Crystal controlled oscillator and frequency multiplier
h—Communications receiver for interpolating between fixed frequency markers.

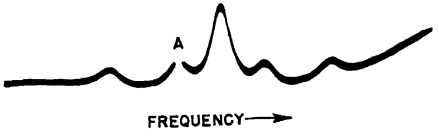


FIG. 16.—Precise measurement of absorption line frequency. The frequency marker A is adjusted to coincide with the peak of an absorption line.

intensifier of the oscilloscope, so that reception of the beat frequency is indicated by a bright spot or a blank spot on the oscilloscope trace. An absorption line frequency may then be measured by adjusting ν_1 until the blank spot coincides with the peak of an absorption line. This type of measurement is illustrated in Fig. 16. A dual beam oscilloscope has

also been found convenient for making frequency measurements and for monitoring the Klystron output.

It has been found possible to repeat frequency readings in the 25,000 Mc region to .01 Mc. With care, it appears that it should be possible to measure the frequency of a sharp absorption line to one part in 10^7, by using a stabilized tunable amplifier in place of an ordinary communications receiver.

The frequency ν_0 may be obtained from a variable frequency oscillator if desired, so that the two pips separated by the distance $2\nu_1$, may be moved about (like a pair of calipers) to compare frequency differences in different parts of the spectrum.

This method may also be used to measure absorption line frequencies of new compounds by comparison with absorption lines of known frequency, such as the well-known lines in the inversion spectrum of ammonia.[89,90] Relative frequency measurements on microwave absorption lines have been described by Carter and Smith.[91] The standard frequency apparatus which is used at the M.I.T. Radiation Laboratory has been described in Chapter 6 of reference 72. A microwave secondary frequency standard which gives markers every 90 Mc from 2970 Mc to above 40,000 Mc has been described by Unterberger and Smith.[92]

REFERENCES

1. Herzberg, G. Infrared and Raman Spectra of Polyatomic Molecules. Van Nostrand, New York, 1945.
2. Cleeton, C. E., and Williams, N. H. *Phys. Rev.*, **45**, 234–37 (1934).
3. Howe, H. S. Ph.D. Thesis, University of Michigan, 1940.
4. Van Vleck, J. H. *Phys. Rev.*, **71**, 413–24 (1947).
5. Van Vleck, J. H. *Phys. Rev.*, **71**, 425–33 (1947).
6. Beringer, R. *Phys. Rev.*, **70**, 53–7 (1946).
7. Becker, G. E., and Autler, S. H. *Phys. Rev.*, **70**, 300–307 (1946).
8. Bleaney, B., and Penrose, R. P. *Nature, Lond.*, **157**, 339L (1946).
9. Good, W. E. *Phys. Rev.*, **69**, 539L (1946).
10. Townes, C. H. *Phys. Rev.*, **70**, 109A (1946).
11. Gordy, W. *Rev. Mod. Phys.*, **20**, 668–717 (1948).
12. Kisliuk, P., and Townes, C. H. *J. Res. Nat. Bur. Stand., Wash.* (to be published).
13. Roberts, A., Beers, L. Y., and Hill, A. G. *Phys. Rev.*, **70**, 112A (1946).
14. Golden, S., and Wilson, E. B., Jr. *J. Chem. Phys.*, **16**, 669–85 (1948).
15. Strandberg, M. W. P., Wentink, T., Jr., and Hill, A. G. *Phys. Rev.*, **75**, 827–32 (1949).
16. Dakin, T. W., Good, W. E., and Coles, D. K. *Phys. Rev.*, **70**, 560L (1946).
17. Coles, D. K., and Good, W. E. *Phys. Rev.*, **70**, 979L (1946).
18. Jen, C. K. *Phys. Rev.*, **74**, 1396–1406 (1948).
19. Jen, C. K. *Phys. Rev.*, **75**, 1318A (1949).
20. Condon, E. U. *Phys. Rev.*, **30**, 781–84 (1927).
21. Condon, E. U., and Shortley, G. H. The Theory of Atomic Spectra. MacMillan, New York, 1935.

22. Casimir, H. B. G. On The Interaction Between Atomic Nuclei and Electrons. E. F., Bohn, Haarlem, The Netherlands, 1936.
23. Van Vleck, J. H. *Phys. Rev.*, **71**, 468A (1947).
24. Dailey, B. P., Kyhl, R. L., Strandberg, M. W. P., Van Vleck, J. H., and Wilson, E. B., Jr. *Phys. Rev.*, **70**, 984L (1946).
25. Simmons, J. W., and Gordy, W. *Phys. Rev.*, **73**, 713–18 (1948).
26. Jauch, J. M. *Phys. Rev.*, **74**, 1262A (1948).
27. Henderson, R. S. *Phys. Rev.*, **74**, 107L and 626L (1948).
28. Bardeen, J., and Townes, C. H. *Phys. Rev.*, **73**, 627–29 and 1204L (1948).
29. Bardeen, J., and Townes, C. H. *Phys. Rev.*, **73**, 97–105 (1948).
30. Bragg, J. K. *Phys. Rev.*, **74**, 533–38 (1948).
Knight, G., and Feld, B. T. *Phys. Rev.*, **74**, 354A (1948).
Bragg, J. K., and Golden, S. *Phys. Rev.*, **75**, 735–38 (1949).
31. Fano, U. *J. Res. Nat. Bur. Stand., Wash.*, **40**, 215–23 (1948).
Karplus, R., and Sharbaugh, A. H. *Phys. Rev.*, **75**, 1325A (1949).
32. Low, W., and Townes, C. H. *Phys. Rev.*, **75**, 1318A (1949).
33. Jauch, J. M. *Phys. Rev.*, **72**, 715–23 (1947).
34. Dakin, R. W., Good, W. E., and Coles, D. K. *Phys. Rev.*, **71**, 640A (1947).
35. Townes, C. H., Holden, A. H., and Merritt, F. R. *Phys. Rev.*, **74**, 1113–33 (1948).
36. King, G. W., Hainer, R. M., and Cross, P. C. *J. Chem. Phys.*, **11**, 27–42 (1943).
37. Dailey, B. P., Golden, S., and Wilson, E. B., Jr. *Phys. Rev.*, **72**, 871L (1947).
38. Golden, S. *J. Chem. Phys.*, **16**, 250–53 (1948) and **17**, 586L (1949).
39. Herzberg, G. Molecular Spectra and Molecular Structure-Diatomic Molecules. Van Nostrand, New York, 1950.
40. Strandberg, M. W. P., and Wentink, T. Jr. *Phys. Rev.*, **73**, 1249A (1948).
41. Townes, C. H., Holden, A. H., and Merritt, F. R. *Phys. Rev.*, **71**, 64L (1947).
42. Wentink, T. Jr., Strandberg, M. W. P., and Hillger, R. *Phys. Rev.*, **73**, 1249A (1948).
43. Townes, C. H., and Dailey, B. P. *J. Chem. Phys.*, **17**, 782–96 (1949).
44. Koehler, J. S., and Dennison, D. M. *Phys. Rev.*, **57**, 1006–21 (1940).
Dennison, D. M., and Burkhard, D. G. Symposium on Molecular Structure and Spectroscopy. (Ohio State Univ., Columbus, Ohio, 1948 and 1949.)
45. Wilson, E. B., Jr. *Chem. Revs.*, **27**, 17–38 (1940).
46. Townes, C. H. *Phys. Rev.*, **71**, 909L (1947).
47. Townes, C. H., Holden, A. N., Bardeen, J., and Merritt, F. R., *Phys. Rev.*, **71**, 644L (1947).
48. Smith, W. V. *Phys. Rev.*, **71**, 126L (1947).
49. Strandberg, M. W. P., Kyhl, R., Wentink, T. Jr., and Hillger, R. E. *Phys. Rev.*, **71**, 326L (1947).
50. Good, W. E., and Coles, D. K. *Phys. Rev.*, **72**, 157A (1947).
51. Nielson, H. H., and Dennison, D. M. *Phys. Rev.*, **72**, 1101–08 (1947).
52. Townes, C. H., Merritt, F. R., and Wright, B. D. *Phys. Rev.*, **73**, 1334–37 (1948).
53. Dailey, B. P. *Anal. Chem.*, **21**, 540–44 (1949).
54. Bleaney, B., and Penrose, R. P. *Proc. Phys. Soc., Lond.*, **60**, 395L (1948).
55. Kittel, C., and Luttinger, J. M. *Phys. Rev.*, **73**, 162–172 (1948).
56. Kip, A. F., and Arnold, R. D. *Phys. Rev.*, **75**, 1199–1205 (1949).
57. Kittel, C. *Phys. Rev.*, **73**, 155–61 (1948).
58. Bleaney, B., and Penrose, R. P. *Proc. Phys. Soc. Lond.*, **59**, 418–28 (1947).
59. Pound, R. V. *Rev. Sci. Instrum.*, **17**, 490–505 (1946); Hershberger, W. D. *Elec. Engng. N.Y.*, **68**, 251A (1949); Chaffee, Fletcher, and Cooke Craft Labora-

tory Progress Reports on Contract N5-ORI-76. Harvard Univ., Cambridge, Mass., 1947–49; Lyons, H. *Elec. Engng. N.Y.*, **68**, 251A (1949).
60. Townes, C. H. *Phys. Rev.*, **70**, 665–71 (1947).
61. Pond, T. A., and Cannon, W. F. *Phys. Rev.*, **72**, 1121L (1947).
62. Schwinger, J. S., and Karplus, R. *Phys. Rev.*, **73**, 1020–26 (1948).
63. Carter, R. L., and Smith, W. V. *Phys. Rev.*, **73**, 1053–58 (1948).
64. Bleaney, B., and Penrose, R. P. *Phys. Rev.*, **70**, 775L (1946).
65. Mizushima, M. *Phys. Rev.*, **74**, 705L (1948).
66. Smith, D. F. *Phys. Rev.*, **74**, 506L (1948).
67. Bleaney, B., and Loubser, J. H. N. *Nature, Lond.*, **161**, 523L (1948); Anderson, P. W. *Phys. Rev.*, **76**, 647–61 (1949); Margenau, H. *Phys. Rev.* **76**, 121–24 (1949).
68. Van Vleck, J. H., and Weisskopf, V. F. *Rev. Mod. Phys.*, **17**, 227–36 (1945).
69. Casimir, H. B. G. *Z. Phys.*, **59**, 623–34 (1929).
70. Cross, P. C., Hainer, R. M., and King, G. W. *J. Chem. Phys.*, **12**, 210–43 (1944); Hainer, R. M., King, C. W., and Cross, P. C. *J. Chem. Phys.*, **17**, 826–36 (1949).
70a. Wesson, L. G. Tables of Electric Dipole Moments. MIT Technology Press, Cambridge, Mass., 1948.
71. Karplus, R. *Phys. Rev.*, **73**, 1120L (1948).
72. Montgomery, C. G. and others. Technique of Microwave Measurements. McGraw-Hill, New York, 1947.
73. Smith, A. G., Gordy, W., Simmons, J. W., and Smith, W. V. *Phys. Rev.*, **75**, 260–63 (1949).
74. Townes, C. H., and Merritt, F. R. *Phys. Rev.*, **72**, 1266L (1947).
75. Karplus, R. *Phys. Rev.*, **73**, 1027–34 (1948).
76. Strandberg, M. W. P., Pearsall, C. S., and Weiss, M. T. *J. Chem. Phys.*, **17**, 429L (1949).
77. Loubser, J. H. N., and Townes, C. H. *Phys. Rev.*, **76**, 178A (1949).
78. Weingarten, I. R., and Kimball, G. E. *Phys. Rev.*, **75**, 1325A (1949).
79. Hamilton, D. R., Knipp, J. K., and Kuper, J. B., H. Klystrons and Microwave Triodes. McGraw-Hill, New York, 1948.
80. Torrey, H. C., and Whitmer, C. A. Crystal Rectifiers. McGraw-Hill, New York, 1948.
81. Van Voorhis, S. N. and others. Microwave Receivers. McGraw-Hill, New York, 1948, Vol. 23.
82. Townes, C. H., and Geschwind, S. *J. Appl. Phys.*, **19**, 795L (1948).
83. Ragan, G. L. and others. Microwave Transmission Circuits. McGraw-Hill, New York, 1948.
84. Hughes, R. H., and Wilson, E. B., Jr. *Phys. Rev.*, **71**, 562L (1947).
85. Hershberger, W. D. *J. Appl. Phys.*, **19**, 411–19 (1948).
86. Gordy, W., and Kessler, M. *Phys. Rev.*, **72**, 644L (1947).
87. Watts, R. J., and Williams, D. *Phys. Rev.*, **72**, 1122L (1947).
88. Ramo, S., and Whinnery, J. R. Fields and Waves in Modern Radio. Wiley, New York, 1944.
89. Strandberg, M. W. P., Kyhl, R., Wentink, T., Jr., and Hillger, R. E. *Phys. Rev.*, **71**, 326L (1947).
90. Good, W. E., and Coles, D. K. *Phys. Rev.*, **71**, 383L (1947).
91. Carter, R. L., and Smith, W. V. *Phys. Rev.*, **72**, 1265L (1947).
92. Unterberger, R., and Smith, W. V. *Rev. Sci. Instrum.*, **19**, 580–5 (1948).

Author Index

Numbers in parentheses are reference numbers. They are included to assist in locating references in which the authors' names are not mentioned in the text. Numbers in italics refer to the page on which the reference is listed in the bibliography at the end of each article.

Example: Aten, A. H. W., 11(2), *44*, indicates that this author's article is reference 2 on page 11 and is listed on page 44.

L

M

Subject Index

F

N